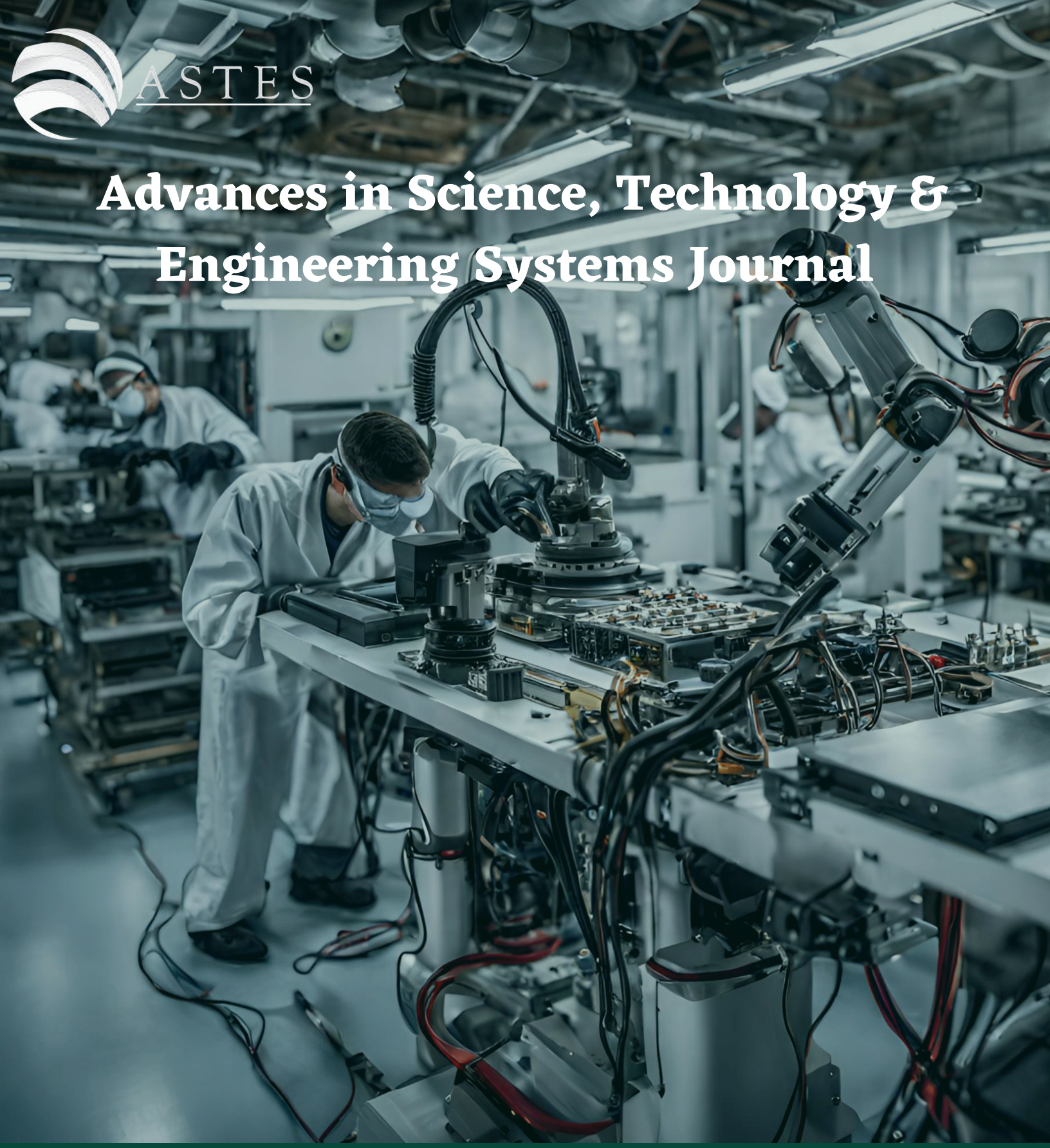


ASTES

Advances in Science, Technology & Engineering Systems Journal

VOLUME 8-ISSUE 4|JUL-AUG 2023

www.astesj.com
ISSN: 2415-6698

Editorial

In this issue, we present a diverse collection of 13 accepted papers covering a wide range of topics within the realms of technology, robotics, education, renewable energy, communication systems, medical imaging, logistics, and storytelling. These papers contribute valuable insights to their respective fields, offering innovative solutions, experimental results, and practical implications.

The first paper addresses the growing demand for connected devices in the Narrow-band Internet of Things (NB-IoT) and presents a novel Symbol Time Compression (STC) based approach to double the number of connected devices while maintaining system performance. The proposed method efficiently utilizes the available bandwidth, reducing the occupied bandwidth of each device without compromising complexity or performance [1].

The second paper delves into the realm of soft robotics, focusing on the design and manufacturing of soft grippers for robotics using injection molding technology. The study introduces a strategy for optimizing injection molding to enable mass production of soft grippers, paving the way for more cost-effective and commercially viable soft robotic solutions [2].

Moving underwater, the third paper introduces a selective modal analysis algorithm for localizing impulsive sound sources in shallow waters. The proposed algorithm utilizes modal dispersions to enhance the accuracy and noise-resistivity of sound source localization, offering a 2D localization method with minimal hardware requirements in a noisy underwater environment [3].

Shifting gears to the field of education, the fourth paper investigates the impression effects of a teacher-type robot equipped with a Perplexion Estimation Method on college students. The study introduces an autonomous learning support approach that estimates students' states of perplexity through facial expressions, demonstrating the potential for enhancing learning environments [4].

Next, the fifth paper presents a fuzzy Maximum Power Point Tracking (MPPT) technique for photovoltaic systems based on custom defuzzification. The proposed method addresses the nonlinearity of solar modules and demonstrates improved performance compared to traditional MPPT methods [5].

The sixth paper takes a circuit designer's perspective to Metal Oxide Semiconductor Field-Effect Transistors (MOSFETs) behavior, offering a comprehensive explanation of MOSFET behavior and providing practical insights into interpreting AC parameters in modern MOSFET models [6].

In the realm of renewable energy, the seventh paper introduces a novel approach for frequency control support of a wind turbine generator using Proportional Derivative (PD) and Proportional Integral Derivative (PID) controllers. The study demonstrates the effectiveness of the proposed controllers in maintaining system stability and frequency within acceptable limits [7].

The eighth paper addresses the distribution management problem in the Moroccan petroleum sector, presenting a heuristic solution for the Vehicle Routing Problem with Time Windows (VRPTW). The study aims to minimize transport costs and optimize the number of trucks to improve the competitiveness of transportation operations [8].

Moving to medical imaging, the ninth paper focuses on MRI semantic segmentation based on an optimized V-net with 2D attention. The study explores deep learning models for accurate tumor segmentation in brain MRI images, demonstrating an effective approach for the diagnosis of oncological brain diseases [9].

The tenth paper shifts to the realm of communication systems, presenting the FPGA implementation of 5G NR LDPC codes. The study showcases the implementation of 5G NR LDPC codes on a FPGA platform, demonstrating high coding gain, throughput, and low power dissipation [10].

The eleventh paper explores the application of lean practices in the food supply chain in Morocco. The study assesses the implementation status of lean practices in agri-food companies, emphasizing their impact on both operational and environmental performance [11].

In the twelfth paper, the authors propose colorized iVAT images for labeled data as a visualization technique for higher-dimensional data sets. The study introduces new types of colorized iVAT images to represent spatial cluster structures and the distribution of labels among clusters in data sets of dimensions 4 or greater [12].

Finally, the thirteenth paper introduces "Seven ReImagined," a transmedia storytelling evolution proposal that reshapes the exploration of the seven deadly sins in a modern context. The project creatively blends traditional art with cutting-edge immersive technologies, fostering a multifaceted dialogue on ancient moral wisdom in contemporary society [13].

In summary, this issue showcases a diverse range of papers that contribute significantly to their respective fields, offering innovative solutions, experimental results, and practical implications for researchers, practitioners, and enthusiasts alike.

References:

[1] A. Mohammed, M.S. Elbakry, H. Mostafa, A.A. Ammar, “Doubling the Number of Connected Devices in Narrow-band Internet of Things while Maintaining System Performance: An STC-based Approach,” Advances in Science, Technology and Engineering Systems Journal, **8**(4), 1–10, 2023, doi:10.25046/aj080401.

[2] H.D. Bryantono, M.R.F. Saduk, J. Hong, M.-H. Tsai, S.-C. Tseng, “Design and Manufacturing of Soft Grippers for Robotics by Injection Molding Technology,” Advances in Science, Technology and Engineering Systems Journal, **8**(4), 11–17, 2023, doi:10.25046/aj080402.

[3] F. Talebpour, S. Mozaffari, M. Saif, S. Alirezaee, “Localization of Impulsive Sound Source in ShallowWaters using a Selective Modal Analysis Algorithm,” Advances in Science, Technology and Engineering Systems Journal, **8**(4), 18–27, 2023, doi:10.25046/aj080403.

[4] K. Okawa, F. Jimenez, S. Akizuki, T. Yoshikawa, “Investigating the Impression Effects of a Teacher-Type Robot Equipped a Perplexion Estimation Method on College Students,” Advances in Science, Technology and Engineering Systems Journal, **8**(4), 28–35, 2023, doi:10.25046/aj080404.

[5] A. Allaoui, M.N. Tandjoui, C. Benachaiba, “Fuzzy MPPT for PV System Based on Custom Defuzzification,” Advances in Science, Technology and Engineering Systems Journal, **8**(4), 36–40, 2023, doi:10.25046/aj080405.

[6] R. Sommer, C.T. Gatermann, F. Vierling, R. Sommer, “A Circuit Designer’s Perspective to MOSFET Behaviour: Common Questions and Practical Insights,” Advances in Science, Technology and Engineering Systems Journal, **8**(4), 41–59, 2023, doi:10.25046/aj080406.

[7] A.A. Jhumka, R.T.F.A. King, C. Ramasawmy, A. Khoodaruth, “Proportional Derivative and Proportional Integral Derivative Controllers for Frequency Support of a Wind Turbine Generator in a Diesel Generation Mix,” Advances in Science, Technology and Engineering Systems Journal, **8**(4), 60–65, 2023, doi:10.25046/aj080407.

[8] Y.F. El Bahi, L. Ezzine, Z. Aman, I. Moussaoui, M. Rahmoune, H. El Moussami, “Distribution Management Problem: Heuristic Solution for Vehicle Routing Problem with Time Windows (VRPTW) in the Moroccan Petroleum Sector,” Advances in Science, Technology and Engineering Systems Journal, **8**(4), 66–72, 2023, doi:10.25046/aj080408.

[9] Z.H.N. Al-azzwi, A.N. Nazarov, “MRI Semantic Segmentation based on Optimize V-net with 2D Attention,” Advances in Science, Technology and Engineering Systems Journal, **8**(4), 73–80, 2023, doi:10.25046/aj080409.

[10] S.F. Abdel-Momen, A.H.A. Zekry, A.Y. Hassan, W.I. Shaban, M.M. Shiple, “FPGA Implementation of 5G NR LDPC Codes,” Advances in Science, Technology and Engineering Systems Journal, **8**(4), 91–100, 2023, doi:10.25046/aj080411.

[11] F. Bouhannana, A. El Korchi, “Application of Lean Practices in Food Supply Chain: The Case of Morocco,” Advances in Science, Technology and Engineering Systems Journal, **8**(4), 101–110, 2023, doi:10.25046/aj080412.

[12] E.D. Hathaway, R.J. Hathaway, “Colorized iVAT Images for Labeled Data,” Advances in Science, Technology and Engineering Systems Journal, **8**(4), 111–121, 2023, doi:10.25046/aj080413.

[13] J. Braguez, “SEVEN ReImagined: A Transmedia Storytelling Evolution Proposal,” Advances in Science, Technology and Engineering Systems Journal, **8**(4), 122–130, 2023, doi:10.25046/aj080414.

Editor-in-chief

Prof. Passerini Kazmersk

ADVANCES IN SCIENCE, TECHNOLOGY AND ENGINEERING SYSTEMS JOURNAL

Volume 8 Issue 4 **July-August 2023**

CONTENTS

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 8, No. 4, 01-10 (2023)
www.astesj.com

ASTES Journal
ISSN: 2415-6698

Doubling the Number of Connected Devices in Narrow-band Internet of Things while Maintaining System Performance: An STC-based Approach

Abdulwahid Mohammed[*,1], Mohamed S. Elbakry [2], Hassan Mostafa [3,4], Abdelhady Abdelazim Ammar[1]

[1]*Department of Electronics and Communication Engineering, Al-Azhar University, Cairo, 11651, Egypt*

[2]*Department of Electronics and Communications Engineering, Institute of Aviation Engineering and Technology, Giza, 25152, Egypt.*

[3]*Department of Electronics and Communication Engineering, Cairo University, , Giza, 12613, Egypt*

[4]*Department of Electronics and Communication Engineering, University of Science and technology, Giza, 12578, Egypt*

ARTICLE INFO

Article history:
Received: 18 January, 2023
Accepted: 01 July, 2023
Online: 21 July, 2023

Keywords:
NB-IoT
5G
LPWAN
OFDM
STC
STC-OFDM

ABSTRACT

Narrow-band Internet of Things (NB-IoT) is a low-power wide-area network (LPWAN) method that was first launched by the 3rd generation partnership project (3GPP) Rel-13 with the purpose of enabling low-cost, low-power, and wide-area cellular connections for the Internet of Things (IoT). As the demand for over-the-air services grows and with the number of linked wireless devices reaching 100 billion, wireless spectrum is becoming scarce, necessitating creative techniques that can increase the number of connected devices within a restricted spectral resource to satisfy service needs. Consequently, it is vital that academics develop efficient solutions to fulfill the quality of service (QoS) criteria of the NB-IoT in the context of 5th generation (5G). This study paves the way for 5G networks and beyond to have increased capacity and data rates for NB-IoT. Whereas, this article suggests a method for increasing the number of connected devices by using a technique known as symbol time compression (STC). The suggested method reduces the occupied bandwidth of each device without increasing complexity, losing data throughput, or affecting bit error rate (BER) performance. The STC approach is proposed in the literature to work with conventional orthogonal frequency-division multiplexing (OFDM) to reduce bandwidth usage by 50% and enhance the peak-to-average power ratio (PAPR). Specifically, an STC-based technique is suggested that exploits the available bandwidth to increase the number of linked devices by double while keeping the complexity and performance of the system. Furthermore, the μ-law companding technique is leveraged to reduce the PAPR of the transmitted signals. The obtained simulation results indicate that the suggested method using the μ-law companding technique doubles the amount of transferred data and lowers the PAPR by 3.22 dB while keeping the same complexity and BER.

1 Introduction

The IoT has changed dramatically in recent years. In particular, the number of IoT devices is rapidly expanding, and various novel IoT applications relating to automobiles, transportation, the power grid, agriculture, metering, and other areas have emerged. According to the Forbes analysis 2017, the worldwide IoT industry will be worth $457 billion by 2020 [1]. However, this amount was already surpassed in 2019 with a worldwide market size of $465 billion (Transforma Insights, 2020) [2]. According to the recent study provided by [2], there were 7.6 billion active IoT devices at the end of 2019, which is expected to rise to 24.1 billion by 2030, representing 11% compound annual growth rate (CAGR), as shown in Figure 1.

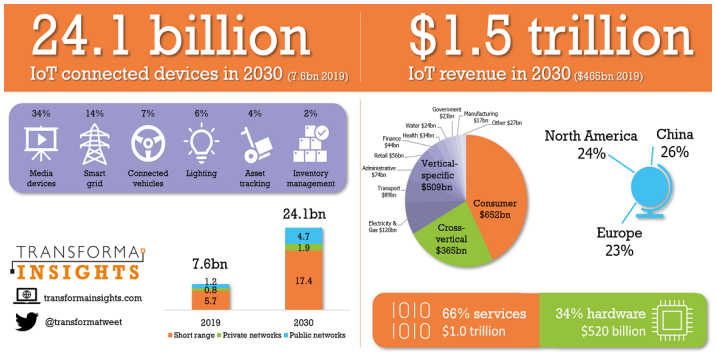


Figure 1: The IOT market 2019-2030 (Transforma Insights, 2020) [2].

[*]Corresponding Author: Abdulwahid Mohammed, Email: awh9140@gmail.com

Several low-power wide area (LPWA) technologies have been developed to meet this enormous demand for data with the goals of increasing network coverage, enhancing power consumption, supporting more users, and reducing device complexity. Different standard-development organizations, including IEEE and 3GPP, work to standardize LPWA technologies. Both cellular and non-cellular wireless technology can be used in LPWA. Machine Type Communication (MTC), Improved Machine Type Communication (eMTC), and Narrowband Internet of Things (NB-IoT) are examples of cellular technologies, while non-cellular technologies include Long Range (LoRa), ZigBee, Bluetooth, Z-Wave, and others[3]. With the fast development of 5G new radio technologies, intensive research on enhanced mobile broadband (eMBB), massive machine-type communications (mMTCs), and ultrareliable low latency communications (URLLCs) have gotten a lot of interest from academics and industry [4]. To satisfy the 5G outlook, it is necessary not only to achieve significant improvements in new wireless technologies but also to take into account the harmonious and fair coexistence of heterogeneous networks and compatibility between 4G and 5G systems [5]. As shown in Figure 2, the unprecedented growth of the IoT creates an enormous demand for MTC, which can be divided into three categories: 1) short-distance MTC (distance = 10 m), 2) medium-distance MTC (distance ranges between [10 m, 100 m]), and 3) long-distance MTC (distance ≥ 100 m).

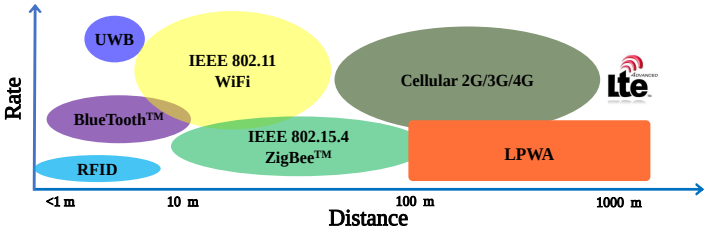


Figure 2: Coverage and transmission rate comparisons of wireless systems [6].

2 Literature Review and Contribution

The maximum likelihood of cross-correlation detection is presented by the authors in [7] as a hardware implementation. The detector attains an average detection delay that can minimize the power needed per time acquisition by up to 34%. However, it is a computationally difficult detection approach. The authors of [8] provide a technique for NB-IoT UEs to decrease power consumption during paging loading and offloading. The suggested approach in [8] has the potential to decrease power consumption by around 80% and enhance energy utilization efficiency by about 30.5%. But in standalone mode, this method is not applicable. In [9], a semi-Markov chain model with four states, namely, power saving mode (PSM), idle, random access (RACH), and transmission (Tx) state, is developed to evaluate the NB-IoT energy consumption and delay for periodic up-link traffic. However, the model does not account for the energy used while switching between the four modes listed above or the repetition effect on power consumption. In [10], the authors compare NB-IoT coverage under various conditions using 15 kHz and 3.75 kHz spacing. When compared to the present LTE technology, the coverage improvement is greater than 20 dB. However, the reported 170 dB of realized maximum coupling loss (MCL) of NB-IoT does not take into account the impairments of channel estimation, carrier offset, or mobility with regard to various configurations. The authors of [11] propose the Link adaptation algorithm, which uses the Shannon theorem to improve coverage by characterizing signal-to-noise ratio (SNR), repetition number, and NB-IoT supported bandwidth. Nevertheless, the influence of channel state information on user equipment (UE) link adaptability was not considered in this study. In [12], the authors propose a method in order to double the number of connected devices by utilizing Fast-OFDM. In comparison to a standard OFDM system, the Fast-OFDM approach decreases the distance between sub-carriers by half, saves 50% of the bandwidth, and prevents the BER from degrading. However, the proposed approach would result in a mismatch in sampling rate and a carrier frequency offset (CFO). Furthermore, this method is still plagued by the PAPR issue. Unlike the previous studies, the suggested method in this work improves system performance by reducing the PAPR issue.

In summary, the advantages of our suggested method are itemized as follows:

- It can decrease the used bandwidth in half by halving the number of sub-carriers, allowing it to transmit twice as much data as a conventional OFDM system.
- When compared to the Fast-OFDM and conventional OFDM systems, it enhances system performance by minimizing the PAPR issue.
- It preserves system performance by preventing BER from deteriorating, as the BER of our suggested technique is precisely equivalent to the BER in the Fast-OFDM and conventional OFDM systems.
- Although it can transport twice as much data as a standard OFDM system, it has nearly the same complexity.
- Since it does not decrease the space between the sub-carriers, it does not result in a sampling rate mismatch or carrier frequency offset (CFO) as in Fast-OFDM.

3 A summary of LPWA technologies

Recently, lots of new LPWA technologies have been proposed for a variety of uses. Some of the most widely used are LoRa, SigFox, eMTC, and NB-IoT [6] chen2017narrowband. In this section, A brief comparison of each technology, as referred to in Table 1.

- **LoRa:** LoRa was developed by Semtech Company and is based on patented spread spectrum algorithms and Gaussian frequency shift keying. It is regarded as the first low-cost wide-area deployment for commercial use. To avoid in-band and out-band interference, LoRa uses chirp spread spectrum (CSS) and Gaussian frequency shift keying (GFSK) modulation, which may work up to 25 dB below the noise level. LoRa's bandwidth requirements might range from 7.8 kHz to 500 kHz. The predicted coverage range in urban areas is 2-5 kilometers, while it is approximately 15 kilometers in suburban areas [14].
- **SigFox:** SigFox also employs unlicensed spectrum through the use of special technologies. It operates at 868 MHz in

Table 1: Comparisons between LoRa, SigFox, eMTC, and NB-IOT technologies [13].

Technologies / Parameters	LoRa	SigFox	eMTC	NB-IoT
Spectrum	Unlicensed	Unlicensed	Licensed	Licensed
Modulation	CSS	UL: DBPSK DL: GFSK	UL: SC-FDMA DL: OFDMA	UL: SC-FDMA DL: OFDMA
Bandwidth	7.8 – 500 kHz	200 kHz	1.08 MHz	180 kHz
Range	Urban : 2–5 km Suburban : ~ 15 km	Urban : 3–10 km Suburban : 30–50 km	Urban : ~ 5 km Suburban : ~ 17 km	Urban : 1 – 8 km Suburban : ~ 25 km
Data rate	< 50 bps	< 100 bps (EU) < 600 bps (USA)	< 1 Mbps	160 – 250 kbps (DL) 160 – 200 kbps (UL)
Battery life	> 10 years	8 – 10 years	5 – 10 years	> 10 years
Price	< 5 $	< 10 $	< 10 $	< 5 $
Note :- **UL:** uplink , **DL:** downlink and **CSS:** chirp spread spectrum				

Europe and 915 MHz in the United States. For up-link and down-link broadcasts, SigFox uses differential binary phase shift keying (DBPSK) and Gaussian frequency shift keying (GFSK) modulation. The bandwidth requirement is around 200 kHz, with a data rate of fewer than 600 kbps in the US and 100 kbps in Europe [15].

- **eMTC:** Enhanced machine-type communication (eMTC) is a novel kind of data transmission that includes one or more entities without the need for humans. The 3GPP has standardized eMTC, which uses LTE infrastructure to operate. In comparison to other LPWAN technologies, eMTC may provide a comparatively high data rate of 1 Mbps at the expense of occupying a substantially wider frequency spectrum of 1.08 MHz inside the LTE band. The battery life of an eMTC can be extended to over 10 years using extended discontinuous reception (eDRX) and power savings management (PSM). However, the cost of eMTC end devices has increased as a result of the comparatively broader coverage and faster data rate, leaving it with no economic benefit [16].
- **NB-IoT:** NB-IOT is a novel 3GPP radio access technology that is intended to perform very well with traditional GSM and LTE technologies [17]. For both down-link and up-link communications, it needs a minimum system bandwidth of 180 kHz, and it may be employed in one of three modes: stand-alone, guard-band, or in-band. For these three operation modes, the down-link transmission technique is based on orthogonal frequency division multiple access (OFDMA) with 15 kHz sub-carrier spacing. NB-IoT typically allows up-link transmission at data rates of 160-200 kHz and down-link transmission at data rates of 160-250 kHz. It can serve regions of 1-8 km in urban environments and 25 km in suburban areas. When compared to the other three LPWA technologies, NB-IoT offers a lower cost of production, a longer working life, and wider coverage, as shown in Table 1.

4 Overview of NB-IoT

For next-generation use cases and applications, the NB-IoT provides LPWA coverage via massive devices [18]. NB-IoT is expected to be one of the technologies of 5G new radio (NR) networks, according to [19]. Figure 3(a) shows the applications of NB-IoT such as smart buildings [20], smart cities [21], intelligent or smart environmental monitoring systems [22], smart metering [23], and intelligent user services [24]. Also, Smart houses, smart wearable gadgets, smart people tracking, and other intelligent and smart user services. Pollution monitoring, intelligent agriculture, water quality monitoring, soil detection, and other aspects of the intelligent or smart environment monitoring system are described in [25, 26]. The main goals of NB-IoT are outlined in 3GPP specifications [3] and depicted in Figure 3(b).

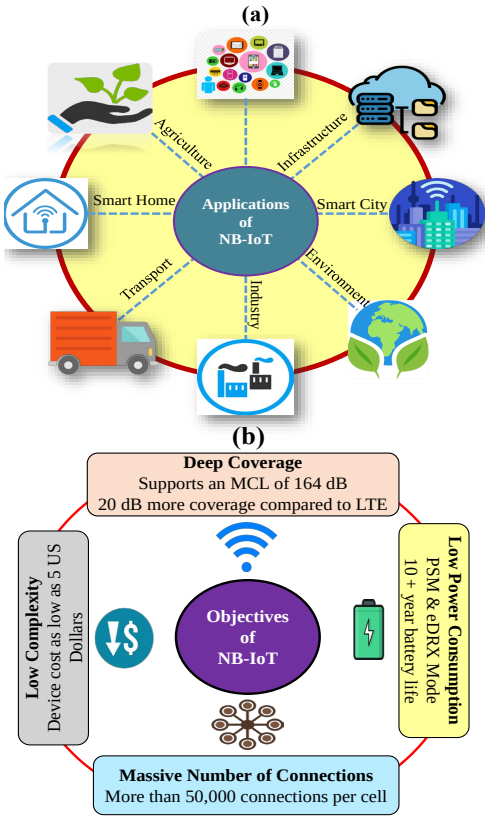


Figure 3: Applications and Objectives for NB-IOT: (a) Applications for NB-IoT and (b) Objectives for NB-IOT.

- **Deep Coverage:** The NB-IoT technology is designed to have both indoor and outdoor deep coverage. When compared to the conventional LTE network, NB-IoT provides up to 20 dB higher coverage [27]. NB-IoT has a maximum coupling loss (MCL) of 164 dB, whereas standard LTE has an MCL

of 144 dB. High coverage may be obtained, according to 3GPP specifications, by reducing bandwidth and increasing the number of data transmission repeats. Reduced bandwidth improves the PSD of the user, resulting in increased coverage. Nevertheless, there are two drawbacks to expanding coverage. Reduced bandwidth affects data throughput, whereas a high number of repeats raises data transfer delay and energy consumption [28].

- **Low Power Consumption:** NB-IoT products are designed to have a battery life of much more than ten years. To guarantee that NB-IoT products have such a long battery life, 3GPP Rel-12 and Rel-13 developed Power Saving Mode (PSM) and enhanced Discontinuous Reception (eDRX) modes. Both of these techniques strive to improve battery life in NB-IoT products by putting them into sleep mode when no data transfer is needed. PSM saves a significant amount of power and has a total sleep duration of 310 hours [29].
- **Low Complexity:** The cost of an NB-IoT device must be kept under $5. The infrastructure of NB-IoT has been simplified and improved to decrease devices price. When compared to the conventional LTE system, the network protocol volume, also the number of channels, signals, and transceivers, are reduced in NB-IoT. For both up-link and down-link transmissions, just one transceiver is used [30]. As NB-IoT only supports low-data-rate applications, it does not need a large memory, which leads to a decrease in the cost of devices.
- **Support of Massive Number of Connections:** The goal of NB-IoT is to serve over 50,000 users per cell. Using the NB-IOT system, the users transfer only a low data rate and delay-tolerant data. As a result, a single cell may effectively serve a massive number of devices. Furthermore, NB-IoT up-link transmission employs sub-carrier level transmission, which improves up-link resource use. For single-tone transmission, NB-IoT offers two frequencies: 15 and 3.75 kHz. When compared to conventional LTE, the signalling overhead of NB-IoT is also simplified [17].

5 System Model of the Suggested Method

Figure 4 shows the transceiver system model for the proposed method using the STC technique with a typical OFDM system (STC-OFDM). The STC-OFDM was initially presented as a wireless approach in [31], which was proven to have similar results to OFDM while saving 50% of bandwidth by compressing OFDM symbols by half. In particular, the STC technique uses a spreading and combining mechanism in the transmitter and a symbol time extension (STE) approach in the receiver to expand the received symbol [32], as shown in Figure 4.

First, the transmitted bits are converted into the polar forms b_0 and b_1. Then, using Walsh codes c_0 and c_1, these polar forms are spread out. Finally, the spread data, b_0c_0 and b_1c_1, are combined. The spreading procedure is accomplished by the use of two spreading codes, which are as follows [33]:

$$c_0 = [1 \quad 1] \quad and \quad c_1 = [1 \quad -1]. \tag{1}$$

Table 2 shows the spreading and combining operations in the STC scheme's transmitter.

Table 2: Spreading and combining process

Transmitted bits	Polar form of transmitted bits (b)	Walsh Code (C)	Spread data (bc)	Combining Data (d^c)
1	$b_0 = 1$	$c_0 = 1 \quad 1$	1 1	0 2
0	$b_1 = -1$	$c_1 = 1 \quad -1$	−1 1	

Following the spreading procedure, the spread data streams for each couple are joined as follows:

$$\begin{aligned} d_1^c &= b_1c_0 + b_2c_1 \\ d_2^c &= b_3c_0 + b_4c_1 \\ &\cdot \\ d_{N/2}^c &= b_{N-1}c_0 + b_Nc_1. \end{aligned} \tag{2}$$

Where b represents a bit stream, c_0 and c_1 are Walsh spreading codes with chip rate $T_c = T_b/2$ and d^c denotes the combining data symbol.

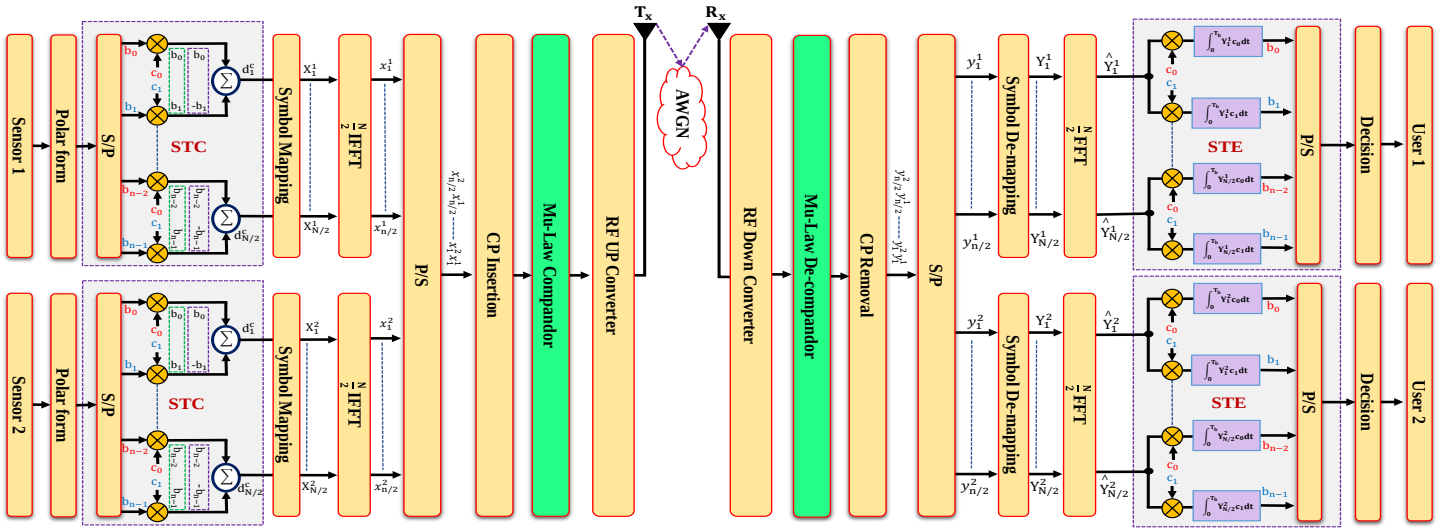


Figure 4: The transceiver system model for the proposed method using STC and $\mu-$law companding technique.

Figure 5 (a) depicts the waveform of the spreading process (b_0c_0) for the transmitted data b_0 and Figure 5 (b) displays the waveform of spreading process (b_1c_1) for the transmitted data b_1. While Figure 5 (c) shows the combined data (d^c).

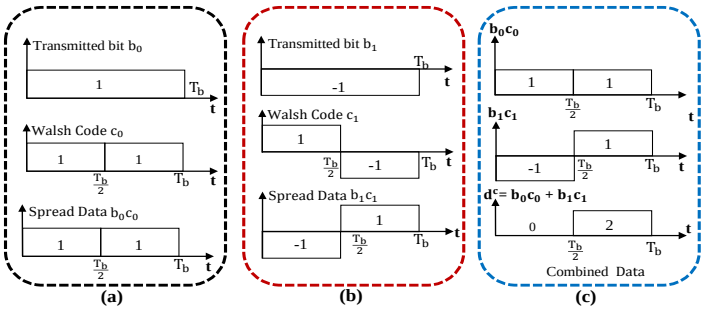


Figure 5: The wave-forms for: (a) spreading process of the transmitted data b_0, (b) spreading process of the transmitted data b_1 and (c) combined data

As displayed in Figure 4, the combining symbols are passed through the symbol mapping block to get the complex data symbols. The complex data symbol on the K^{th} sub-carrier is denoted by X_k^c, $k = 1, 2, ..., N/2$. The resulting $N/2$ signals are applied to the $N/2$ input ports of an inverse fast Fourier transform (IFFT) processor. A discrete-time OFDM symbol after IFFT is expressed as follows [33]:

$$x_k^c = \frac{2}{N}\sum_{m=0}^{\frac{N}{2}-1} X_m^c e^{j2\pi km/\frac{N}{2}}, \quad 0 \le k \le \frac{N}{2} - 1. \tag{3}$$

Where k represents the time index, N is the number of sub-carriers, x_k^c is the k^{th} OFDM symbol and X_m represents the m^{th} transmitted data symbols. The cyclic prefix (CP) is inserted in front of each OFDM signal as a guard interval (GI) between subsequent OFDM symbols in OFDM systems to keep the orthogonality criterion and prevent inter-carrier interference (ICI) and inter-symbol interference (ISI). Where the length of CP must be greater than the maximum multi-path channel delay spread ($T_{cp} \ge \tau_{max}$). A parallel-to-serial (P/S) converter is applied to the resulting time domain symbols. A cyclic prefix (CP) of a suitable length (L_{cp}) is added to combat the effect of multi-path propagation and the transmitted OFDM symbol with CP is defined as follows:

$$x_k^c(cp) = \frac{2}{N}\sum_{m=0}^{\frac{N}{2}-1} X_m^c e^{j2\pi km/\frac{N}{2}}, -L_{cp} \le k \le \frac{N}{2} - 1 \tag{4}$$

To recover the transmitted data, the transmitter processing is functionally reversed in reverse order at the receiver side, as shown in Figure 4, to yield an approximated form of the binary information sequence. The k^{th} received compressed OFDM symbol in the frequency domain is expressed as follows:

$$Y_k^c = \frac{2}{N}\sum_{m=0}^{\frac{N}{2}-1} y_m^c e^{-j2\pi km/\frac{N}{2}}, \quad 0 \le k \le \frac{N}{2} - 1 \tag{5}$$

Following symbol de-mapping, the STE technique is used to retrieve the data streams, as shown in Table 3.

Table 3: De-spreading and combining processes

R_x Data (d^c)	$d_0^c * c_0$	$\hat{b}_0 = \frac{1}{2}\sum d^c * c_0$	$d_1^c * c_1$	$\hat{b}_1 = \frac{1}{2}\sum d^c * c_1$
0 2	0 2	1	0 −2	−1

To de-spread the received data, it is multiplied by the same code as the transmitter (c_0 and c_1). Then, integration throughout the bit period is as follows:

$$\begin{aligned} \hat{b}_1 &= \int_0^{T_b} \hat{Y}_1^c c_0 dt, \quad \hat{b}_2 = \int_0^{T_b} \hat{Y}_1^c c_1 dt \\ \hat{b}_3 &= \int_0^{T_b} \hat{Y}_2^c c_0 dt, \quad \hat{b}_4 = \int_0^{T_b} \hat{Y}_2^c c_1 dt \\ &\vdots \\ \hat{b}_{N-1} &= \int_0^{T_b} \hat{Y}_{N/2}^c c_0 dt, \quad \hat{b}_N = \int_0^{T_b} \hat{Y}_{N/2}^c c_1 dt. \end{aligned} \tag{6}$$

where T_b denotes the bit duration and $\hat{b}_N$ is the detected N^{th} bit. The suggested approach is based on employing the STC technique twice in order to send data in the unused bandwidth, therefore doubling the number of connected devices. On the other hand, STE technique is employed twice on the receiver side to recover the data transferred from the transmitter to its original form, as shown in Figure 4.

6 Simulation Results and Discussion

This section discusses the numerical simulation and results for the proposed design. Table (4) provides the utilized simulation inputs of the system. Enhancement in PAPR, deterioration in BER, and power spectral density (PSD) are utilized as interesting performance metrics. Binary Phase Shift Keying (BPSK) modulation is used in this study for Fast-OFDM, STC-OFDM, and typical OFDM. Fast-OFDM is only applicable with one-dimensional modulation schemes, as was mentioned in [12]. Fast-OFDM and STC-OFDM systems are unable to handle higher-order modulation forms like Quadrature Phase Shift Keying (QPSK), which NB-IoT supports. Nevertheless, a developed non-orthogonal wave-forms known as "SEFDM" [34], may employ modulation schemes up to 16QAM at the cost of larger and more complexity. The improvement in PAPR is the difference between the PAPR of the conventional OFDM signal (original signal) and the PAPR of the STC-OFDM signal (i.e. Improvement in PAPR = PAPR of conventional OFDM - PAPR of STC-OFDM signal).Similarly, the difference between the BER of the original signal and the BER of the STC-OFDM signal is what is known as the BER deterioration.

Table 4: Simulation Parameters [12]

Parameter	OFDM	Fast-OFDM	STC-OFDM
Occupied Channel BW	180 kHz	90 kHz	90 kHz
Bit rate (kbit/s)	180 Kbps	180 Kbps	180 Kbps
Spacing frequency, Δf	15 kHz	7.5 kHz	15 kHz
Sampling frequency, f_s	1.92 MHz	1.92 MHz	1.92 MHz
FFT size, N	128	128	64
CP length,N_{CP}	32 samples	32 samples	16 samples
Modulation type	BPSK	BPSK	BPSK
Channel model	AWGN	AWGN	AWGN

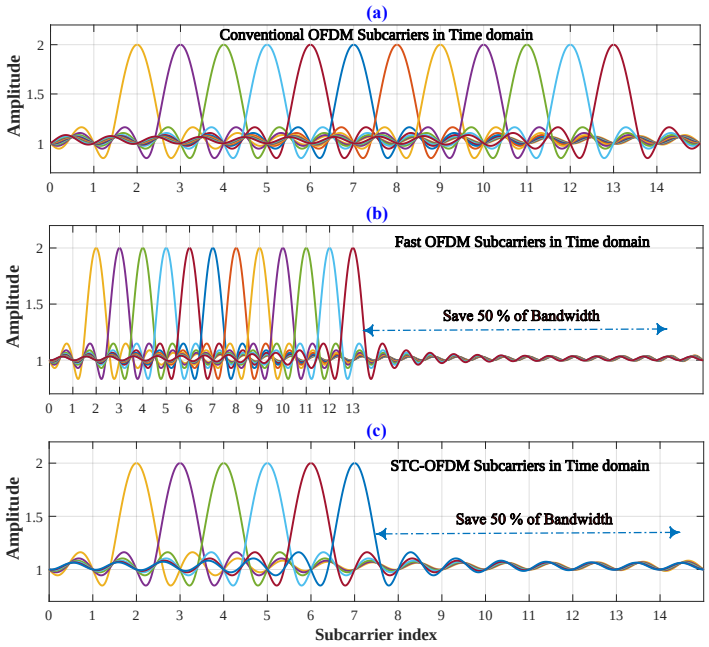


Figure 6: Sub-carrier allocation schemes for: (a) Typical OFDM, (b) Fast OFDM [12] and (c) STC-OFDM System [33].

In Figure 6, the sub-carriers for the traditional OFDM technique, Fast-OFDM, and the STC-OFDM are compared. Figure 6(a) shows a standard OFDM spectrum with orthogonally packed sub-carriers. The number of sub-carriers for Fast-OFDM is the same as typical OFDM, but the space between the sub-carriers is decreased to half as compared to typical OFDM as depicted in Figure 6(b). STC-OFDM reduces the number of sub-carriers by half while keeping the sub-carrier spacing, as it is clear in Figure 6(c). As a consequence, the bandwidth in Fast-OFDM and STC-OFDM is reduced by half, and the conserved bandwidth might be utilized to power more devices.

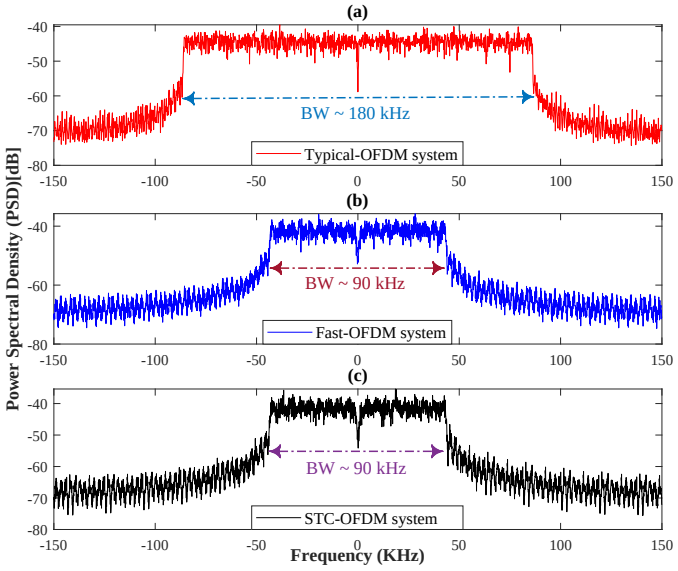


Figure 7: PSD for: (a) Typical OFDM, (b) Fast OFDM [12] and (c) STC-OFDM System [33].

Figure 7 shows the spectrum of STC-OFDM, Fast-OFDM, and typical OFDM. The first spectrum is for conventional OFDM and offers a bandwidth of 180 kHz; the second is for Fast-OFDM and compresses a bandwidth to 90 kHz; and the third is for STC-OFDM and also has a bandwidth of 90 kHz. The Fast-OFDM can transmit the same amount of data with half the bandwidth as compared to typical OFDM by decreasing the sub-carriers spacing to half, while the STC-OFDM sends the same amount of data with half the bandwidth as compared to typical OFDM by decreasing the number of sub-carriers to half, as it is seen in Figure 7. This means that they can save 50% of the bandwidth.

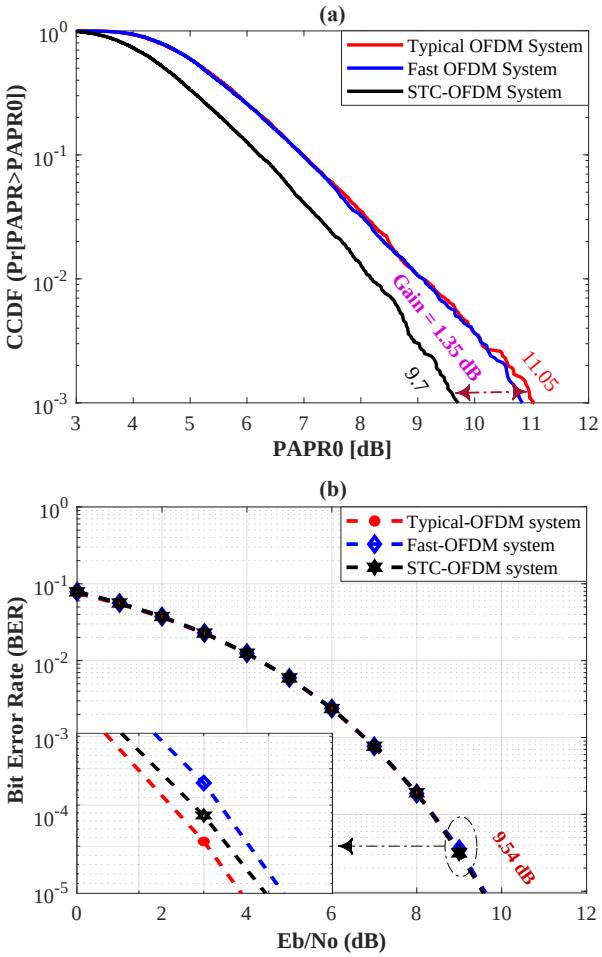


Figure 8: Performance comparison between Typical OFDM, Fast OFDM [12] and STC-OFDM System [33] based on (a) PAPR and (b) BER.

The STC-OFDM system uses half the bandwidth of the typical OFDM system to transmit the same amount of data. As a result, as illustrated in Figure 7(c), 50% of the bandwidth will be available for use to transmit the same amount of data. Furthermore, the STC-OFDM system increases system performance by decreasing the PAPR issue and avoiding BER deterioration. Figs. 8(a) and 8(b) indicate that the PAPR gain is 1.35 dB utilizing the STC-OFDM system with no BER deterioration. As for the Fast-OFDM, it saves half the bandwidth (Figure 7(b)) and prevents the BER degradation (Figure 8(b)), but it does not reduce PAPR, as seen in Figure 8(a).

From Figure 9 and Figure 10, it is clear that the transmitted data can increase to double using the same bandwidth. Figure 9 shows the transmitted picture, received picture, and PSD for the conventional OFDM system, while Figure 10 displays the transmitted pictures, received pictures, and PSD for the proposed scheme. The suggested approach concurrently transmits data from two sources (in our example, two sensors) with the same bandwidth as a standard OFDM system. Figure 9(c) and Figure 10(c)

illustrate that the traditional OFDM technique and the suggested method have the same bandwidth, but the suggested approach has the benefit of being able to transfer twice as much information as the conventional OFDM technique, as seen in Figs. 9 and 10.

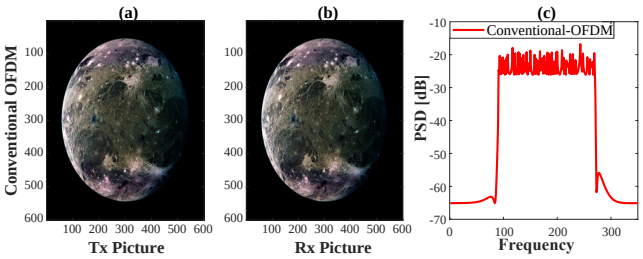


Figure 9: Send and receive a picture using a conventional OFDM system.

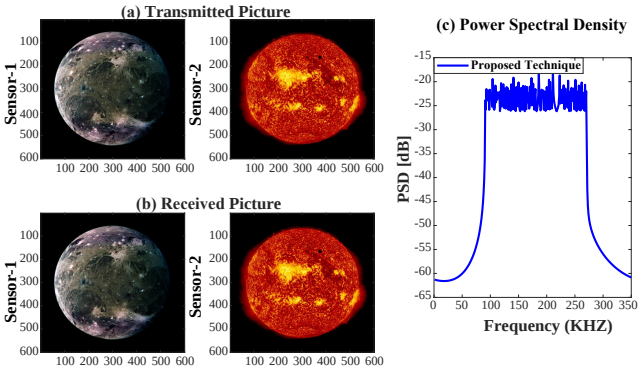


Figure 10: Send and receive two pictures using the proposed technique.

The time domain of the transmitted signal is measured after IFFT in the transmitter side for both the traditional OFDM system and the STC-OFDM method. When compared to a conventional OFDM system, the STC-OFDM scheme reduces OFDM symbol time by half for the same amount of data, as displayed in Figure 11(a). The transmitted information of a traditional OFDM system with 128 samples is transferred in 64 samples utilizing the STC-OFDM method, as it is clear in Figure 11(a). STC-OFDM is found to have similar performance to conventional OFDM, but with a 50% bandwidth savings. The number of IFFT is reduced by half when using the STC-OFDM technique and the PAPR is reduced as well. When employing the STC-OFDM technique, the PAPR improves by 1.58 dB, where the PAPR of typical OFDM is 11.35 and the PAPR of STC-OFDM is 9.77, as it is depicted in Figure 11(b). In addition to the improvement in PAPR, the STC-OFDM approach ensures that the BER of traditional OFDM and STC-OFDM is the same (i.e. the degradation in BER =0), as shown in Figure 11(c).

The STC-OFDM approach can save 50% of bandwidth, which means there is 50% of capacity that is not being used. As shown in Figure 4, the suggested method employs two STC techniques to use the entire bandwidth and double the amount of data transferred. Figure 12(a) shows that the suggested method has the same OFDM symbol time as a typical OFDM system, but it can send twice as much data, as displayed in Figure 10. The suggested approach offers the same PAPR as traditional OFDM and the same BER performance. The PAPR of a typical OFDM system and the suggested approach are shown in Figure 12(b), and it is evident that they are identical. Despite the fact that the suggested approach can transmit twice as much data as the typical OFDM system, it has the same BER performance, as illustrated in Figure 12(c).

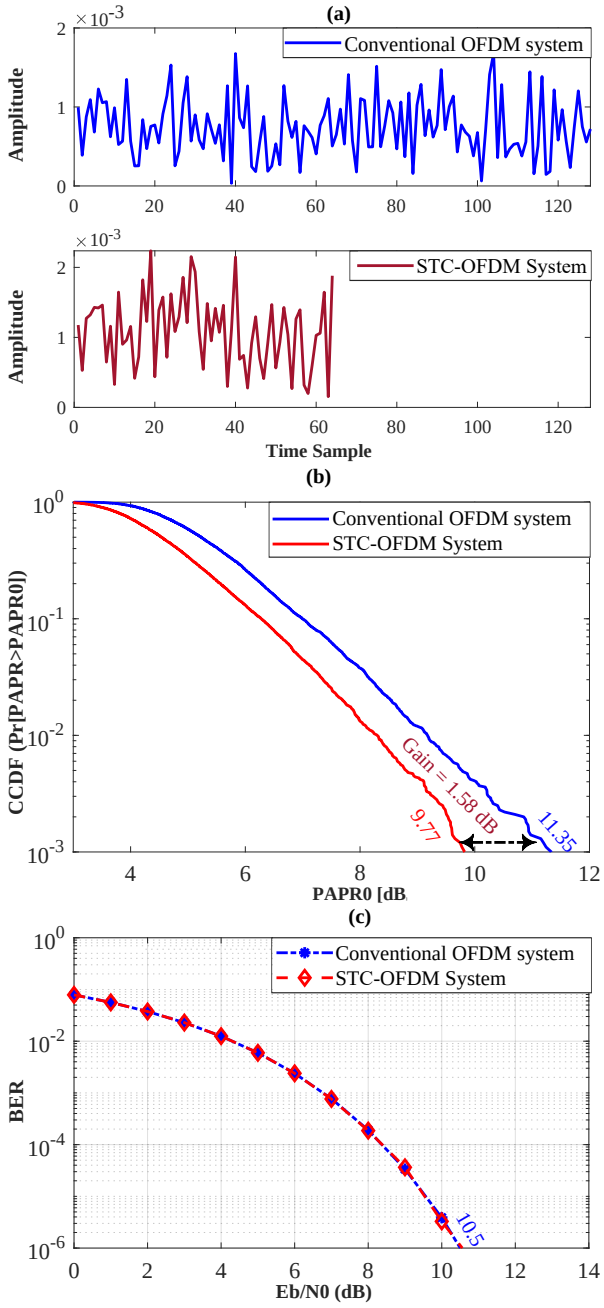


Figure 11: Performance comparison between typical OFDM and STC-OFDM technique [33] based on: (a) Time domain of Tx signal, (b) PAPR and (c) BER.

The μ-law companding approach reduces the high PAPR by enlarging small signals [35]. And therefore, it will be used with the proposed method to enhance system performance and reduce PAPR. Figure 13 shows the PAPR and BER for the suggested method using the $\mu-$law companding technique. When the μ parameter is set to one ($\mu = 1$), the PAPR improves by 3.22 dB with nearly no BER deterioration (BER degradation ≈ 0), as seen in Figure 13(a) and (b). Increasing the value of the μ parameter leads to decrease the PAPR and increase the enhancement in PAPR, as displayed in Figure 13(a). However, this enhancement will be at the expense of degradation in BER, as it is clear in Figure 13(b). Consequently, there is a trade-off between the enhancement in PAPR and the deterioration in BER. Table 5 summarizes the performance comparison of the proposed approach using $\mu-$law Companding and traditional OFDM based on PAPR and BER.

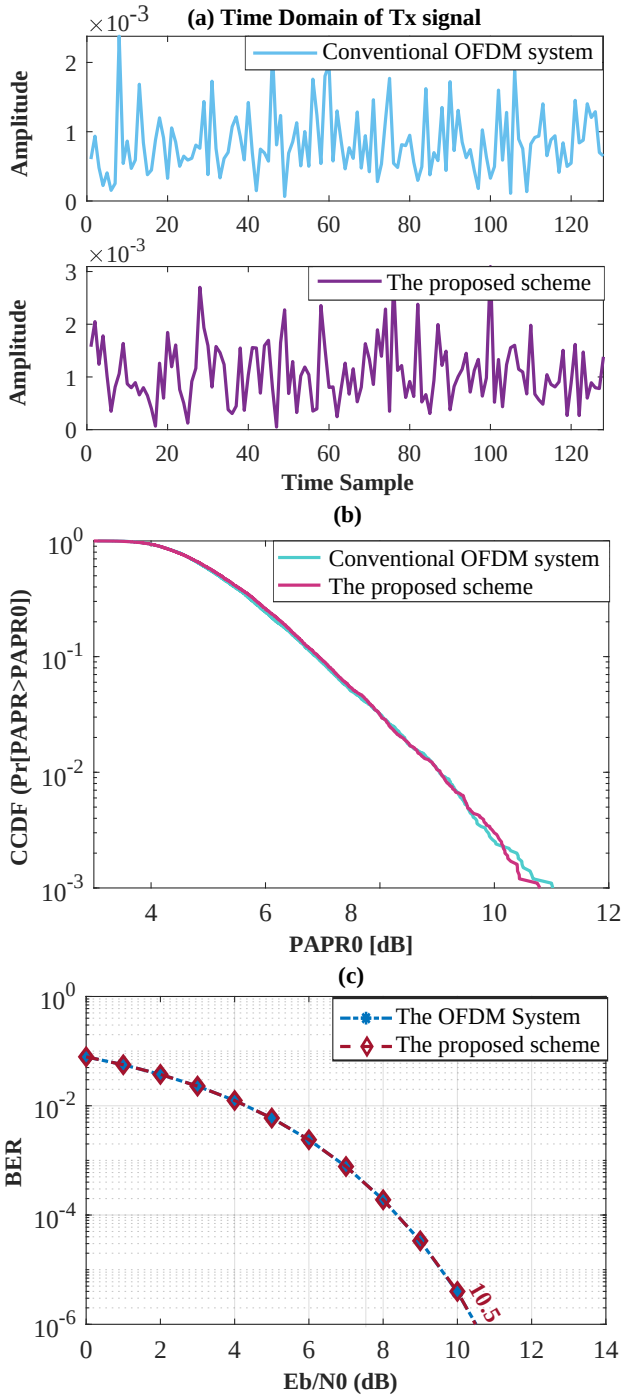


Figure 12: Performance comparison between conventional OFDM and the proposed method based on: (a) Time domain of Tx signal, (b) PAPR and (c) BER.

Table 5: PAPR improvement and BER deterioration for the suggested method using the Mu-Law technique.

μ−law	Improvement in PAPR	SNR at BER 10^{-6} degradation
μ = 1	11.17 - 7.95 = 3.22 dB	10.5 - 10.5 = 0 dB
μ = 4	11.17 - 5.1 = 6.07 dB	11.4 - 10.5 = 0.9 dB
μ = 10	11.17 - 3.28 = 7.89 dB	12.4 - 10.5 = 1.9 dB
μ = 100	11.17 - 2.17 = 9 dB	17.5 - 10.5 = 7 dB

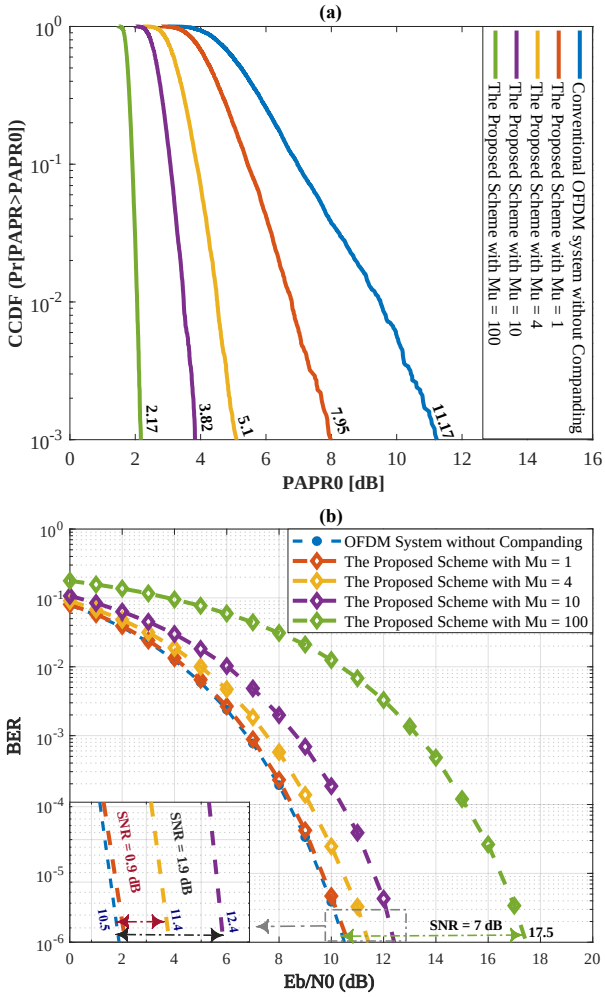


Figure 13: A typical OFDM system against the suggested method using μ-Law companding techniques: (a) PAPR and (b) BER.

7 Algorithms and Analysis of Complexity

This section explains the algorithms that utilized for the STC-OFDM and the proposed method. Algorithm 1 is designed to address the STC scheme and its inverse STC^{-1}, while Algorithm 2 represents the transceiver of the suggested method (Transmitter and Receiver). Table 6 represents a comparison between the computational complexity of the proposed method using μ−law companding technique and the other techniques. The following assumptions are applied to calculate complexity: i) the complexity of subtraction equals the complexity of addition and ii) the complexity of division equals the complexity of multiplication. In the conventional OFDM system, it requires a total of $(3N * \log_2 N - N)$ additions and $(2N * \log_2 N - 2N)$ multiplications [36]. The Fast OFDM system has the same sub-carriers compared to the conventional OFDM system. Consequently, the Fast OFDM system has the same complexity as the conventional OFDM system. The STC-OFDM system has half the sub-carriers as compared to the conventional OFDM and Fast OFDM systems. Therefore, the complexity of FFT in the STC-OFDM system requires a total of $(\frac{3}{2}N \log_2 \frac{N}{2} - \frac{N}{2})$ additions and $(N \log_2 \frac{N}{2} - N)$ multiplications. To exploit the whole bandwidth, the STC approach is implemented twice in the proposed system. As a result, the complexity of the suggested approach requires a total of $(3N * \log_2 N - N)$ additions and $(2N * \log_2 N - 2N)$ multiplications. For the complexity of μ−law technique, it needs N multiplications and $4N$ addition [37]. And therefore, the complexity of proposed scheme using μ−law requires a total

of $(3N * \log_2 N + 3N)$ additions and $(2N * \log_2 N - N)$ multiplications. It is clear from Table 6 that the complexity of the proposed method using $\mu-$law technique is nearly the same as the conventional OFDM. Consequently, this indicates that the proposed method does not cause an increase in the complexity of the system.

Table 6: Computational Complexity Analysis.

Techniques	No. Multiplications	No. Additions
Conventional OFDM [36]	2N Log_2N - 2N	3N Log_2N - N
Fast OFDM	2N Log_2N - 2N	3N Log_2N - N
STC-OFDM [36]	N $\text{Log}_2 \frac{N}{2}$ - N	$\frac{3}{2}$N $\text{Log}_2 \frac{N}{2}$ - $\frac{N}{2}$
The proposed scheme	2N Log_2N - 2N	3N Log_2N - N
The proposed scheme using μ-law [37]	2N Log_2N - N	3N Log_2N + 3N

Algorithm 1: : STC & STC^{-1} functions

At the Transmitter STC function

D= $reshape(Data, length(D)/2, 2)$; → Converting generated data from serial to parallel (Polar Matrix form)
W= $hadamard(n)$;→ Generate walsh code & n = 2
SpreadData = $D * W$; → Spreading Data
CombData= $comb(SpreadData)$; → Combining Data
Norm= $CombineData/2$;
$f(X)$ = Norm(:,1) + j*Norm(:,2);

At the Receiver STC^{-1} function

W= $hadamard(n)$;→ Generate walsh code & n = 2
R_x = Received Data in complex form ;
RealR_x = real (R_x); → Peal Part
ImagR_x = imag (R_x); → Imaginary part
R_y = [$RealR_x$; $ImagR_x$];
Despread_1 = R_y*W(1,:); → De-spreading Data
Despread_2 = R_y*W(2,:); → De-spreading Data
for $k = 1 : length(R_y)$ **do**
    CombData_1 = sum (Despread_1(k,:)) → Combining Data
    CombData_2 = sum (Despread_2(k,:)) → Combining Data
end for
Y = [$CombData_1$; $CombData_2$];
$f^{-1}(X)$ = (Y+1)/2;

Algorithm 2: : The Transceiver of Proposed Technique

Error = zeros $(1, length(EbN0dB))$;
nloop = 100; → Number of simulation loops
nsym = 1000; → Number of OFDM symbols for one loop
EbN0dB = 0:12; → Bit to noise ratio (Eb/N0)
for $i = 1 : nloop$ **do**
    D_1 = Generate data for first source
    D_2 = Generate data for second source
    X_1 = STC(D_1) → Apply STC Tech on the 1^{st} source
    X_2 = STC(D_1) → Apply STC Tech on the 2^{nd} source
    x_1 = **IFFT**(X_1); → Convert to time domain
    x_2 = **IFFT**(X_2); → Convert to time domain
    $y_1 = x_1 + cp_1$; → Add cyclic prefix
    $y_2 = x_2 + cp_2$; → Add cyclic prefix
    y = [y_1; y_2]; → Data of the two sources
    PAPR = zeros $(1, N)$; → N is the number of IFFT,
    for $k = 1 : nsym$ **do**
        Peak-power = **max**$(|y|^2)$;
        Avg-power = **mean**$(|y|^2)$;
        PAPR(k) = 10×$\log_{10}$(Peak-power/Avg-power);
    end for
    for $q = 1 : length(EsN0dB)$ **do**
        $r_x = y$+ noise; → Received under AWGN channel
        r_{cp1} = RemoveCP (r_{x1}); →Remove CP of the 1^{st} source
        r_{cp2} =RemoveCP (r_{x2}); →Remove CP of the 2^{nd} source
        R_{y1} =**FFT**(r_{cp1}); → Convert to frequency domain
        R_{y2} =**FFT**(r_{cp2}); → Convert to frequency domain
        R_{y1} =$\textbf{STC}^{-1}(r_{cp1})$; → Apply the inverse STC Tech to 1^{st} source
        R_{y2} =$\textbf{STC}^{-1}(r_{cp2})$; →Apply the inverse STC Tech to 2^{nd} source
        R_y = [R_{y1}; R_{y2}]; → Data of the two sources
        for $m = 1 : length(R_y)$ **do**
            if $R_y(m) > 0.5$ **then**
                $R_y(m) = 1$
            else
                $R_y(m) = 0$
            end if
        end for
        Error = BER (y , Output); → BER calculation
    end for
    D_1 = $R_y(1 : length(R_y/2))$; → Received Data of the 1^{st} source
    D_2 = $R_y(length(R_y/2) + 1 : end)$; → Received Data of the 2^{nd} source
end for

8 Conclusion

The major issue of low power wide area networks (LPWAN) is supporting a large number of devices while employing a limited spectrum of resources. This difficulty can be solved by employing narrow-band transmissions using NB-IoT. As a result, exponentially connected sensor nodes may be combined with additional benefits like better SNR and expanded coverage. Nonetheless, as the need for IoT services grows, more devices must be connected. The total number of linked devices is restricted because of the limited spectrum resources. The simulation results demonstrated that the STC-OFDM technique has the same performance as the standard OFDM system while conserving 50% of bandwidth, using half the number of sub-carriers to transmit the same data as the conventional OFDM system. According to simulation studies, the STC-OFDM scheme decreases OFDM symbol time by half for the same amount of data when compared to the standard OFDM system. When employing the STC-OFDM approach, the number of IFFT is decreased by half, as is the PAPR, where the PAPR improved by 1.58 dB with zero degradation in BER and nearly the same complexity.

In this study, the proposed method was employed to exploit the unused bandwidth in order to double the number of connected devices without requiring more bandwidth while still maintaining the system's performance. However, the suggested approach had the same PAPR and BER performance as the conventional OFDM system. When comparing STC-OFDM with Fast-OFDM, the two techniques had the same performance in sending the same amount of data with half the bandwidth (50% of the bandwidth), as compared to the typical OFDM. Additionally, they had the same BER compared to the typical OFDM. The STC-OFDM, however, outperformed the Fast-OFDM by 1.35 dB in terms of the PAPR problem. To improve the system performance in the proposed method and lower PAPR, the μ-law companding technique was combined with the proposed scheme. The μ value was carefully modified to get a good improvement in PAPR with no deterioration in BER. Based on the simulation results, the μ parameter was set to one ($\mu = 1$) to improve the PAPR by 3.22 dB with almost no BER deterioration (BER degradation ≈ 0). In addition to the benefits of reduced PAPR, improved system performance, and double the number of connected devices, the proposed method using the $\mu-$Law technique had roughly the same complexity as the conventional OFDM. Finally, raising the value of

the μ parameter reduced PAPR and increased the improvement in PAPR. However, this improvement came at the cost of BER deterioration. As a result, there was a trade-off between the enhancement in PAPR and the deterioration in BER.

References

[1] Forbes, "2017 Roundup Of Internet Of Things Forecasts," https://www.forbes.com/sites/louiscolumbus/2017/12/10/2017-roundup-of-internet-of-things-forecasts/?sh=39f80e1b1480, 2017.

[2] M. SYSTEMS, "Global IoT market to grow to 24.1 billion devices in 2030, generating $1.5 trillion annual revenue," https://transformainsights.com/news/iot-market-24-billion-usd15-trillion-revenue-2030, 2020.

[3] E. Rastogi, N. Saxena, A. Roy, D. R. Shin, "Narrowband Internet of Things: A comprehensive study," Computer networks, **173**, 107209, 2020.

[4] S. Liu, L. Xiao, Z. Han, Y. Tang, "Eliminating NB-IoT interference to LTE system: A sparse machine learning-based approach," IEEE Internet of Things Journal, **6**(4), 6919–6932, 2019.

[5] Y. Niu, C. Gao, Y. Li, L. Su, D. Jin, A. V. Vasilakos, "Exploiting device-to-device communications in joint scheduling of access and backhaul for mmWave small cells," IEEE Journal on Selected Areas in Communications, **33**(10), 2052–2069, 2015.

[6] J. Chen, K. Hu, Q. Wang, Y. Sun, Z. Shi, S. He, "Narrowband internet of things: Implementations and applications," IEEE Internet of Things Journal, **4**(6), 2309–2314, 2017.

[7] H. Kroll, M. Korb, B. Weber, S. Willi, Q. Huang, "Maximum-likelihood detection for energy-efficient timing acquisition in NB-IoT," in 2017 IEEE Wireless Communications and Networking Conference Workshops (WCNCW), 1–5, IEEE, 2017.

[8] J. Liu, Q. Mu, L. Liu, L. Chen, "Investigation about the paging resource allocation in NB-IoT," in 2017 20th International Symposium on Wireless Personal Multimedia Communications (WPMC), 320–324, IEEE, 2017.

[9] H. Bello, X. Jian, Y. Wei, M. Chen, "Energy-delay evaluation and optimization for NB-IoT PSM with periodic uplink reporting," IEEE Access, **7**, 3074–3081, 2018.

[10] A. Adhikary, X. Lin, Y.-P. E. Wang, "Performance evaluation of NB-IoT coverage," in 2016 IEEE 84th Vehicular Technology Conference (VTC-Fall), 1–5, IEEE, 2016.

[11] P. Andres-Maldonado, P. Ameigeiras, J. Prados-Garzon, J. J. Ramos-Munoz, J. Navarro-Ortiz, J. M. Lopez-Soler, "Analytic analysis of narrowband IoT coverage enhancement approaches," in 2018 Global internet of things summit (GIoTS), 1–6, IEEE, 2018.

[12] T. Xu, I. Darwazeh, "Non-orthogonal narrowband Internet of Things: A design for saving bandwidth and doubling the number of connected devices," IEEE Internet of Things Journal, **5**(3), 2120–2129, 2018.

[13] M. Mahbub, "NB-IoT: applications and future prospects in perspective of Bangladesh," International Journal of Information Technology, **12**(4), 1183–1193, 2020.

[14] R. S. Sinha, Y. Wei, S.-H. Hwang, "A survey on LPWA technology: LoRa and NB-IoT," Ict Express, **3**(1), 14–21, 2017.

[15] Z. Zhang, "ZCNET: Achieving high capacity in low power wide area networks," in 2020 IEEE 17th International Conference on Mobile Ad Hoc and Sensor Systems (MASS), 702–710, IEEE, 2020.

[16] Y. Chen, Y. A. Sambo, O. Onireti, M. A. Imran, "A Survey on LPWAN-5G Integration: Main Challenges and Potential Solutions," IEEE Access, **10**, 32132–32149, 2022.

[17] R. Marini, K. Mikhaylov, G. Pasolini, C. Buratti, "Low-Power Wide-Area Networks: Comparison of LoRaWAN and NB-IoT Performance," IEEE Internet of Things Journal, 2022.

[18] A. Tusha, S. Doğan, H. Arslan, "IQI mitigation for narrowband IoT systems with OFDM-IM," IEEE Access, **6**, 44626–44634, 2018.

[19] G. A. Akpakwu, B. J. Silva, G. P. Hancke, A. M. Abu-Mahfouz, "A survey on 5G networks for the Internet of Things: Communication technologies and challenges," IEEE access, **6**, 3619–3647, 2017.

[20] B. Qolomany, A. Al-Fuqaha, A. Gupta, D. Benhaddou, S. Alwajidi, J. Qadir, A. C. Fong, "Leveraging machine learning and big data for smart buildings: A comprehensive survey," IEEE Access, **7**, 90316–90356, 2019.

[21] V. Javidroozi, H. Shah, G. Feldman, "Urban computing and smart cities: Towards changing city processes by applying enterprise systems integration practices," IEEE Access, **7**, 108023–108034, 2019.

[22] R. Du, P. Santi, M. Xiao, A. V. Vasilakos, C. Fischione, "The sensable city: A survey on the deployment and management for smart city monitoring," IEEE Communications Surveys & Tutorials, **21**(2), 1533–1560, 2018.

[23] L. Wan, Z. Zhang, J. Wang, "Demonstrability of Narrowband Internet of Things technology in advanced metering infrastructure," EURASIP Journal on Wireless Communications and Networking, **2019**(1), 1–12, 2019.

[24] V. Nair, R. Litjens, H. Zhang, "Optimisation of NB-IoT deployment for smart energy distribution networks," Eurasip journal on wireless communications and networking, **2019**(1), 1–15, 2019.

[25] W. Zhong, R. Yu, S. Xie, Y. Zhang, D. H. Tsang, "Software defined networking for flexible and green energy internet," IEEE Communications Magazine, **54**(12), 68–75, 2016.

[26] K. Wang, Y. Shao, L. Shu, C. Zhu, Y. Zhang, "Mobile big data fault-tolerant processing for ehealth networks," IEEE Network, **30**(1), 36–42, 2016.

[27] Y. D. Beyene, R. Jantti, O. Tirkkonen, K. Ruttik, S. Iraji, A. Larmo, T. Tirronen, J. Torsner, "NB-IoT technology overview and experience from cloud-RAN implementation," IEEE wireless communications, **24**(3), 26–32, 2017.

[28] J. Xu, J. Yao, L. Wang, Z. Ming, K. Wu, L. Chen, "Narrowband internet of things: Evolutions, technologies, and open issues," IEEE Internet of Things Journal, **5**(3), 1449–1462, 2017.

[29] S. K. Sharma, X. Wang, "Toward massive machine type communications in ultra-dense cellular IoT networks: Current issues and machine learning-assisted solutions," IEEE Communications Surveys & Tutorials, **22**(1), 426–471, 2019.

[30] Y. D. Beyene, R. Jantti, K. Ruttik, S. Iraji, "On the performance of narrowband Internet of Things (NB-IoT)," in 2017 ieee wireless communications and networking conference (wcnc), 1–6, IEEE, 2017.

[31] M. S. El-Bakry, H. A. El-Shenawy, A. E.-H. A. Ammar, "A time inversion and symbol time compression (TI-STC) scheme for ICI cancellation in high mobility OFDM systems," in 2017 Japan-Africa Conference on Electronics, Communications and Computers (JAC-ECC), 82–85, IEEE, 2017.

[32] M. S. Elbakry, A. Mohammed, T. Ismail, "Throughput improvement and PAPR reduction for OFDM-based VLC systems using an integrated STC-IMADJS technique," Optical and Quantum Electronics, **54**(7), 1–17, 2022.

[33] M. S. El-Bakry, H. A. El-Shenawy, A. E.-H. A. Ammar, "A symbol time compression for ICI reduction in high mobility OFDM systems," in 2017 29th International Conference on Microelectronics (ICM), 1–4, IEEE, 2017.

[34] T. Xu, I. Darwazeh, "Transmission Experiment of Bandwidth Compressed Carrier Aggregation in a Realistic Fading Channel," IEEE Transactions on Vehicular Technology, **66**(5), 4087–4097, 2017, doi:10.1109/TVT.2016.2607523.

[35] A. Mohammed, M. Shehata, A. Nassar, H. Mostafa, "Performance Comparison of Companding-Based PAPR Suppression Techniques in OFDM Systems," in 2019 8th International Conference on Modern Circuits and Systems Technologies (MOCAST), 1–4, IEEE, 2019.

[36] M. Y. I. Afifi, M. S. Elbakry, E.-S. S. Soliman, A. Ammar, "An efficient technique for out-of-band power reduction for the eliminated CP-STC-shaped system for 5G requirements," International Journal of Electrical and Computer Engineering, **10**(5), 5306, 2020.

[37] A. Mohammed, T. Ismail, A. Nassar, H. Mostafa, "A novel companding technique to reduce high peak to average power ratio in OFDM systems," IEEE Access, **9**, 35217–35228, 2021.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 8, No. 4, 11-17 (2023)

www.astesj.com

Special Issue on Innovation in Computing, Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Design and Manufacturing of Soft Grippers for Robotics by Injection Molding Technology

Helmy Dewanto Bryantono[1], Melsiani Rosdiani Fillipin Saduk[1,2], Jiaqi Hong[1], Meng-Hsun Tsai[1], Shi-Chang Tseng[1,*]

[1]*Mechanical Engineering Department, National Yunlin University of Science and Technology, Douliu, 640301, Taiwan*

[2]*Mechanical Engineering Department, State Polytechnic of Kupang, Kupang, 85361, Indonesia*

ARTICLE INFO

Article History:
Received: 25 March, 2023
Accepted: 22 June, 2023
Online: 21 July, 2023

Keywords:
Deformation
Injection molding
Soft gripper
Soft Robot

ABSTRACT

Soft robots have softer, more flexible, and more compliant characteristics than traditional rigid robots. These qualities encourage more secure relationships between people and machines. Nevertheless, traditional robots continue to rule the commercial sector. Due to the high cost of gripper production, soft robots are very far from being commercially feasible. This research focuses on fabricating a soft robotic gripper with the potential for mass production using injection molding technology. The material used for manufacture is Thermoplastic Elastomer (TPE). This study gives an injection molding optimization strategy by using Moldex 3D simulation to minimize production time for soft grippers. Furthermore, using an Ansys workbench, this study simulated soft gripper deflections with variable pressures by finite element analysis and then compared it with the actual experiment. The simulation results of TPE warpage volume shrinkage are 11.969% and 11.96% in the molding experiment. Thus, the shrinkage and warpage for the simulation and actual experiment are similar. According to the simulation outcome, the success of TPE hollow injection molding facilitates soft gripper creation. The maximum pressure used in the FEM simulation of the bending experiment was achieved at the pressure of 50 kPa with 152.02 mm of deformation and compared to the experimental data, 145,03 mm. This error is less than 5%. Finally, a better soft gripper design was achieved by Ansys simulation, and the soft gripper appears to be ready for mass-produced by TPE injection molding.

1. Introduction

Robots have typically been made of stiff materials in order to generate accurate, predictable movements with minor deformation. As a result, rigid robots can accomplish activities with great precision and accuracy in controlled environments. However, nature regularly utilizes compliant and flexible structures in unstructured conditions to generate movement and manipulation. Therefore, because of such organic designs, soft robotics is becoming increasingly popular [1–6].

Soft grippers are needed for soft robots to generate motion and force. Soft grippers can develop the flexibility and adaptation necessary to construct extraordinarily adaptive robotics for soft interactions in the absence of an inflexible skeleton in soft robots. Because of its versatility in filling the gap left by typical rigid robots, the soft gripper is gaining popularity. Therefore, one area where soft robotics can have a substantial impact is the development of soft grippers. Soft grippers are one kind of soft robotics that uses softness to provide highly compliant and adjustable grasping capability [7].

The utilization of rigid grippers designed to handle delicate objects is frequently problematic or limited. As one potential solution, a soft device that can passively adjust to the object and its surroundings could replace the gripper's role. As a result, numerous soft mechanical architectures have been proposed, and extensive research on soft robotics has been conducted [8–10]. Soft robotics, unlike rigid robotics, can adapt to changing environmental conditions and interact with people more securely and flexibly. As a result, soft robot design and fabrication have significantly progressed over the last ten years, and new actuator design and fabrication processes have been created and improved [11,12].

The development of novel materials and soft components is advancing in order to create lighter, simpler general grippers. Compliance has always been considered important in grasping. Shocks induced by contact with a firm gripper and a rigid object, if not adequately regulated, can injure the product or force it outside of the targeted route. Incorporating sensitive materials into robotic end effector grabbing portions is a basic but only partially successful alternative (for example, in the shape of basic rubber pads). Interaction between bodies, on the other hand, impedes their mobility [13–15]. Soft grippers use the responsiveness and elasticity of materials to build highly flexible robotic arms that allow for secure contact between devices and the surrounding. Therefore, when making soft grippers, it is vital to consider the materials and production processes [16–18].

Soft grippers, in contrast to standard robots, are often made from materials with Young's moduli equivalent to soft

*Corresponding Author: Shi-Chang Tseng, Douliu, *640301*, Taiwan, tsengsc@yuntech.edu.tw

www.astesj.com
https://dx.doi.org/10.25046/aj080402

biological materials such as muscles, tendons, and skin, typically about 1e-7GPa-1GPa. Compatibility and "softness" are produced using morphology and material characteristics in naturally pliable materials. However, soft materials have limitations regarding potential manufacture, non-linear response, modeling complexity, self-repair, and fatigue performance. Therefore, elastomers and rubbers, which can sustain reversibly (>100%) due to tremendous loads, are the most frequently utilized soft grip materials [16,19]. Thermoplastic elastomers have emerged as the most popular option for industrial applications in this study because of their low toxicity, robustness, and mechanical dampening characteristics and their simplicity of manufacture [20].

However, due to the high cost of gripper production, soft robots are very far from commercially feasible. When it comes to mass production, forming is the most cost-efficient and has the lowest unit cost of the three basic types of manufacturing: forming, machining (subtractive), and additive manufacturing [21].

In addition, the intricacy of the fabrication procedure may result in a more extended fabrication period that may last from hours to days, and the high expense of typical gripper production methods all contribute to this. Therefore, an efficient fabrication process is required to ensure acceptable mechanical performance while reducing fabrication time and complexity.

Thus, this research focuses on fabricating soft robotic grippers that have the potential for mass production. Consequently, this research concentrates on the material used for soft grippers and a unique process for manufacturing soft grippers with high throughput and mass production potential. For this reason, TPE hollow injection molding was chosen because it is a high-speed, automatic technique that can produce parts with a wide range of sizes and extremely complicated geometries in large quantities.

2. Theoretical Background

There are numerous mathematical models for viscosity accessible in Moldex3D. TPE is used in all injection molding simulations in this study. Moldex3D's suggested viscosity model for this material is the Modified Cross Model 3, often called the Cross-WLF model. In this model, the melt viscosity is dependent on the shear rate $\dot{\gamma}$, temperature T, and pressure p, according to:

$$\eta(\dot{\gamma},T,p)=\frac{\eta_0(T,p)}{1+\left(\frac{\eta_0(T,p)\dot{\gamma}}{\tau}\right)^{1-n}} \quad (1)$$

where η_0 (zero-shear rate viscosity), n (power-law index with a value between 0 and 1), τ^* (relaxation stress). The viscosity versus shear rate obtained from the Cross Model using the TPE material is depicted in Figure 1 for temperatures of 190°C, 215°C, and 240°C, respectively.

Table 1: Cross-WLF model data for TPE melt viscosity [22]

n	0.264907	-
Taus	153557	dyne/cm^2
B	4.30664e-06	g/(cm.sec)
Tb	10819	K
D	0	cm^2/dyne

The shear-dependent parameters are identical to those in the Modified Cross Model. Correlating low-temperature viscosity with the Cross-WLF model is often more accurate. The Cross-WLF model outperforms the Cross-Exp model at temperatures below Tg+100°C.

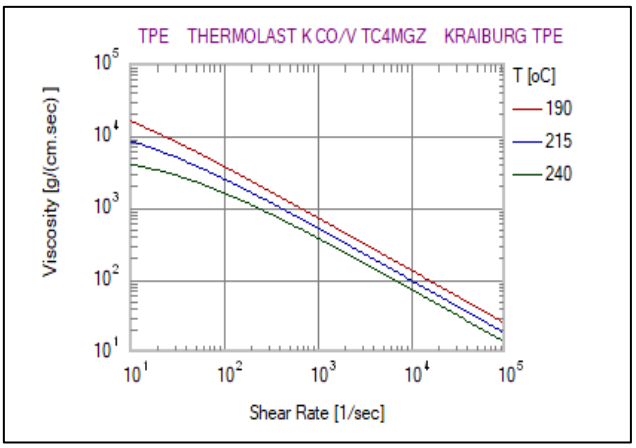


Figure 1: Viscosity versus shear rate for TPE

Modified Tait Model 2, used in this work and offered by Moldex 3D, can be used to explain the PVT model.

$$v(T,p)=v_0(T)\left[1-Cln\left(1+\frac{p}{B(T)}\right)\right]+v_t(T,p) \quad (2)$$

$v(T,p)$ is the specific volume (contrarywise proportional to density) at Temperature, T. Pressure, P, $v_0(T)$ is a specific volume at zero gauge pressure, C is a constant of 0.0894 and $B(T)$ accounts for the material pressure sensitivity, as shown in Figure 1. For the upper-temperature region $(T > T_t)$ and the lower-temperature region $(T \leq T_t)$ respectively. Figure 2 illustrates the specific volume dependency for TPE on Temperature and pressure.

Table 2: PVT data for TPE [22]

b1L [0.8415, 1....	1.03922	cc/g
b2L [0.0005017...	0.000713614	cc/(g.K)
b3L [5.13e+08, ...	9.5127e+08	dyne/cm^2
b4L [1e-06, 0.0...	0.00474979	1/K
b1S [0.8263, 1....	1.0367	cc/g
b2S [0.000352, ...	0.000548239	cc/(g.K)
b3S [5.2e+08, ...	1.01167e+09	dyne/cm^2
b4S [0.002665, ...	0.00302857	1/K
b5 [338.1, 493.1]	448.973	K
b6 [1.9e-09, 04...	2.75172e-08	cm^2.K/dyne
b7 [0, 0.0711]	0	cc/g
b8 [0, 0.41]	0	1/k

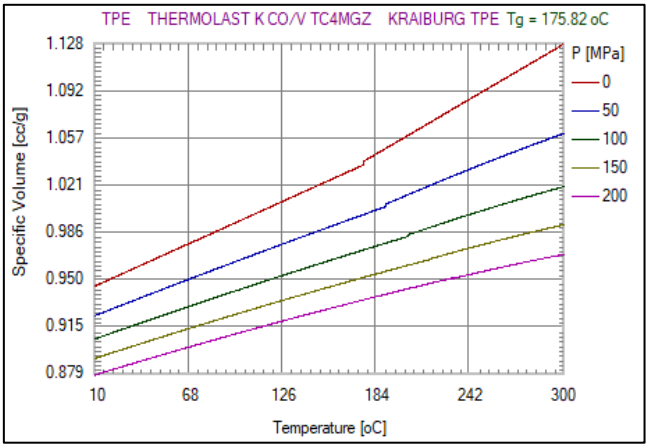


Figure 2: PVT diagram for TPE

Because these are equilibrium states, the PVT characteristics of polymers in the molten state may be appropriately described by the equation of state (EoS).

The generalized Newtonian fluid (GNF) theory predicts that the polymer melt will behave [23,24]. Because the GHS flow model depicts the flow within two plates that are close together, the width of the gap is considered to be substantially slighter than the other flow parameters. This notion concludes that local geometry has the greatest influence on flow at a single spot. Because velocity in the gap-wise orientation is ignored and pressure is simply an expression of planar coordinates, the lubrication approximation may be used. The following equations describe the melt flow of the Hele-Shaw polymer:

Continuity equation:

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \mathbf{v}) = 0 \quad (3)$$

Where ρ (density) and t (Time), and **v** (velocity vector). Equation (3 is a standard continuity equation.

Momentum equation:

$$\frac{\partial}{\partial t}(\rho \mathbf{v}) + \nabla \cdot (\rho \mathbf{v}\mathbf{v}) = -\nabla \mathrm{p} + \rho g + \nabla . \tau \quad (4)$$

Where g (gravity) and τ (shear stress).

Energy equation:

$$\frac{\partial}{\partial t}\left(\rho C_p T\right) + \nabla .\left(\rho C_p \mathbf{v} T - k \nabla \mathrm{T}\right) = \eta \dot{\gamma}^2 \quad (5)$$

where T (temperature field), C_p (heat capacity), k (thermal conductivity coefficient), η (viscosity of the fluid), and $\dot{\gamma}$ (shear rate).

The function used in the elasticity problem is the structure's total potential energy. In the first step, assume an estimated form for the solution, and the answer should be acceptable and meet the crucial internal compatibility and boundary restrictions. The optimal constant values are then obtained using the stationary potential energy theorem. The theory states that among all acceptable arrangements of a conservative system, the ones that satisfy the equations of equilibrium render the potential energy constant when allowable displacement fluctuations are taken into consideration., i.e., equilibrium configurations are defined when:

$$\frac{\partial \Pi_p}{\partial a_i} = 0 \quad (6)$$

where Π_p (potential) and a_i (ith dof).

$$\Pi_p = U + \Omega \quad (7)$$

where U (strain energy) and Ω (potential of the load system). In a stressed body, strain energy is provided by

$$U = \frac{1}{2}\int \{\varepsilon\}^T [E]\{\varepsilon\} dv \quad (8)$$

where $[E]$ (constant elastic matrix). It expresses the connection between stress $\{\sigma\}$ and strain $\{\varepsilon\}$.

The potential of the load system is

$$\Omega = -\{u^e\}^T\{F^e\} \quad (9)$$

The total potential of the system is,

$$\Pi_p = U + \Omega = \frac{1}{2}\int \{\varepsilon\}^T [E]\{\varepsilon\} dv - \{u^e\}^T\{F^e\} \quad (10)$$

A shape function, $[N]$, is used to relate the vector of displacements in the element, $\{u\}$, to the vector of nodal displacements, $\{u^e\}$.

After that, the strain vector may be calculated by partial differentiation of the displacement vector, giving

$$\{\varepsilon\} = \left\{\frac{du}{dx}\right\} = \left[\frac{dN}{dx}\right]\{u^e\} = [B]\{u^e\} \quad (11)$$

where $[B]$ is the strain-displacement matrix. Substituting (10) into (9),

$$\Pi_p = \frac{\{u^e\}^T}{2}\int [B]^T[E][B] dv \{u^e\} - \{u^e\}^T\{F^e\} \quad (12)$$

Applying the stationary potential energy theorem,

$$\frac{\partial \Pi_p}{\partial \{u^e\}} = 0 = \int [B]^T[E][B] dv \{u^e\} - \{F^e\} \quad (13)$$

$$\{F^e\} = \int [B]^T[E][B] dv \{u^e\} \quad (14)$$

And as $\{F^e\} = [K^e]\{u^e\}$:

$$[K^e] = \int [B]^T[E][B] dv \quad (15)$$

The soft gripper body in this research is made of a thermoplastic elastomer (TPE), a hyper elastic material. Because the material feature of a hyper elastic material is non-linear under external forces, the strain energy density function is extensively employed to describe the mechanical features of TPE material. In this paper, the Yeoh model proposes a non-linear association between material stress and strain. Yeoh thought that the strain tensor invariants I_2 had no effect on the strain energy and might be ignored entirely. Here is the simplified strain energy density function:

$$W = \sum_i^N C_{i0}(\overline{I_1} - 3)^i + \sum_{k=1}^N \frac{1}{d_k}(J-1)^{2k} \quad (16)$$

where N, C_{i0}, and d_k are material constants discovered during the material experiment and $J = 1$ for incompressible materials. The typical two-parameter form is

$$W = C_{10}(I_1 - 3)^1 + C_{20}(I_1 - 3)^2 \quad (17)$$

The 2nd Yeoh Parameter hyper-elastic material model was used to model the TPE material and fit its average stress-strain data.

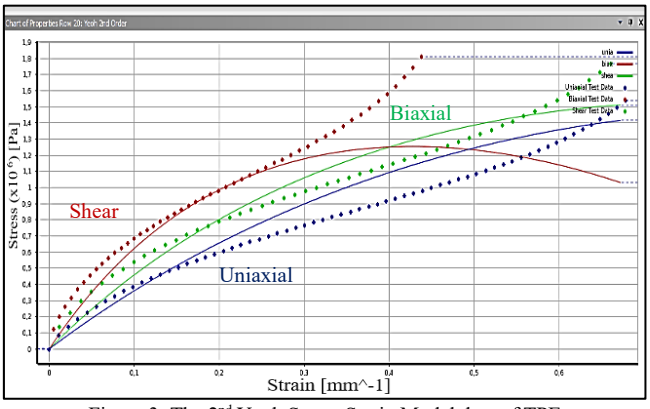


Figure 3: The 2nd Yeoh Stress-Strain Model data of TPE

Table 3: TPE Hyper-Elastic Material Model Constants

Model for TPE Hyper-Elastic Materials	Constant Material	Value (Pa)
The 2nd Yeoh Parameter	C10	6.6275+05
	C20	-62451
	Incompressibility Parameter D1	0
	Incompressibility Parameter D2	0

3. Design and Experimental of TPE Soft Gripper

3.1 Soft Gripper Design

A soft gripper form was created in this study, as illustrated in Figure 4. The TPE soft grippers' air chamber is enclosed within the gripper body. Once air pressure is applied and motion is formed, this air chamber enlarges to accomplish bending deformation. The dimension of the TPE soft gripper is 114 x 19 x 16.5 mm [1]. This design has been proven can accomplish pleasant bending deformation by previous studies [25].

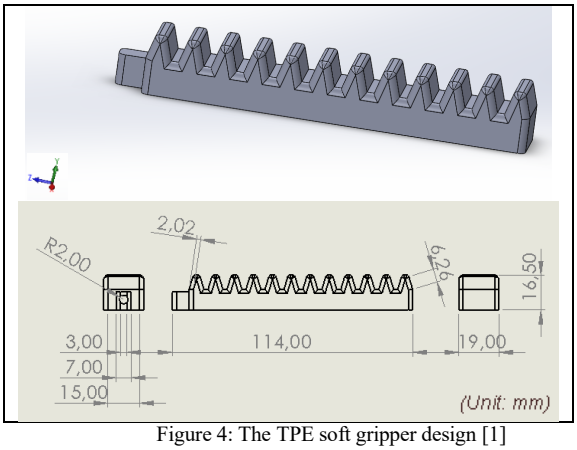


Figure 4: The TPE soft gripper design [1]

3.2 Soft Gripper Mold Design

The soft gripper's mold design was achieved after designing a soft gripper body, as shown in Figure 5.

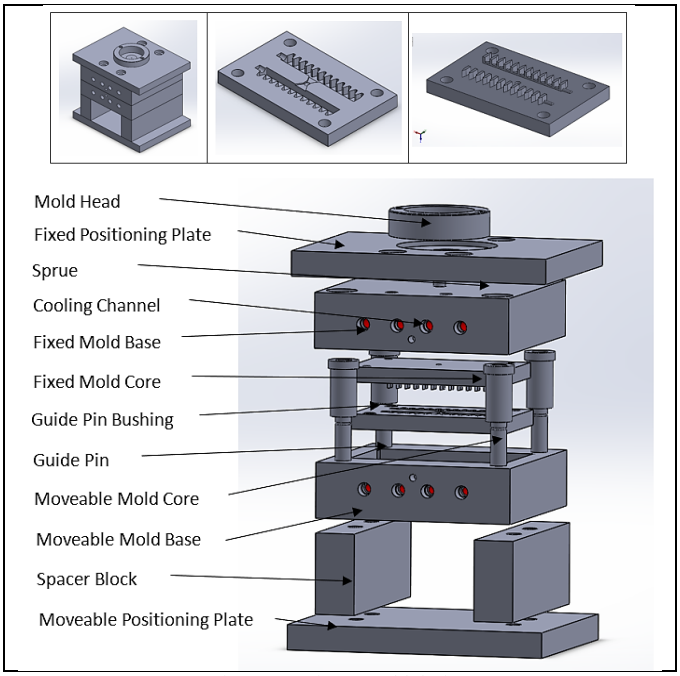
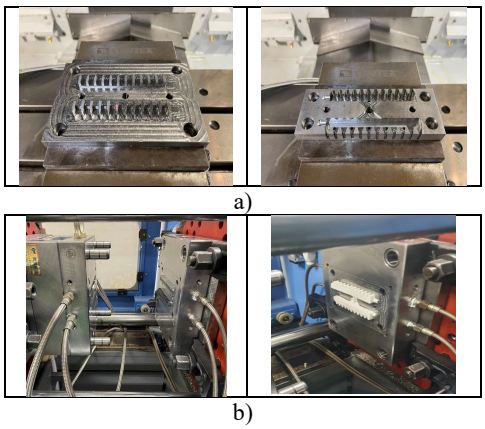


Figure 5: Gripper mold design

3.3 Mold Manufacturing

A High-Speed/Closed-Loop Hybrid Injection Molding Machine produces the soft gripper (AF Series). The soft gripper is meant to be flexible in this experiment, and TPE was employed. Figure 6 depicts the soft gripper mold used to create the soft gripper.



a)

b)

Figure 6: TPE Hollow Gripper a) Mold [1] and b) Injection Molding Machine

3.4 The Simulation of Soft Gripper

Simulations carried out in ANSYS Workbench using a Static Structural Analysis. The finite element diagram in Figure 7 illustrates and describes the processes necessary to model a soft gripper.

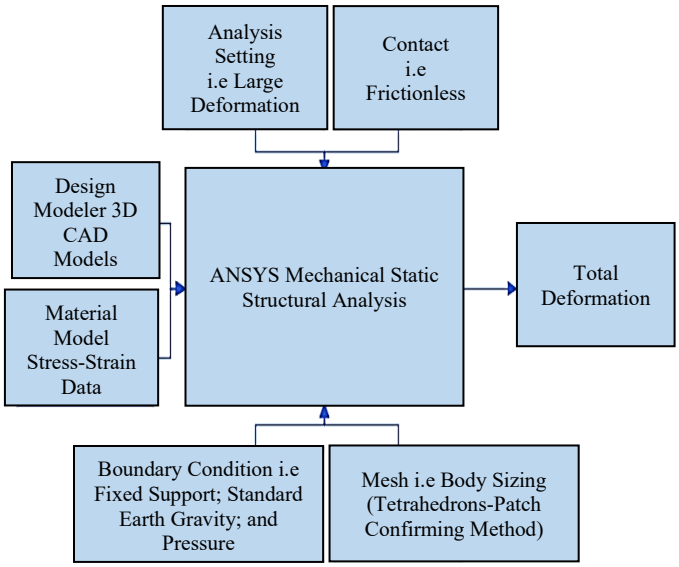


Figure 7: Diagram of finite element simulation

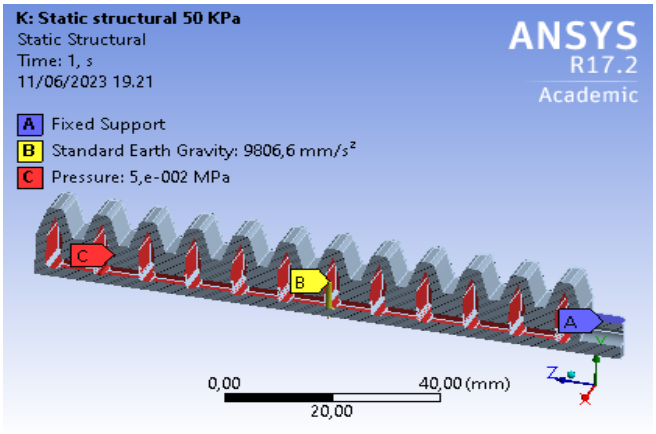


Figure 8: Simulation boundary condition diagram

TPE soft gripper uses tetrahedrons meshed, with the element size adjusted to 1 mm. The whole deformation is caused by pressure given to the TPE soft gripper, exhibiting the gripper is stretched at various phases of the pressure applied. Figure 8 displays a boundary condition simulation of the soft gripper.

Aside from the numerical simulations used (Ansys software), further unique and powerful numerical approaches for strain-stress analysis of isotropic or anisotropic media have recently been developed. The "Gaussian Quadrature" and "Bezier" techniques, among others, demonstrated more stability and accuracy than previous numerical approaches. Alternatively, these approaches may forecast soft gripper deflections with variable pressure [26,27].

3.5 Grasping Experiment

The grasping experiment is applied in this research to examine the condition of the gripping force. Figure 9 depicts the fixture design, which includes the length, pitch between the chambers, and air pressure. As a result, the plastic ball object is grasped well without any defect on the object, as shown in Figure 15.

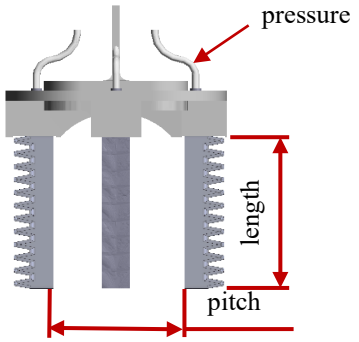


Figure 9: The fixture design: length, pitch, and the air pressure

4. Result and Discussion

4.1 Injection Molding Simulation

Moldex 3D has been used to simulate Injection Molding tests of TPE soft gripper. First, as shown in Table 4, the results were achieved by simulating the injection molding process. Following that, the filling phase simulation is seen in Figure 10.

Table 4: Soft Gripper Injection Molding Simulation

Parameter	Value
Melt Temperature (°C)	210
Cooling Channel Temperature (°C)	20
Mold Temperature (°C)	50
Maximum Injection Pressure (MPa)	30
Maximum Packing Pressure (MPa)	20
Cooling Flow Rate (cm^3/sec)	120
Filling Time (sec)	2
Cooling Time (sec)	30
Holding Time (sec)	2
Mold-Open Time (sec)	5
Eject Temperature (°C)	70

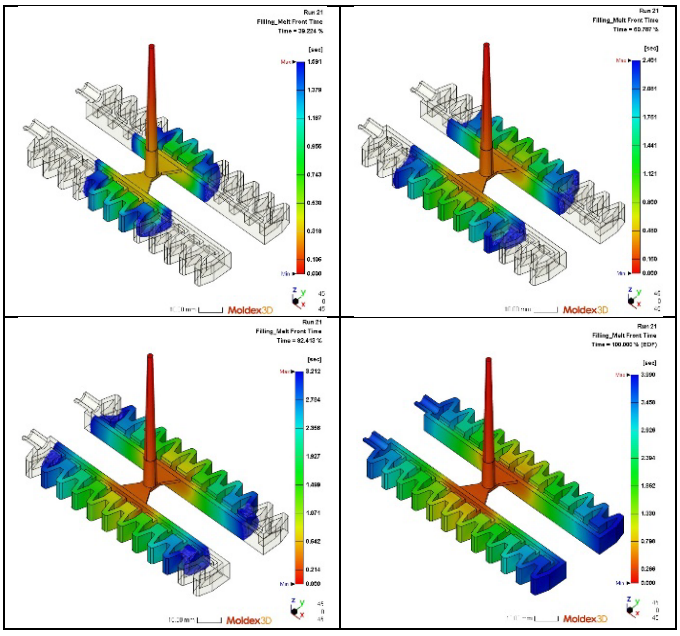


Figure 10: Filling phase simulation of the TPE 40%, 60%, 80%, and 100%

4.2 Soft Gripper Manufacturing

This manufacturing of the soft gripper is displayed in Figure 11. To prevent the TPE material from absorbing moisture and causing pores in the finished product. The material must be dried before being injected into the molding machine. After determining the injection volume, the finished product has shrinkage and slight overflow. Therefore, the mold temperature has to adjust to reduce the mold temperature and increase the packing pressure to prevent the overflow and shrinkage of the finished product. The finished product is displayed below in Figure 11.

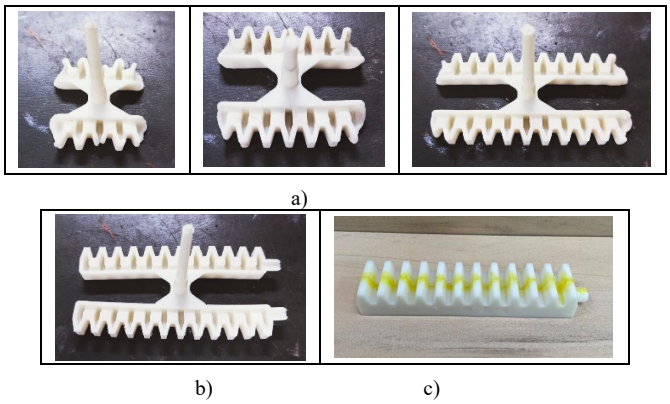


Figure 11: a) Filling experiment of the soft gripper 40%, 60%, and 80%, b) Gripper ejected from the mold after 100% filling, and c) Final gripper after glued

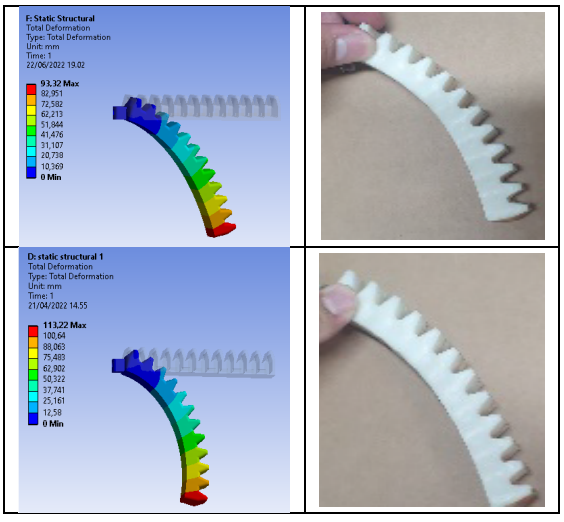
Compared to traditional production, the production time might range from hours to a day. However, most soft gripper manufacturers use 3D printing, which takes 35 minutes [28]. Therefore, the production time of one soft gripper is less than 5 minutes in this work, and we obtained an optimization technique for lowering time and expenses in creating soft grippers.

4.3 Warpage and Shrinkage Comparison

By the result of the total warpage displacement of Moldex 3D simulation and the experiment of the injection molding machine, the final of the simulation and experiment of the soft gripper. According to the result of Moldex 3D, warpage occurs, and the total warpage volumetric shrinkage is 11.969% based on the simulation. The original design size of the soft gripper was 114x19x16.5 mm. However, due to the warpage, the dimension of the soft gripper changed to 103x18.64x16.43 mm. Thus, the warpage volume shrinkage of the actual experiment is 11.96%. Remarkably similar for both simulation and experiment results.

4.4 Bending Comparison

The TPE soft gripper's bending experiment performs well in the bending pressure experiment, both positive and negative pressure. Furthermore, this design works as expected, considering the bending restriction from the previous research is eliminated as the gripper has no edge touching for the negative bend.



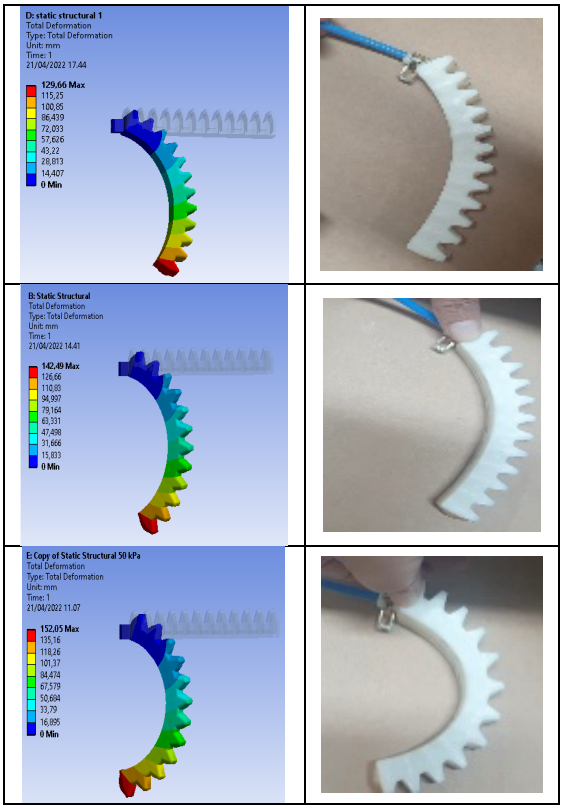


Figure 12: Actual positive bending deformation of soft gripper

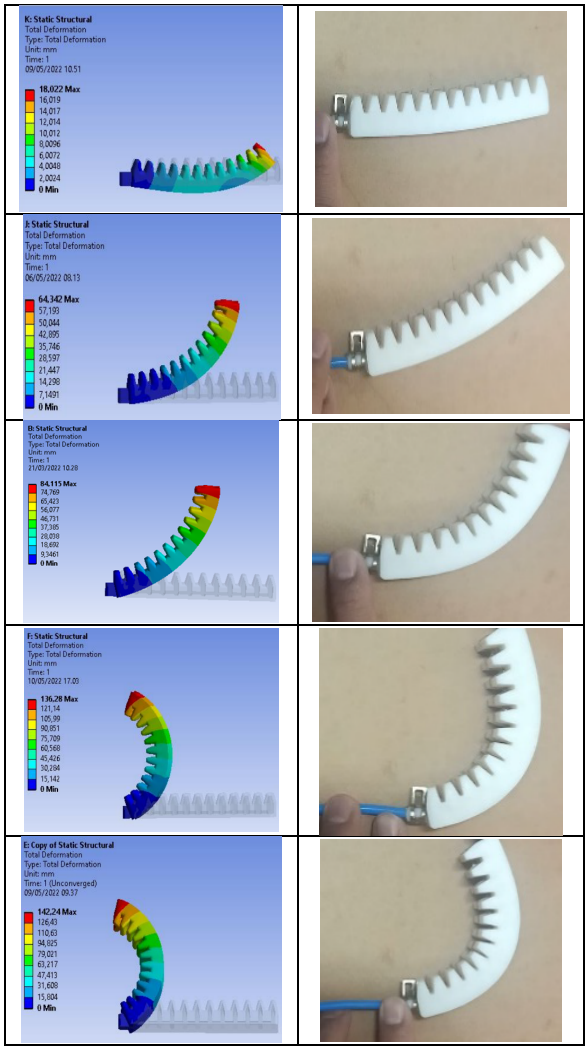


Figure 13: Actual negative bending deformation of soft gripper

However, this product's weakest sections can also be identified as the areas where fastening processes are used. Thus, processes involving high pressure or heavy loads increase the product's propensity to failure. Under excessive pressure, it will lead to failure and may cause the material to burst, rip, or deform permanently. Due to the fact that this soft gripper design must be glued after being manufactured, significant air pressure may cause separation between the two sections of the gripper finger as well. Furthermore, the chemicals employed in bonding can potentially harm the environment. Therefore, in this experiment, the range of air pressure applied only between -50 to 50 KPa, as shown in Figures 12 and 13.

Based on the comparison between simulation and actual experiments of bending deformation of the soft gripper with several pressure differences, Figure 14 shows the influence of the pressure.

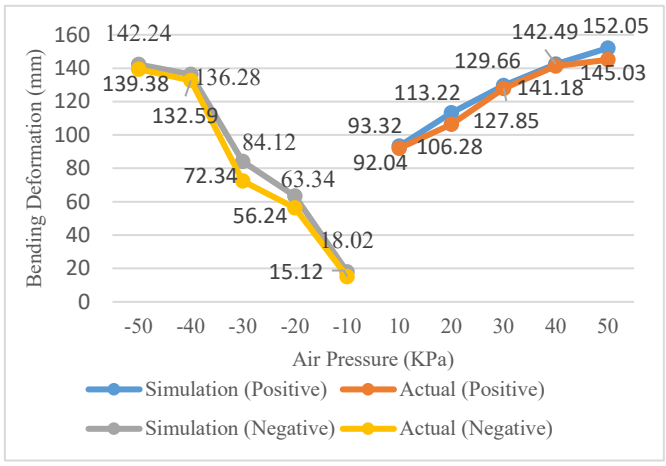


Figure 14: TPE bending deformation comparison influenced by the pressure

4.5 *Grasping Experiment*

Three soft grippers have been constructed using the fixture design in Figure 9 and utilized to pick up a softball as the goal is to examine the grasping force, as shown in Figure 15. The air pressure applied to the gripper is 40 KPa in order to get enough bending deformation to grasp the object. A 3D print weight object was also attached to the ball to examine the bending strength to encounter the additional weight beside the ball.

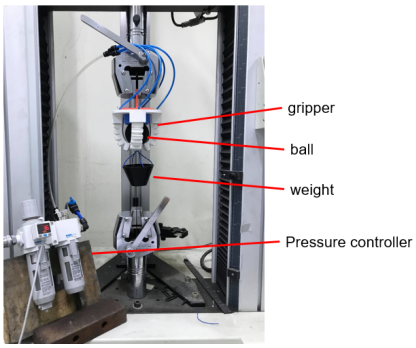


Figure 15: Grasping experiments of the soft gripper [1]

5. Conclusion

The numerical analysis of injection molding was done by Moldex 3D. The soft gripper deformation was successfully predicted by Ansys software, then proved by actual experiment. In both positive and negative pressure conditions, higher air pressure causes more widespread deformation of the soft gripper. Injection molding and FEM simulations indicated that creating a soft gripper is straightforward. Furthermore, the

modeling and real outcomes are fairly similar. In Moldex 3D simulation, the value of shrinkage is 11.969%, compared to the experiment of the soft gripper is 11.96%. Thus, the percentage warpage volumetric shrinkage for the simulation and actual experiment is similar. However, for future research, the directions may be to discover other convenient designs or material combinations to fulfill manufacture on a mass scale following the needs of Industry 4.0. It is recommended to use an injection molding machine as the manufacturing method since it is a popular technique for producing plastic components on a massive scale.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

The authors would like to thank the Ministry of Science and Technology (MOST) in Taiwan for the financial support of MOST-107-2221-E-224-045.

References

[1] M.R.F. Saduk, H.D. Bryantono, S.C. Tseng, "Engineering Design of Soft Gripper and Manufacturing by TPE Hollow Injection Molding," in Proceedings - 2022 IET International Conference on Engineering Technologies and Applications, IET-ICETA 2022, Institute of Electrical and Electronics Engineers Inc., 2022, doi:10.1109/IET-ICETA56553.2022.9971524.

[2] M.R. Fillipin Saduk, H. Jiaqi, S.C. Tseng, "Modeling and Experiments of a Soft Gripper for Robotics Arms Applications," in Proceedings of the 2022 8th International Conference on Applied System Innovation, ICASI 2022, Institute of Electrical and Electronics Engineers Inc.: 179–182, 2022, doi:10.1109/ICASI55125.2022.9774440.

[3] C. Majidi, Soft Robotics: A Perspective - Current Trends and Prospects for the Future, Soft Robotics, **1**(1), 5–11, 2014, doi:10.1089/soro.2013.0001.

[4] D. Rus, M.T. Tolley, Design, fabrication and control of soft robots, Nature, **521**(7553), 467–475, 2015, doi:10.1038/nature14543.

[5] D. Trivedi, C.D. Rahn, W.M. Kier, I.D. Walker, "Soft robotics: Biological inspiration, state of the art, and future research," Applied Bionics and Biomechanics, **5**(3), 99–117, 2008, doi:10.1080/11762320802557865.

[6] S. Kim, C. Laschi, B. Trimmer, Soft robotics: A bioinspired evolution in robotics, Trends in Biotechnology, **31**(5), 287–294, 2013, doi:10.1016/j.tibtech.2013.03.002.

[7] J. Shintake, V. Cacucciolo, D. Floreano, H. Shea, Soft Robotic Grippers, Advanced Materials, **30**(29), 2018, doi:10.1002/adma.201707035.

[8] P. Bharath Vamsi, V. Ragavendra Rao, "Design and fabrication of soft gripper using 3D printer," in IOP Conference Series: Materials Science and Engineering, Institute of Physics Publishing, 2018, doi:10.1088/1757-899X/402/1/012026.

[9] T. Hassan, M. Manti, G. Passetti, N. d'Elia, M. Cianchetti, C. Laschi, "Design and development of a bio-inspired, under-actuated soft gripper," in 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE: 3619–3622, 2015.

[10] R. Maruyama, T. Watanabe, M. Uchida, "Delicate grasping by robotic gripper with incompressible fluid-based deformable fingertips," in 2013 IEEE/RSJ International Conference on Intelligent Robots and Systems, IEEE: 5469–5474, 2013.

[11] Y. Yang, Y. Wu, C. Li, X. Yang, W. Chen, "Flexible Actuators for Soft Robotics," Advanced Intelligent Systems, **2**(1), 1900077, 2020, doi:10.1002/aisy.201900077.

[12] J. Pinskier, D. Howard, "From Bioinspiration to Computer Generation: Developments in Autonomous Soft Robot Design," Advanced Intelligent Systems, **4**(1), 2100086, 2022, doi:10.1002/aisy.202100086.

[13] M. Kaur, W.S. Kim, "Toward a Smart Compliant Robotic Gripper Equipped with 3D-Designed Cellular Fingers," Advanced Intelligent Systems, **1**(3), 1900019, 2019, doi:10.1002/aisy.201900019.

[14] F. Schmitt, O. Piccin, L. Barbé, B. Bayle, Soft robots manufacturing: A review, Frontiers Robotics AI, **5**(JUN), 2018, doi:10.3389/frobt.2018.00084.

[15] Z. Wang, Y. Torigoe, S. Hirai, "A Prestressed Soft Gripper: Design, Modeling, Fabrication, and Tests for Food Handling," IEEE Robotics and Automation Letters, **2**(4), 1909–1916, 2017, doi:10.1109/LRA.2017.2714141.

[16] J. Hughes, U. Culha, F. Giardina, F. Guenther, A. Rosendo, F. Iida, Soft manipulators and grippers: A review, Frontiers Robotics AI, **3**(NOV), 2016, doi:10.3389/frobt.2016.00069.

[17] S. Zaidi, M. Maselli, C. Laschi, M. Cianchetti, "Actuation Technologies for Soft Robot Grippers and Manipulators: A Review," Current Robotics Reports, **2**(3), 355–369, 2021, doi:10.1007/s43154-021-00054-5.

[18] G. Alici, "Softer is harder: What differentiates soft robotics from hard robotics?," in MRS Advances, Materials Research Society: 1557–1568, 2018, doi:10.1557/adv.2018.159.

[19] L. Wang, F. Iida, "Deformation in Soft-Matter Robotics: A Categorization and Quantitative Characterization," IEEE Robotics and Automation Magazine, **22**(3), 125–139, 2015, doi:10.1109/MRA.2015.2448277.

[20] Z. Wang, K. Or, S. Hirai, "A dual-mode soft gripper for food packaging," Robotics and Autonomous Systems, **125**, 2020, doi:10.1016/j.robot.2020.103427.

[21] W. Zhou, J. Lin, D.S. Balint, T.A. Dean, "Clarification of the effect of temperature and strain rate on workpiece deformation behaviour in metal forming processes," International Journal of Machine Tools and Manufacture, **171**, 2021, doi:10.1016/j.ijmachtools.2021.103815.

[22] M.-L. Wang, R.-Y. Chang, C.-H.D. Hsu, Molding simulation: Theory and practice, Carl Hanser Verlag GmbH Co KG, 2018.

[23] H. Schlichting, K. Gersten, H. Schlichting, K. Gersten, "Fundamentals of boundary-layer theory," Boundary-Layer Theory, 29–49, 2000.

[24] M.R. Barone, C. Tucker, Fundamentals of computer modeling for polymer processing, Hanser, 1989.

[25] Y. Hao, T. Wang, Z. Ren, Z. Gong, H. Wang, X. Yang, S. Guan, L. Wen, "Modeling and experiments of a soft robotic gripper in amphibious environments," International Journal of Advanced Robotic Systems, **14**(3), 2017, doi:10.1177/1729881417707148.

[26] S.E. Mousavi, N. Sukumar, "Generalized Gaussian quadrature rules for discontinuities and crack singularities in the extended finite element method," Computer Methods in Applied Mechanics and Engineering, **199**(49–52), 3237–3249, 2010, doi:10.1016/j.cma.2010.06.031.

[27] H. Kabir, M.M. Aghdam, "A generalized 2D Bézier-based solution for stress analysis of notched epoxy resin plates reinforced with graphene nanoplatelets," Thin-Walled Structures, **169**, 2021, doi:10.1016/j.tws.2021.108484.

[28] B.N. Peele, T.J. Wallin, H. Zhao, R.F. Shepherd, "3D printing antagonistic systems of artificial muscle using projection stereolithography," Bioinspiration and Biomimetics, **10**(5), 2015, doi:10.1088/1748-3190/10/5/055003.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 8, No. 4, 18-27 (2023)
www.astesj.com
Special Issue on Computing, Engineering and Multidisciplinary Sciences

ASTES Journal
ISSN: 2415-6698

Localization of Impulsive Sound Source in Shallow Waters using a Selective Modal Analysis Algorithm

Faraz Talebpour, Saeed Mozaffari, Mehrdad Saif, Shahpour Alirezaee*

Department of Electrical and Computer Engineering University of Windsor, Windsor, Ontario, Canada

ARTICLE INFO

Article history:
Received: 19 December, 2022
Accepted: 22 June, 2023
Online: 21 July, 2023

Keywords:
Passive Underwater Localization
Underwater Signal Processing
Passive Environmental Monitoring
Modal Analysis

ABSTRACT

Passive remote monitoring applications of underwater signal processing in a shallow water environment are an impactful area of research for environmental and marine-life monitoring. The majority of the sound source localization techniques require carefully placed synchronized hydrophone arrays, which can be complicated and hard to maintain. In this paper, we utilized the modal dispersions of a signal to derive a localization method for a noisy, shallow water environment. Our proposed algorithm employs modal selection to process the most noise-resistive dispersion curves, improving the accuracy and noise-resistivity of the existing methods. Moreover, we proposed a 2D localization method with multiple unsynchronized hydrophones and minimal hardware requirements and limitations. Furthermore, we analyzed the effects of underwater ambient noise on the accuracy of the proposed method, using simulated and real recorded explosion and whale sounds, and compared our algorithm's localization performance with others. Simulation results show increased localization accuracy of 30m for the recorded explosion sound and 360m for the Whale sound.

1 Introduction

This paper extends our previous work presented in CCECE 2022 [1] by introducing a selective-modal algorithm architecture for localizing impulsive sound sources in shallow waters. Our proposed algorithm improves performance in lower signal-to-noise ratio (SNR) scenarios by selecting the best modal pairs. In this paper, we provide a more detailed explanation of the localization formulas, propose a 2D unsynchronized localization scheme, analyze the performance of our algorithms using real recorded signals, and compare them with existing works. This paper extends our previous work presented in CCECE 2022 [1] by introducing a selective-modal algorithm architecture for localizing impulsive sound sources in shallow waters. Our proposed algorithm improves performance in lower signal-to-noise ratio (SNR) scenarios by selecting the best modal pairs. In this paper, we provide a more detailed explanation of the localization formulas, propose a 2D unsynchronized localization scheme, analyze the performance of our algorithms using real recorded signals, and compare them with existing works.

The field of underwater acoustics encompasses the primary modality of underwater sensing and communication, which is sound. Early research in underwater signal processing focused on mathematical models and the behavior of acoustic sounds in the underwater environment [2]. Over time, advancements in adaptive signal processing and sensor technology led to practical applications in underwater signal processing. Sonar systems, particularly underwater sonars, have undergone rapid developments in the past two decades, driven by increased processing capability and the implementation of more computationally intensive techniques. The underwater environment presents unique challenges, including increased human-made noise due to the growing number of vessels in the ocean. Marine mammals heavily rely on vocalization for communication and locating other mammals, making them sensitive to sounds generated by human activities such as geophysical explorations, offshore extraction, shipping, and active sonar applications[3]. As a result, researchers have been motivated to develop remote monitoring techniques to study marine mammal behavior and monitor environmental changes. Underwater localization techniques can be broadly classified into passive and active categories. Passive sonar processes received signals without signal transmission, while active sonar involves both signal transmission and reception [4]–[5]. Researchers have proposed various passive underwater localization methods, including time-frequency difference of arrival (TDOA), received signal strength (RSS), and modal-based analysis [6]–[7]. The

*Corresponding Author: Shahpour Alirezaee, CEI, University of Windsor, Windsor,ON, CA, 519-253-3000 ext. 7472 & alirezae@uwindsor.ca

underwater medium is a dynamic multi-path channel where sound-waves travel through multiple paths with different speeds [8, 9]. TDOA algorithms utilize time differences between received signals, while RSS algorithms focus on received signal power. However, implementing TDOA-based techniques often requires synchronized hydrophone arrays and prior information, resulting in increased costs, complexity, and high error levels in low SNR environments. In [10], the authors conducted experiments under real test conditions with sensor nodes and observed that the sensors constantly move due to varying water surface conditions, resulting in unsynchronized sensor nodes. To address this issue, the authors in [11] proposed a self-calibration technique utilizing a shift-keying pulse and composite transducers. Similarly, in [12], it was demonstrated that the use of maximum likelihood estimators (MLE) in TDOA methods led to non-linearity problems. In response, the authors in [13] formulated TDOA target motion analysis as a least-square optimization problem, solving it in polynomial time. Furthermore, [14] investigated the performance of TDOA techniques under different noise levels and highlighted the significant impact of white noise on the accuracy of TDOA algorithms.

To improve the accuracy and noise resistivity, [7] introduced a hybrid localization technique based on the direction of arrival (DOA) and received signal strength (RSS) using a vector and an isotropic acoustic hydrophone. Phased array-based localization was proposed in [15] to enhance noise resistivity. However, TDOA-based methods, while accurate, often require arrays of synchronized hydrophones and prior information, resulting in higher implementation costs, increased complexity, and reduced accuracy in low SNR environments. In [16], the authors suggested the utilization of the Kronecker product operation to extract the two-dimensional power distribution matrix from the beam power function, reducing the number of required hydrophones and improving noise resistivity.

Despite extensive efforts in the field, achieving sensor node synchronization and fulfilling the multi-hydrophone requirements of TDOA-based techniques can still pose significant challenges and incur high costs. To overcome these limitations, modal analysis-based localization was introduced as a solution, eliminating the need for source prior information, multiple hydrophones, and hydrophone synchronization [16, 17]. In the underwater environment, acoustic waves consist of multiple modes that travel through water with varying velocities. As a result of these differing velocities, the modes disperse during propagation through the water channel [6, 18]. In [19] , the authors proposed a modal analysis-based approach specifically designed for localizing mammal sounds. Furthermore, in [11], accuracy was enhanced by expanding the localization frequency range and considering additional modes during the localization process. Additionally, [20] proposed a nonlinear-based warping technique for modal filtering.

In this paper, we build upon our previous work published in [1] and introduce novel advancements to the field of underwater localization. Specifically, we extend our research by incorporating the utilization of multiple hydrophones for two-dimensional localization. Unlike previous approaches, our proposed techniques are independent and standalone, enabling each hydrophone to perform separate target localization in an unsynchronized manner.

To lay the groundwork for our methodology, we begin by introducing a shallow underwater channel model based on the theory of normal modes in Section 2. Additionally, we present a comprehensive model for the channel's ambient noise and derive the modal functions necessary for modal analysis.

In Section 3, we take a significant step forward by deriving a selective noise-resistive modal-based localization method that exhibits improved resistance to noise. This novel approach addresses a crucial challenge in underwater localization and enhances the accuracy of our algorithm.

To evaluate the performance of our proposed method, we present the obtained results in Section 4 and highlight the significance of modal selection for achieving superior performance. Furthermore, we thoroughly investigate the impact of noise on the accuracy of our algorithm within the 30dB < SNR < 45dB range, providing insightful comparisons with existing approaches.

In addition, we conduct an in-depth analysis and comparison of the accuracy and noise resistivity of our proposed method with other techniques using real recorded explosions and north Atlantic sounds. By doing so, we establish a comprehensive understanding of the strengths and limitations of our approach in realistic scenarios.

Finally, we evaluate the performance of our proposed 2D Localization and tracking method by comparing it with state-of-the-art techniques, demonstrating the advancements we have made in the field of underwater localization.

2 Normal Mode Propagation

Normal mode theory is suitable for modeling shallow underwater environments with respect to normal-Modes propagation. While modal-based channel models are not the most accurate model currently available, they can accurately model shallow underwater environments for passive sound source localization and monitoring applications.

2.1 *Underwater Acoustic Propagation*

Let us consider the model description presented in Figure 1 where an acoustic sound source is located at (x_s, y_s, z_s) that produces a continuous-time signal. After propagation, the signal is picked up by a hydrophone placed on a buoy at (x_h, y_h, z_h). For ease of use, we have considered the hydrophone on the right end of Figure 1 as the point of origin in the Cartesian and cylindrical coordinates. The displacement caused by the propagating source is time-harmonic, governed by Helmholtz law, and is given as [6, 21, 17, 22, 23, 24]

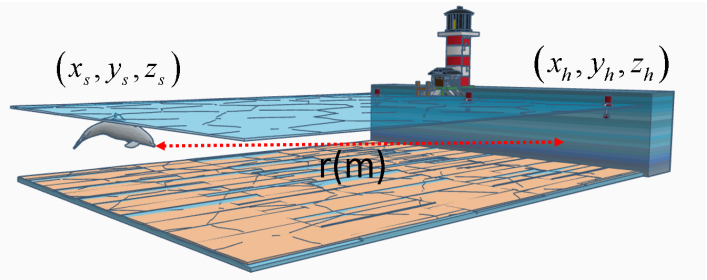


Figure 1: Model description

$$\left[\nabla^2 + K^2(\vec{r})\right] p(\vec{r}) = -4\pi f(\vec{r}) \tag{1}$$

$K(\vec{r})$ is the medium wave number at radial frequency ω,∇ gradient operator, and $p(\vec{r})$ is the pressure. We can further simplify this equation to form the Helmholtz equation in two dimensions, as the sound speed and density depends only on depth z

$$\frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial p}{\partial r}\right) + \rho(z)\frac{\partial}{\partial z}\left(\frac{1}{\rho(z)}\frac{\partial p}{\partial z}\right) + \frac{\omega^2}{c^2}p = \frac{\delta(r)\,\delta(z-z_s)}{-2\pi r} \tag{2}$$

where r is the distance to the source, ρ is the medium density, c is the propagation speed, δ is the Dirac delta, and ω is angular velocity [6]. Using the separation of variables, we look for a depth-related pressure solution in the form of $p(r,z) = \phi(r)\psi(z)$, which will result in

$$\frac{1}{\phi}\left[\frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial \phi}{\partial r}\right)\right] + \frac{1}{\psi}\left[\rho(z)\frac{\partial}{\partial z}\left(\frac{1}{\rho(z)}\frac{\partial \psi}{\partial z}\right) + \frac{\omega^2}{c^2}\psi\right] = 0 \tag{3}$$

where ϕ is the volume displacement, and Ψ is the general modal depth function. Terms in square brackets of the equation (3) are functions of $r$and $z$respectively. To satisfy the equation (3), each term should be equal to a constant [24]. Considering the Pekeris waveguide -where water is considered equal columns with varying speeds of propagation- We can drive the modal equation by considering the K_{rm}^2 as a separation constant [25].

$$\begin{aligned} &\rho(z)\tfrac{\partial}{\partial z}\left(\tfrac{1}{\rho(z)}\tfrac{\partial \psi_m(z)}{\partial z}\right) + \left[\tfrac{\omega^2}{c^2} - K_{rm}^2\right]\psi_m(z) = 0 \\ &\psi(0) = 0 \quad , \quad \tfrac{d\psi}{dz}\Big|_{z=D} = 0 \end{aligned} \tag{4}$$

where $\psi_m(z)$ is the particular modal function $\psi(z)$ obtained with horizontal wave-number K_{rm} as separation constant. The boundary condition of the equation (4) considers each water column a pressure release surface ($z = 0$) and a perfectly rigid seabed at $z = D$ ($D < 100$) which translates to no changes in the volume at surface and seabed resulting in $d\psi/dz = 0$ [25].
Equation (4) is a classical Sturm-Liouville eigenvalue problem [24]. Applying the orthogonality of the modal Sturm-Liouville problem, we can write

$$\int_0^D \frac{\psi_m(z)\,\psi_n(z)}{\rho(z)}dz = 0 \qquad m \neq n \tag{5}$$

Equation (3), the solutions of modal equations are arbitrary to multicaptive constants; therefore, we can further simplify the results using equation (5) as

$$\int_0^D \frac{\psi_m^2(z)}{\rho(z)}dz = 1 \tag{6}$$

Moreover, modes transmit as a complete set, resulting in an arbitrary function as a sum of all normal modes, which will yield the pressure function as:

$$p(r,z) = \sum_{m=1}^{\infty} \phi_m(r)\,\psi_m(z) \tag{7}$$

Substituting equation (7) in equation (2) provides :

$$\begin{aligned} &\textstyle\sum_{m=1}^{\infty}\left\{\frac{1}{r}\frac{d}{dr}\left(r\frac{d\phi_m(r)}{dr}\right)\psi_m(z) + \phi_m(r)\left[\rho(z)\frac{d}{dz}\left(\frac{1}{\rho(z)}\frac{d\psi_m(z)}{dz}\right) + \frac{\omega^2}{c^2}\psi_m(z)\right]\right\} \\ &= -\frac{\delta(r)\delta(z-z_s)}{2\pi r} \end{aligned} \tag{8}$$

After applying the operator equation (9)

$$\int_0^D (\cdot)\frac{\psi_n(z)}{\rho(z)}dz \tag{9}$$

Furthermore, considering the orthogonality property stated in the equation (5), only n terms of the sum remain.

$$\frac{1}{r}\frac{d}{dr}\left(r\frac{d\phi_n(r)}{dr}\right) + K_{rn}^2\phi_n(r) = -\frac{\delta(r)\,\psi_n(z_s)}{2\pi r\rho(z_s)} \tag{10}$$

the solution to the equation (10) is provided in terms of the Hankel function as:

$$\phi_n(r) = \frac{1}{4\rho(z_s)}\left(\psi_n(z_s)\,H_0^{(1,2)}(K_{rn}r)\right) \tag{11}$$

The signal's energy radiates outwards, and therefore the solution will be $H_0^{(1)}$. Considering the radiation conditions, after substituting (11) in (7), we can derive the pressure equation based on the modal function as

$$p(r,z) = \frac{1}{4\rho(z_s)}\sum_{m=1}^{\infty}\psi_m(z_s)\,\psi_m(z)\,H_0^{(1)}(K_{rm}r) \tag{12}$$

we can further simply equation (12) by using the asymptotic approximation to the Henkal's function, yielding:

$$p(r,z) \simeq \frac{i}{\rho(z_s)\sqrt{8\pi r}}e^{-\frac{i\pi}{4}}\sum_{m=1}^{\infty}\psi_m(z_s)\,\psi_m(z)\,\frac{e^{iK_{rm}r}}{\sqrt{K_{rm}}} \tag{13}$$

provides us with the pressure function, based solely on modal functions and depth.

2.2 Solution to Wave Equation

We must simplify the displacement equations further to perform channel modeling in simulation software. The non-homogeneous differential equation (1) can be solved using the Green's function method and expanded as the displacement equation as [6, 23]

$$\begin{aligned} &\rho(z)\frac{d}{dz}\left[\frac{1}{\rho(z)}\frac{dg(z)}{dz}\right] + \left[\frac{\omega^2}{c^2} - K_r^2(z)\right]g(z) = \frac{\delta(z-z_s)}{-2\pi} \\ &\delta(z-z_s) = \sum_m a_m\psi(z_s) \\ &\delta(z-z_s) = \sum_m \frac{\psi_m(z_s)\,\psi_m(z)}{\rho(z_s)}, a_m = \frac{\psi_m(z_s)}{\rho(z_s)} \end{aligned} \tag{14}$$

substituting $g(z) = \sum_m a_m\psi_m(z)$ In equation (14) provides the depth related modal function as:

$$\begin{aligned} &\textstyle\sum_{m=1}^{\infty} b_m\left[\left(K_{rm}^2 - K_r^2\right)\psi_m(z)\right] = \\ &\frac{1}{-2\pi}\sum_m \frac{\psi_m(z_s)\,\psi_m(z)}{\rho(z_s)}, K_z^2 = K_{rm}^2 - K_r^2 \end{aligned} \tag{15}$$

where K_{rm}, K_z, K_r are the angular, vertical and horizontal wavenumbers [26, 6]. applying Green's solution to equation (15) would provide the general modal function:

$$g(z) = -\frac{1}{2\pi\rho(z_s)} \sum_m \frac{\psi_m(z_s)\psi_m(z)}{K_r^2 - K_{rm}^2} \tag{16}$$

with general solutions and eigenfunctions as follows

$$\begin{aligned} &\psi_m(z) = A\sin(K_z z) + B\cos(K_z z) \\ &K_z = \sqrt{(K_{rm})^2 - K_r^2} \qquad K_{rm} = \frac{\omega}{c} \\ &\psi_m(z) = \sqrt{2(\rho/D)}\sin(K_z z) \\ &v_m(\omega) = \frac{\omega}{K_{rm}} \\ &\frac{\omega}{c_{Seabed}} < \text{Number of Modes} < \frac{\omega}{c_{Water}} \end{aligned} \tag{17}$$

where $v_m(\omega)$ is the velocity of mode m at angular frequency ω.

2.3 Underwater Ambient Noise

Noise in a shallow underwater environment can be categorized into two main types, ambient noise caused by the channel characteristics and artificial noises created by external sources such as ships and marine life. Many studies consider the noise a simple added white noise; however, underwater ambient noise can be more accurately modeled as colored noise. The underwater channel's behavior is best described as a low-pass filter. It can be modeled as a white noise sequence filtered using a Butterworth IIR low-pass filter with 30dB attenuation in stopband and normalized stopband frequency of 0.05 $half cyle/sample$ per sample ad 0.9 $half cyle/sample$ Respectively [27]. Figure 2 presents the signal and noise in the time domain with SNR=45dB.

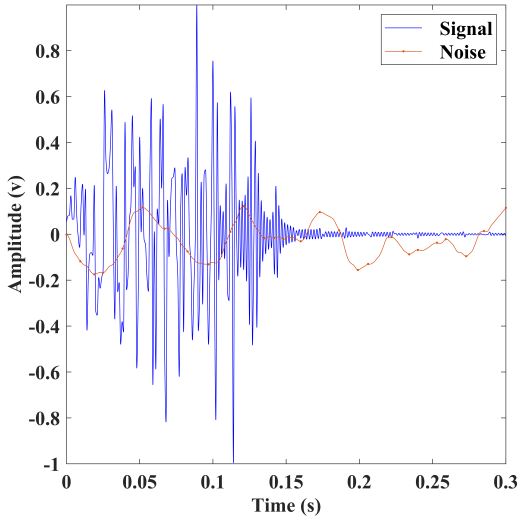


Figure 2: Signal (blue), Noise (red)

3 Modal Analysis based localization

In the previous section, we introduced the channel model and modal functions. In this section, we will derive the necessary equations for the localization of impulsive sound sources using modal functions. The modal-based localization methods are based on the dispersion of the natural frequencies as they propagate underwater.

3.1 Modal Dispersion

As stated earlier, modes travel at different speeds (equation (17)), resulting in dispersion at the receiver. Let us consider the simulation scenario of Figure.1, where the Normal mode theory with ambient noise is used to model the channel. Considering an impulsive sound source at a depth of D_s=20m, 4000 meters away from the hydrophone, ($\rho_{(Seabed)}$=1000(Kg/m^3), $\rho_{(Water)}$=1000(Kg/m^3), $c_{(Seabed)}$=1500(m/s), $c_{(Water)}$=1600(m/s)), the propagated signal will have the time-frequency (TF) representation provided in Figure.3, which illustrates the dispersion caused by the difference in propagation speeds. One can employ the dispersion of modes to localize the sound source through modal analysis after filtering them.

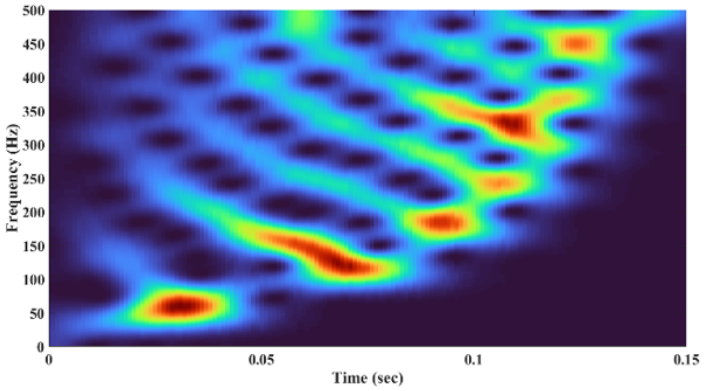


Figure 3: TF analysis, Modal dispersion , $f_{(Max)}$=600Hz

As the TF analysis graph illustrates, the dispersion curve's frequencies overlap between the modes and render conventional filtering techniques inert. The overlapped frequencies are the product of the nonlinear phase characteristics in the equation (16).To address this issue, considering the pressure signal in the time domain as

$$P(t) = \sum_m \psi_m(t)\, e^{2j\pi\nu_c(m)\zeta(t)} \tag{18}$$

Where $\zeta(t)$is the dispersity function $\zeta(t)$ in the time domain is given as

$$\zeta(t) = \sqrt{t^2 - t_r^2} = \sqrt{t^2 - (r/v_g)^2} \tag{19}$$

Using $\zeta(t)$, we can warp the signal by linearizing the phase using a warping function [17]:

$$\begin{cases} \zeta = \sqrt{t^2 - (r/v_g)^2} \\ \zeta\zeta^{-1} = 1 \end{cases} \rightarrow \zeta^{-1}(t) = \sqrt{t^2 + \left(r/v_g\right)^2} \tag{20}$$

Applying the warping function ζ^{-1} linearizes the phase. The TF graph of the linearized signal is presented in Figure 4.

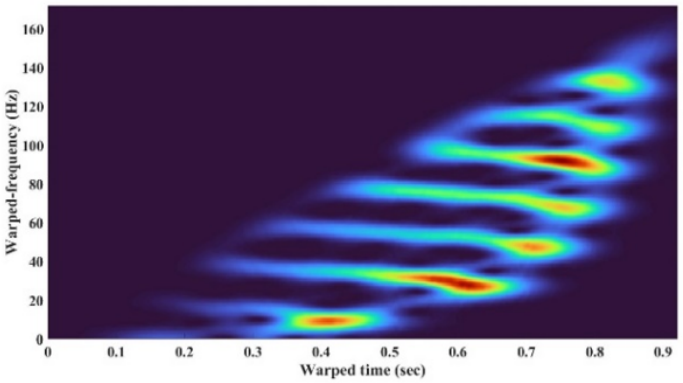


Figure 4: TF graph of warped signal

3.2 Localization Algorithm

Since the modal dispersion is directly related to the speed of propagation and distance, we can develop localization algorithms based on the TDOA concept. After filtering each of the dispersion curves; in an accurate channel model, the following expression will be true

$$\underset{Estimated}{\tau_n}(r,c) - \underset{Measured}{\tau_n} \approx 0 \qquad \forall n \tag{21}$$

Where τ_n is the dispersion curve. Measured τ_n can be obtained by warping and filtering each of the modals and by substituting the relationship between velocity and distance in the equation (17), the estimated τ_n can be obtained as:

$$\underset{Estimated}{\tau_n}(r,c) = \frac{r}{v_g(f,n)} \tag{22}$$

Where τ_n is the estimated dispersion curve for mode n transmitted over the range R with seabed sound speed c and group velocity $\upsilon_g(f,n)$. To localize the signal, we are looking for a range r that minimizes the statement (21). In other words

$$[\hat{r}] = \arg\min_{[\hat{r}]}\left(\left(\underset{Estimated}{\tau_m}(r,c_{seabed}) - \underset{Estimated}{\tau_n}(r,c_{seabed})\right) - \ldots \left(\underset{Measured}{\tau_m} - \underset{Measured}{\tau_n}\right)\right) \tag{23}$$

Where m and n can be any of the modes, summing over all frequency bins will yield

$$\sum_n \sum_m \sum_f \left[\left(\underset{Estimated}{\Delta\tau_{n,m}}(r,c)\right) - \left(\underset{Measured}{\Delta\tau_{n,m}}\right)\right] \approx 0 \qquad \forall n,m \tag{24}$$

Equation (24) results in a $m \times n$ matrix of dispersion curve differences and are used to derive the following cost function

$$\eta(r,c,n,f) = \sum_r \sum_n \sum_m \sum_f \left[\left(\left(\underset{Estimated}{\Delta\tau_{n,m}}(r)\right) - \left(\underset{Measured}{\Delta\tau_{n,m}}\right)\right)\right]^2 \tag{25}$$

We employed a grid search algorithm to minimize the cost function η for values of r.

Algorithm 1 presents our proposed method where μ_r defines the localization step size in the search boundary $[r_{min}, r_{max}]$ and ε is the accuracy of the estimated range. The Localization is performed in two steps; first, seabed and water parameters are defined based on the environment, and search boundaries for range and propagation speeds in seabed and sea are set. Next, the tensor of order 3, as shown in Figure 5, is formed to find the pairs of dispersion curves with the best performance (lowest value). Then, the cost function is formed only for the selected pairs of modes. Using a grid search algorithm, the location of the source can be estimated.

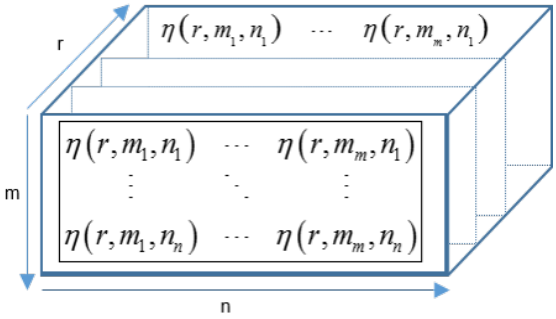


Figure 5: Cost function $[\ ]_{r\times m\times n}$

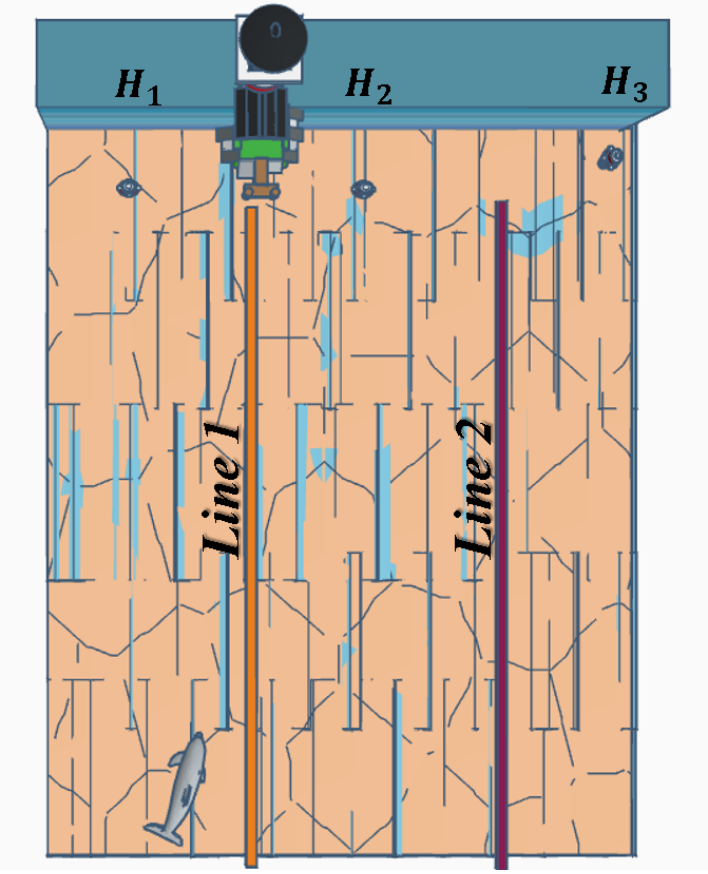


Figure 6: Model Description, multi hydrophone (H1,H2,H3)

Algorithm 1: Proposed localization Algorithm

Result: r
Initialization $\hat{r} = (r_{\min} : r_{\min} + \mu_r : r_{\max})$, $\rho_{water}, \rho_{seabed}$;
Warp Input signal;
Extract τ_n using TF ;
while $(r_{\min} < \hat{r} < r_{\max})$ **do**
  Form MPC;
  Select best modal pair;
  Minimize cost function;
  if $\Delta r, \leq \varepsilon$ **then**
    return:$[\hat{r}]$;
  else
    Change μ_r;
  end
end

3.3 2D Localization

While most modal-based localization methods proposed by literature perform ranging, we propose a method for unsynchronized 2D localization with minimal hardware requirements. In the case of 2D-localization requirements, buoys (each with a single hydrophone) can transmit the received signals to a base station on shore or a vessel to be analyzed in a central processor. Although utilizing multiple hydrophones would require sensor synchronization in other methods, the proposed modal-based localization analyzes modes picked up by each hydrophone separately. Moreover, given the high-range localization capabilities, buoys can be placed far apart, reducing implementation costs. Figure 6 illustrates the model description for 2D localization, where lines $Line1$ and $Line2$ are assumed at coordinates $[(x_{H2} - x_{H1})/2], [(x_{H2} - x_{H3})/2]$. Given the distances of each buoy, each hydrophone's average power of received modes is different. Hydrophones with the highest levels of received signal power are closest to the target. In the model description presented in Figure.6 ; $P(B_{H3}) < P(B_{H2}) < P(B_{H1})$ places the estimated latitude of the source $Line1_{H1H2} > x_s$. Based on the estimated location of the source and three calculated ranges from each buoy, we can perform 2D triangulation and track an object without needing a synchronized sensor array.

4 Results and Discussions

This section includes numerical experiments to illustrate the proposed localization method and discusses the effects of ambient noise on the accuracy of the proposed algorithm. Modal analysis is suitable for processing underwater signals in long distances ($r > 1000m$) based on only one hydrophone without synchronization.

4.1 Simulated Sound

We consider an impulsive sound source is placed at Cartesian coordinates (4000,45,0) with a maximum frequency of 500 Hz. An Omni-directional hydrophone is located at (0,15,0). We assume the speed of propagation in the seabed $c_b = 1600m/s$, speed of propagation in water c_w=1500m/s, density in water ρ_w=1000kg/m^3, and density in the seabed ρ_b=1500kg/m^3. The performance is evaluated based on the cost function's mean square error (MSE) and the estimated range's Root Mean Square Error (RMSE). Moreover, the result of this study is compared with those of [17], which has used the same approach in localization.

Figure 7 (a),(b), and (c) illustrates the RMSE of the cost function for SNR=45dB,35dB, and 30dB values for each modal pair. As we can see, considering the low-pass filter nature of the ambient noise, the noise than others would more influence pairs of first and last modes. This is mainly due to both filter boundaries' relatively low stop-band attenuation. This effect can compromise localization accuracy in low SNR environments. To address this issue, we proposed employing the cost-function MSE matrix of Figure 7, using the equation (25) to identify the best and most noise-resistant pairs of modes (lowest values), resulting in the lowest MSE. After identifying the best modal pairs (2 pairs in this study), we can find the estimated location of the acoustic sound source through the equation (23).

Figures 7 (a), (b), and (c) depict the Root Mean Square Error (RMSE) of the cost function for SNR values of 45dB, 35dB, and 30dB, respectively, for each pair of modes. It can be observed that, due to the low-pass filter characteristics of ambient noise, certain modal pairs are more influenced by noise compared to others. This effect is particularly prominent in the first and last mode pairs, primarily because of the relatively low stop-band attenuation at the boundaries of the filter. In low SNR environments, this influence can significantly compromise localization accuracy. Furthermore, Figure 7 demonstrates that the choice of modal pairs significantly impacts the error levels, as different pairs yield varying levels of error. The study presented in [17] solely employs modal pairs with sequential wavenumbers numbers, disregarding the performance of different pairs. To address this issue, we propose utilizing the MSE matrix of Figure 33 as the cost function, employing equation (25) to identify the most noise-resistant and optimal pairs of modes (with the lowest values). This selection process leads to lower MSE and enables us to determine the estimated location of the acoustic sound source using equation (23).

Figure 8a showcases the Root Mean Square Error (RMSE) of our proposed cost function for range estimation at different SNR levels, and it compares these results with the localization outcomes presented in [17]. Figures 8a and 8b clearly demonstrate that our proposed method exhibits superior performance in both low and high SNR environments. This improvement can be attributed to the fact that the localization method employed in [17] does not incorporate mode pair evaluation or selection. Instead, they utilize pairs of modes with consecutive mode numbers in their localization algorithm. However, as indicated in Figure 7, sequential mode numbers do not necessarily yield better localization results. By performing mode evaluation and selection, as shown in Figures 8a and 8b, the localization algorithm becomes more resilient to high levels of noise and achieves greater accuracy.

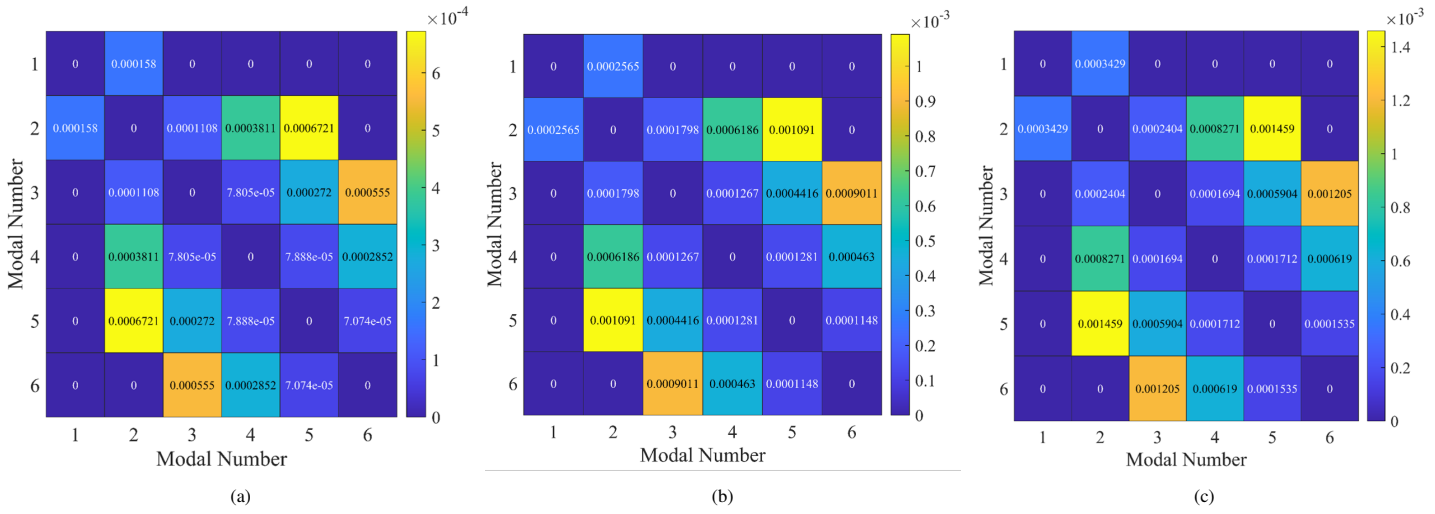


Figure 7: Cost function MSE for (a)SNR:45dB, (b)35dB, (c)30dB

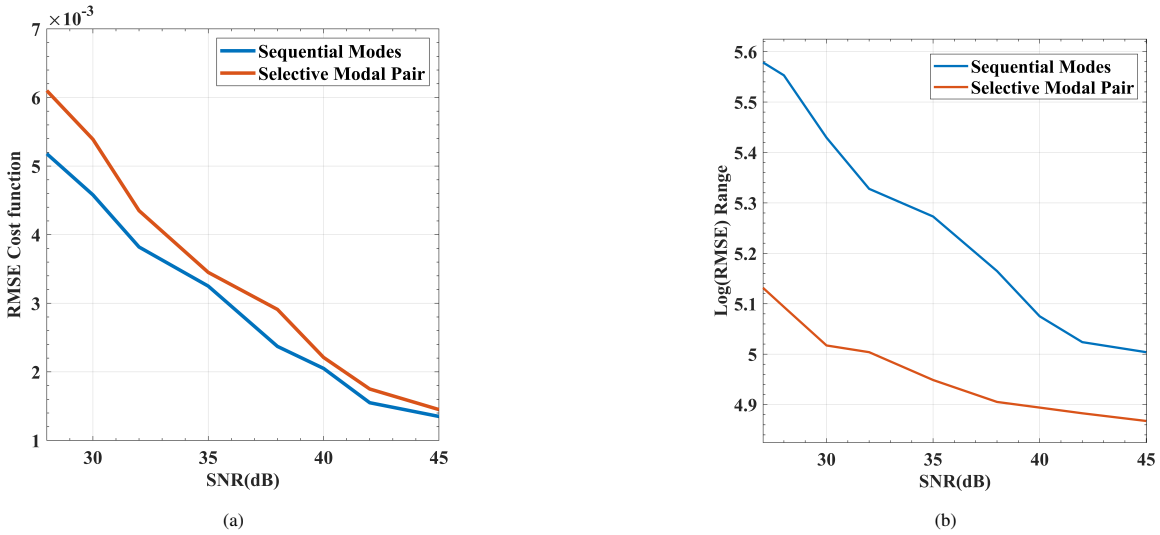


Figure 8: Time Frequency Representation (TFR) of (a) Cost Function RMSE (b) Log(RMSE) of estimated range for: $28dB < SNR < 45dB$

4.2 *Recorded North Atlantic Whale and Explosion Sound Localization*

In this section, we conduct a comprehensive evaluation of our proposed selective weighted algorithm using two distinct sound sources: the sound of a North Atlantic Right Whale and an explosion sound. Figure 9 depicts the time series and time-frequency (TF) analysis of these signals transmitted over different distances: 4.5 Km (z=20m) for the explosion sound and 8.7 Km (z=66m) for the whale sound. The TF analysis reveals that the noisy signal representing the explosion has a maximum frequency of 450 Hz, while the whale sound exhibits a lower maximum frequency of approximately 350 Hz. Furthermore, it is evident that certain modes are more susceptible to interference, highlighting the significance of modal selection and weighting functions in our approach.

We proceeded to localize the two signals and compared our results with our previous work and other existing methods. Table 1 presents the localization outcomes, demonstrating notable improvements compared to other proposed methods. Our Selective-modal based localization (SMP) approach achieved an error rate of 2.6% for both the recorded explosion sound and whale sound, while the Sequential Pair-Mode Analysis (SM) method yielded error rates of 3.11% and 6.2% for the respective signals. The superior performance of our proposed SMP method can be attributed to employing a larger number of dispersion curves (as opposed to only six sequential dispersion curves in SM) and performing initial modal selection.

Despite these improvements, it is important to note, as indicated in the TF analysis of Figure 9 and discussed in Section 3, that noise and channel effects vary across different modes. Consequently, each mode exhibits different weights and importance in the localization process, a consideration that is addressed in our approach.

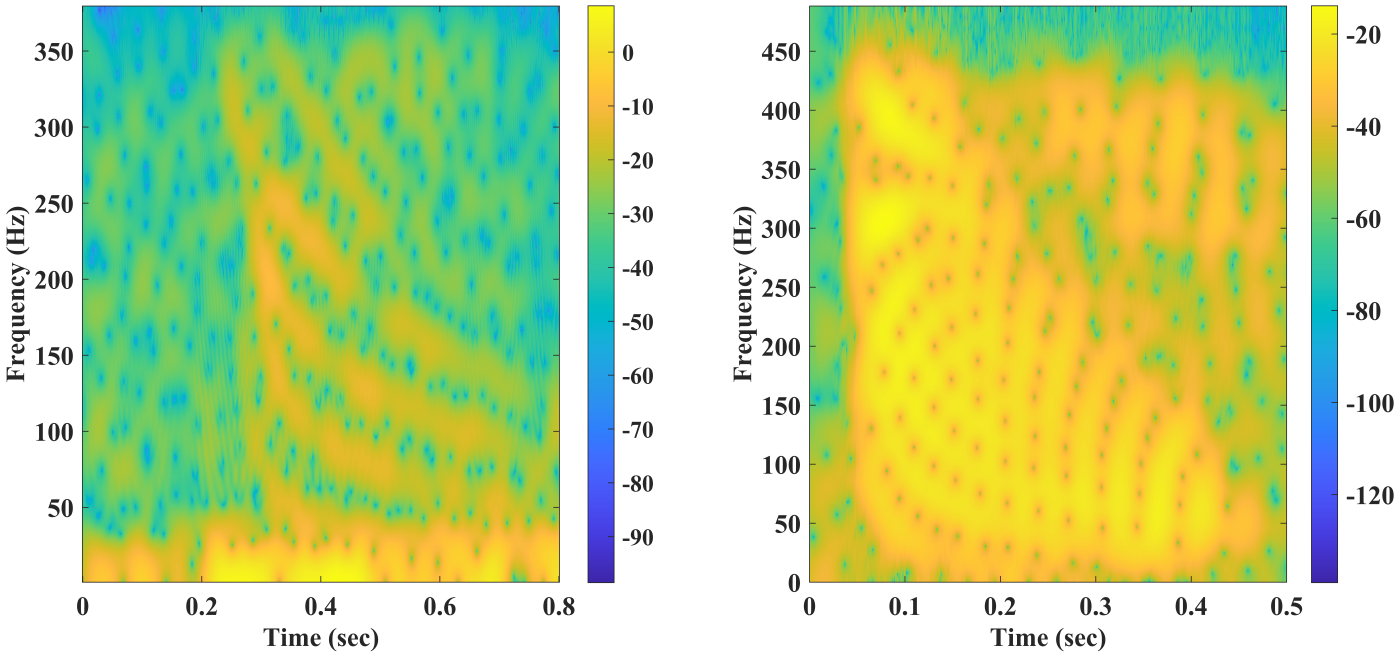


Figure 9: TF analysis: (a) North Atlantic Whale (b) Underwater Explosion

Table 1: Localization of Recorded North Atlantic whale (r=8700m) and explosion sound(r=4500)

Signal Source	Method	Number of Modes Used	Range (m)	Error (%)	References
Explosion Sound	Sequential Pair-Mode Analysis (SM)	6	4351	3.11	[17]
	Selective modal-Pair Analysis	9	4383	2.6	Proposed
North Atlantic Whale	Sequential Pair-Mode Analysis	4	9240	6.2	[17]
	Mode analysis	2	9225	6.03	[28]
	Downhill simplex algorithm	2	8884	2.11	[29]
	TOA	2	8950	2.87	[30]
	Selective modal-Pair Analysis	4	8881	2.06	Proposed

4.3 2D Localization

In this section, we conduct a comparative analysis of the 2D tracking performance of our localization algorithm in relation to other methods. Using the model description outlined in Figure 6, we employed a simulated non-stationary impulsive sound source that closely resembles the characteristics of a traveling whale following a sinusoidal path along the (x, y) axis.

Our 2D localization approach involves estimating the range of the sound source to each buoy, followed by triangulation based on the approximate direction of arrival and the intersection point of circles with a radius of r_h. The localization results for both the Sequential modes (SM) and our proposed Selective-modal based localization (SMP) are depicted in Figure 9, along with the true location of the sound source. It is evident from the results that SMP exhibits a closer adherence to the true range line compared to SM. This improved performance can be attributed to the modal selection function we introduced in this paper, which enables more accurate localization of the sound source.

5 Conclusion

In this study, we presented a passive impulsive sound source localization approach specifically designed for shallow underwater environments. Our method utilized the normal mode channel model and ambient noise to achieve accurate localization. A key contribution of this paper is the introduction of a localization scheme that incorporates modal pair selection, enabling enhanced noise resistance and improved accuracy.

Additionally, we proposed a 2D localization technique suitable for unsynchronized hydrophones, which aligns with the requirements of existing remote monitoring systems. To evaluate the performance of our algorithm, we conducted extensive analyses under various signal-to-noise ratio (SNR) conditions, comparing its noise resistance capabilities with other methods.

Furthermore, we validated our algorithm by testing it with actual recorded whale and explosion sounds. The results demonstrated its effectiveness in accurately tracking impulsive sound sources in a 2D space. Overall, our proposed approach showcases advancements in impulsive sound source localization and offers notable improvements over existing techniques.

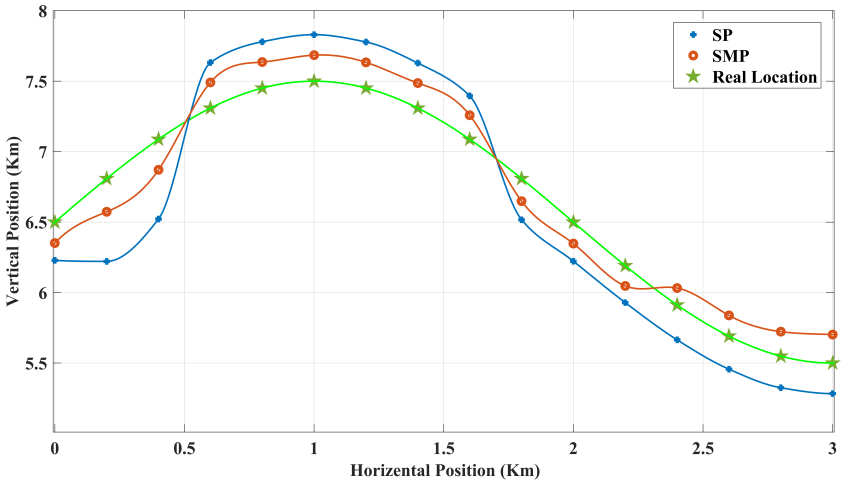


Figure 10: 2D localization and tracking of an impulsive sound source

References

[1] F. Talebpour, S. Mozaffari, M. Saif, S. Alirezaee, "Multi-Modal Signal Analysis for Underwater Acoustic Sound Processing," in 2022 IEEE Canadian Conference on Electrical and Computer Engineering (CCECE), 300–305, IEEE.

[2] X. Su, I. Ullah, X. Liu, D. Choi, "A Review of Underwater Localization Techniques, Algorithms, and Challenges," Journal of Sensors, **2020**, 1–24, 2020, doi:10.1155/2020/6403161.

[3] R. Vaccaro, "The past, present, and the future of underwater acoustic signal processing," IEEE signal processing magezine, **15**(4), 21–51, 1998.

[4] V. Kavoosi, M. J. Dehghani, R. Javidan, "Underwater Acoustic Source Positioning by Isotropic and Vector Hydrophone Combination," Journal of Sound and Vibration, **501**, 116031, 2021, doi:10.1016/j.jsv.2021.116031.

[5] K. F. Bou-Hamdan, A. H. Abbas, "Utilizing ultrasonic waves in the investigation of contact stresses, areas, and embedment of spheres in manufactured materials replicating proppants and brittle rocks," Arabian Journal for Science and Engineering, **47**(9), 11635–11650, 2022.

[6] D. A. Abraham, Underwater Acoustic Signal Processing: Modeling, Detection, and Estimation, Springer, 2019.

[7] D. Neupane, J. Seok, "A Review on Deep Learning-Based Approaches for Automatic Sonar Target Recognition," Electronics, **9**(11), 1972, 2020, doi:10.3390/electronics9111972.

[8] W. A. Kuperman, J. F. Lynch, "Shallow-Water Acoustics," Physics Today, **57**(10), 55–61, 2004, doi:10.1063/1.1825269.

[9] S. M. Wiggins, M. A. McDonald, L. Munger, S. E. Moore, J. A. Hildebrand, "Waveguide propagation allows range estimates for North Pacific right whales in the Bering Sea," Canadian acoustics, **32**, 146–154, 2004.

[10] M. Sanguineti, J. Alessi, M. Brunoldi, G. Cannarile, O. Cavalleri, R. Cerruti, N. Falzoi, F. Gaberscek, C. Gili, G. Gnone, D. Grosso, C. Guidi, A. Mandich, C. Melchiorre, A. Pesce, M. Petrillo, M. G. Taiuti, B. Valettini, G. Viano, "An automated passive acoustic monitoring system for real time sperm whale (Physeter macrocephalus) threat prevention in the Mediterranean Sea," Applied Acoustics, **172**, 107650, 2021, doi:10.1016/j.apacoust.2020.107650.

[11] L. An, L. Chen, "A real-time array calibration method for underwater acoustic flexible sensor array," Applied Acoustics, **97**, 54–64, 2015, doi:10.1016/j.apacoust.2015.04.008.

[12] F. B. Jensen, W. A. Kuperman, M. B. Porter, H. Schmidt, A. Tolstoy, Computational ocean acoustics, volume 794, Springer, 2011.

[13] S. Khazaie, X. Wang, P. Sagaut, "Localization of random acoustic sources in an inhomogeneous medium," Journal of Sound and Vibration, **384**, 75–93, 2016, doi:10.1016/j.jsv.2016.08.004.

[14] G. Wang, S. Cai, Y. Li, M. Jin, "Second-Order Cone Relaxation for TOA-Based Source Localization With Unknown Start Transmission Time," IEEE Transactions on Vehicular Technology, **63**(6), 2973–2977, 2014, doi:10.1109/tvt.2013.2294452.

[15] Y. Zou, H. Liu, Q. Wan, "An Iterative Method for Moving Target Localization Using TDOA and FDOA Measurements," IEEE Access, **6**, 2746–2754, 2018, doi:10.1109/access.2017.2785182.

[16] P. Wu, S. Su, Z. Zuo, X. Guo, B. Sun, X. Wen, "Time Difference of Arrival (TDoA) Localization Combining Weighted Least Squares and Firefly Algorithm," Sensors (Basel), **19**(11), 2554, 2019, doi:10.3390/s19112554.

[17] J. Bonnel, A. Thode, D. Wright, R. Chapman, "Nonlinear time-warping made simple: A step-by-step tutorial on underwater acoustic modal separation with a single hydrophone," J Acoust Soc Am, **147**(3), 1897, 2020, doi:10.1121/10.0000937.

[18] E. Xu, Z. Ding, S. Dasgupta, "Source Localization in Wireless Sensor Networks From Signal Time-of-Arrival Measurements," IEEE Transactions on Signal Processing, **59**(6), 2887–2897, 2011, doi:10.1109/tsp.2011.2116012.

[19] R. Diamant, L. Lampe, "Underwater Localization with Time-Synchronization and Propagation Speed Uncertainties," IEEE Transactions on Mobile Computing, **12**(7), 1257–1269, 2013, doi:10.1109/tmc.2012.100.

[20] A. Thode, J. Bonnel, M. Thieury, A. Fagan, C. M. Verlinden, D. Wright, J. Crance, C. L. Berchok, "Using nonlinear time warping to estimate North Pacific right whale calling depths and propagation environment in the Bering Sea," Journal of the Acoustical Society of America, **142**(4), 2711–2712, 2017.

[21] H. Jia, X. Li, "Underwater reverberation suppression based on non-negative matrix factorisation," Journal of Sound and Vibration, **506**, 116166, 2021, doi:10.1016/j.jsv.2021.116166.

[22] Y. Tian, M. Liu, S. Zhang, T. Zhou, "Underwater multi-target passive detection based on transient signals using adaptive empirical mode decomposition," Applied Acoustics, **190**, 108641, 2022.

[23] C. T. Tindle, A. Stamp, K. Guthrie, "Virtual modes and the surface boundary condition in underwater acoustics," Journal of Sound Vibration, **49**(2), 231–240, 1976.

[24] E. Costa, L. Godinho, W. Mansur, Peters, "Prediction of Acoustic Wave Propagation in Underwater Step Problems via the Method of Fundamental Solutions," European Acoustics Association, 2016.

[25] C. L. Pekeris, Theory of propagation of explosive sound in shallow water, Geological Society of America, 1948.

[26] R. P. Hodges, Underwater acoustics: Analysis, design and performance of sonar, John Wiley and Sons, 2011.

[27] Q. Yang, K. Yang, "Seasonal comparison of underwater ambient noise observed in the deep area of the South China Sea," Applied Acoustics, **172**, 107672, 2021, doi:10.1016/j.apacoust.2020.107672.

[28] C. Gervaise, S. Vallez, Y. Stephan, Y. Simard, "Robust 2d localization of low-frequency calls in shallow waters using modal propagation modelling," Canadian Acoustics, **36**(1), 153–159, 2008.

[29] F. Desharnais, M. Côté, C. J. Calnan, G. R. Ebbeson, D. J. Thomson, N. E. Collison, C. A. Gillard, "Right whale localisation using a downhill simplex inversion scheme," Canadian Acoustics, **32**(2), 137–145, 2004.

[30] M. H. Laurinolli, A. E. Hay, "Localisation of right whale sounds in the workshop Bay of Fundy dataset by spectrogram cross-correlation and hyperbolic fixing," Canadian Acoustics, **32**(2), 132–136, 2004.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 8, No. 4, 28-35 (2023)
www.astesj.com
Special Issue on Innovation in Computing, Engineering Science & Technology

ASTES Journal
ISSN: 2415-6698

Investigating the Impression Effects of a Teacher-Type Robot Equipped a Perplexion Estimation Method on College Students

Kohei Okawa[*,1], Felix Jimenez[2], Shuichi Akizuki[3], Tomohiro Yoshikawa[4]

[1]*Graduate School of Information Science and Technology,Aichi Prefectural University, 1522-3 Ibaragabasama, Nagakute-shi, Aichi, 480-1198, Japan*

[2]*School of Information Science and Technology,Aichi Prefectural University, 1522-3 Ibaragabasama, Nagakute-shi, Aichi, 480-1198, Japan*

[3]*School of Engineering, Chukyo University, 101-2 Yagoto Honmachi, Showa-ku, Nagoya, Aichi, 466-8666, Japan*

[4]*Faculty of Medical Engineering, Suzuka University of Medical Science, 1001-1, Kishioka, Suzuka-city, Mie, 510-0293, Japan*

ARTICLE INFO

Article history:
Received: 30 March, 2023
Accepted: 08 July, 2023
Online: 25 July, 2023

Keywords:
Deep Learning
The Educational support robot
The Impression
Collaborative learning

ABSTRACT

In recent years, the adoption of ICT education has increased in educational settings. Research and development of educational support robots have garnered considerable interest as a promising approach to inspire and engage students. Conventional robots provide learning support through button operations by the learners. However, the frequent need for button operation to request support may lead to a tedious impression on the learner and lower the efficiency of the learning process. Therefore, in this study, we developed a Perplexion Estimation Method that estimates the learner's state of perplexity by analyzing their facial expressions and provides autonomous learning support. We verified the impact of a teacher-type robot (referred to as the proposed robot) that autonomously provides learning support by estimating the learners' perplexity states in joint learning with university students. The results of a subject experiment showed that the impression of the proposed robot was not different from that of the conventional robot. However, the proposed robot demonstrated the ability to provide optimal support timing compared to the conventional robot. Based on these results, it is expected that the utilization of the perplexion estimation method with teacher-type robots can create a learning environment similar to human-to-human interaction.

1 Introduction

This paper is an extension of the one presented at SCIS [1]. In this conference, we presented the results of the life quality and the timing of support given to the learner by a supervised robot equipped with the proposed puzzling estimation method. In addition to these results, this paper provides additional analytical information on the results of subject experiments.

In recent years, the introduction of ICT education has become increasingly active in the field of education. The introduction of ICT education using educational big data that collects the learning status of individual students is being promoted in order to realize "fair, individualized and optimized learning that leaves no one behind" for children who have difficulty learning with other children due to reasons such as not attending school, children with developmental disabilities, and other children who are becoming increasingly diverse [2]. By utilizing information technology, it is possible to provide individualized educational support. In this ICT education, collaborative learning with robots that have a "presence" in the real world is shown to be effective in creating a learning environment where people can learn from each other, as well as in stimulating interest in learning [3]. It has also been reported that the robot's advise is superior to that of an on-screen agent [4]. Therefore, we believe that educational support by robots is more effective than on-screen agents in educational settings. For these reasons, research and development of robots that can play an active role in educational settings (hereinafter referred to as "educational support robots") has been attracting attention in Japan and abroad [5].

Educational robots include "teacher-type robots" that instruct learners like a teacher. The role of a supervised robot in conventional research is to teach the learner how to solve a problem[6]. For example, Yoshizawa et al, proposed a supervised robot that

[*]Corresponding Author: Kohei Okawa, Aichi Prefectural University, 1522-3 Ibaragabasama, Nagakute-shi, Aichi, 480-1198, JAPAN, 0561-76-8826, im222002@cis.aichi-pu.ac.jp

switches learning support according to the number of correct answers given by the learner. Experimental results showed that the robot that switches the learning support has the potential to provide high learning effectiveness for university students.

However, conventional supervised robots have a problem in that they cannot autonomously provide learning support in response to the learner's state of being unable to solve a problem (hereafter referred to as "perplexion"). For example, Yoshizawa et al.'s teacher-type robot provides learning support by pressing a button of the learner (hereinafter referred to as a "conventional robot"). The learner presses a button every time he/she needs learning support.

Conversely, enhancing the provision of support in learning systems is a critical focus in educational psychology research. In an environment where learners can request unlimited support, excessive use of hints [7] and search for hint patterns [8] occur regardless of the need for support. The use of artificial intelligence techniques has been reported to be effective in preventing these problems [9].

Therefore, the goal of this study creates a learning environment where a teacher-type robot provides autonomous learning support without the need for learners to press buttons, using deep learning techniques. We believe that it is effective for a teacher-type robot to estimate the learner's state of perplexion and autonomously provide learning support at the most appropriate timing, instead of providing learning support at the push of a button by the learner. Furthermore, when estimating the learner's state of perplexion, it is important to ensure that no burden is placed on the learner. We believe that this will prevent the learner from becoming dependent on the assistance and will enable smooth interaction with the robot, thereby improving the effectiveness of the teacher-type robot on the learner.

Regarding the learner's state estimation, Matsui has conducted previous research on autonomous learning support through perplexity state estimation [10]. Matsui's research attempted to estimate the learner's mental state by combining biometric devices, back-and-forth facial movements, and mouse movements. By using biological signals, we can obtain a state that is closer to the learner's raw data, which allows us to accurately estimate the perplexion. On the other hand, when using measurement equipment, the burden on the learner is large, and the data obtained is likely to contain noise in the real perplexity data. In the case of back-and-forth facial movements and mouse movements, the influence of the learner's posture and thinking habits is considered to be significant. Few studies have focused on estimating perplexion solely based on facial expression, despite the existing research on combining it with biometric signals.

In this study, we focus on research on human facial expression recognition [11] to provide autonomous learning support through teacher-type robots. In particular, methods based on deep learning have been widely used in research on facial expressions, and have shown high performance in image recognition and image classification Convolutional neural networks (hereafter, this is called CNN) have been proposed [12]. However, conventional research on facial expression recognition has focused only on the seven basic emotions of anger, disgust, fear, happy, sad, surprise, and neutral (hereafter referred to as the seven basic emotions), and has not focused on perplexion.

Therefore, in this study, we constructed the proposed method by transfer learning, using the seven basic emotion estimation methods of [13] as a base model. The proposed method is a perplexion estimation method that classifies two classes of perplexion state and non-perplexion state. In this study, we first extended the seven emotion estimation method to the eight emotion estimation method, including the perplexion state (67% recall).

Then, we conducted a subjective experiment on the impression effect of a teacher-type robot equipped with the eight emotion estimation method (hereinafter referred to as "the eight emotion robot") and discovered that it elicited a similar level of the impression as a conventional robot [14]. On the other hand, the accuracy of the robot was not sufficient, as some subjects commented in a questionnaire that "the timing of support is too fast" and "the robot is noisy when it repeatedly speaks. To solve this problem, we constructed a proposed method to improve the estimation accuracy (88% recall). Nevertheless, we have been unable to make a direct comparison between the impressions and support timing of the robot equipped with the proposed method and the eight emotion robot.

In this paper, we verify the impression effect on university students and the support timing of joint learning with a teacher-type robot that provides learning support autonomously (hereinafter referred to as "the proposed robot") equipped with the proposed method specialized for estimating the perplexion on a teacher-type robot. In the experiment, we will conduct a comparison between the proposed robot and the eight emotion robot.

TThis paper begins with a description of the proposed method in Chapter 2. Then, in Chapter 3, we verify the impression that the proposed robot gives to the learner and the timing of support through experiments with participants. Chapter 4 discusses the results and Chapter 5 summarizes them.

2 Perplexioin Estimation Method

Eight emotion estimation methods are designed to estimate the seven basic emotions together with the state of perplexion. However, including the estimation of the eight emotions reduces the accuracy of the estimation. Since the primary focus during learning is to determine whether the state is perplexed or not, estimating the eight emotions becomes unnecessary. In this chapter, we have developed a method for estimating perplexity using transfer learning, specifically focusing on estimating only the perplexed and non-perplexion states.

2.1 How to collect perplexed facial expression data

Perplexed state data were collected by capturing learners' facial expressions (at a resolution of 1920 and 1080 pixels, 30 frames per second) while they interacted with software related to the Technology Passport Examination (IT Passport) or software dealing with mathematical graphics difficulties and the Computer Aptitude Battery (CAB) [15] provided by SHL Japan. The participants in this study were university students, and the learning software was specifically designed to present them with moderately difficult problems that required careful thinking, based on their university lectures and high school mathematics knowledge.

The software presented a question with a hint button underneath. Participants were informed that they could press the button as many

times as they wanted to get pedagogical help. This facilitated the annotation of facial expression data during puzzles.

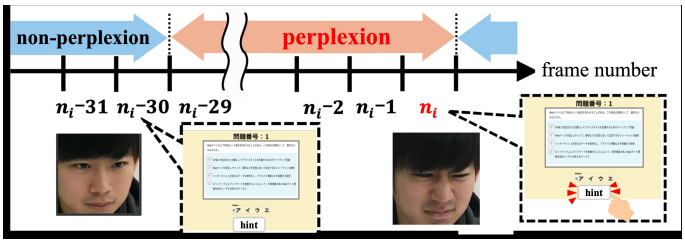


Figure 1: Definition of perplexion

In this study, facial expression data captured within 1 second (frames $n_i - 29$ at 30 frames per second) before the i-th button press (frame n_i) during learning with the learning software were categorized as perplexed state data, while other facial expression data were classified as non-perplexed state data (Fig. 1). The decision to define facial expression data up to 1 second ago as perplexed state data was based on the assumption that the expression of perplexity would be prominently evident during that period. Including facial expressions starting from 2 seconds ago may introduce variability in the strength of the puzzled state, leading to potential errors. Therefore, in this study, perplexed facial expressions are defined as the facial expressions that occurred 1 second before. The non-perplexed state is defined as the data collected from the frame immediately following the i-th button press (frame n_{i+1}) to the frame just before the onset of the perplexed state at the time of the i-th button press (frames n_{i+i} to $n_{i+i} + 30$).

2.2 Method Overview

In the field of deep learning, when the dataset available for training is small, transfer learning [16] is often utilized to leverage features learned by pre-existing models. In our study, in order to effectively capture subtle changes in facial expressions such as perplexion and non-perplexion, we employed seven basic emotion recognition models based on the FER2013 dataset as the base models (referred to as the "base model" in Fig. 2).

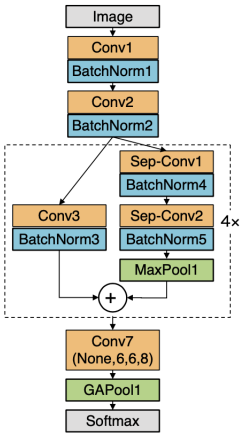


Figure 2: CNN model configuration

Specifically, we utilized the emotion recognition model using CNN developed by O. Arriaga et al [13]. By adapting the base model's features to focus on facial expressions, we anticipated an enhancement in the accuracy of perplexion estimation. Consequently, we constructed a dedicated model for perplexion estimation. In this model, the final layer generates the likelihood of two classes based on the perplexion state and the non-perplexion states. he perplexion estimation method of this paper was constructed using only facial expression data of college students. Therefore, the present method is specific to college students. To construct the perplexion estimation method, we utilized 36 face images out of the total of 52 collected in Section 2.1 as training data for perplexion/non-perplexion states. Of these, 26 images were used as training data (teacher data), while 10 images were kept aside for testing. Furthermore, since the method involves the binary classification of perplexion and non-perplexion states, we also included data from the FER2013 dataset representing other emotions into the non-perplexion data category (Table 1). Table 2 presents the number of data samples obtained from the aforementioned datasets.

The estimation of perplexion state is performed by inputting the face image to be estimated into a pre-trained model. The output of the model, denoted as y_c, represents the probability distribution over each class, with c indicating corresponding perplexion or non-perplexion state.

$$\arg\max_{1 \le c \le 2} y_c = 2 \tag{1}$$

Therefore, the state of perplexion is determined when the value of the formula (1) is "2", while the state of non-perplexion corresponds to all other cases. The process of estimating the perplexion state, referred to as "perplexion state estimation," commences after an average duration of time, denoted as x, has elapsed since the learner requests a hint. This approach prevents erroneous recognition that may occur when detecting, for example, a furrowed brow immediately upon the learner encountering a problem.

Table 1: Breakdown of the learning data.

	breakdown
Perplexion	Perplexion data collected
Non-Perplexion	Basic seven emotions from FER2013 Non-Perplexion data collected

Table 2: Configuration and number of learning data

Data Name	Data source	Number of data
anger	FER-2013	1997
disgust	FER-2013	218
fear	FER-2013	2048
happy	FER-2013	3607
sad	FER-2013	2415
surprise	FER-2013	1585
neutral	FER-2013	2482
non-perplexion	By Section 2.1	3504
perplexion	By Section 2.1	3475

To illustrate, the perplexion state estimation initiates after 30 seconds if the average time is set to be 30s.

3 participant experiment

3.1 Robot overview

For the experiment, we utilized Tabot (Figure 3), a tablet-type robot with a tablet serving as its head, capable of displaying agents and expressing different facial expressions. Tabot's body consists of 3 degrees of freedom (DOFs) in the neck, 5 DOFs in an arm, and 1 DOF in the legs, for a total of 14 DOFs that allow different body movements.

In this experiment, as shown in Figure 3, a camera was installed on the robot's head to capture facial expressions during the learning process. The camera records the student's facial expressions as they look at the tablet screen during the learning process. The camera footage is regulated and processed by a processing PC. If a state of perplexion is detected from the captured facial expression data, an instruction is sent to Tabot via Local Area Network communication. The camera has a resolution of 640 and 480 pixels and a frame rate of 30 frames per second.

3.2 Learning system

Participants study using the learning system shown in Figure 4. In the pre-training phase, subjects study, only with the learning system installed on their PCs, while in the collaborative phase, they study with the learning system displayed at the bottom of the tabot. The learning system displays the screen shown in Figure 4(a) when the subject proceeds to study. In Figure 4(a), there is a question and a button below the question that provides a hint. We informed the participants that they could press the button as many times as they wanted and that they would receive a hint by pressing the button. This makes it possible to label the facial expression data in the perplexoin. However, the hint button is installed only in the pre-training phase, and is removed in the collaborative learning phase. When an answer is given on the screen shown in Figure 4(a), the user is taken to the screen shown in Figure 4(b). The system repeats this process for the number of questions to reach the final screen (Figure 4(c)).

The learning system comprises two types of challenges that are designed to involve a puzzle-like element or spark.

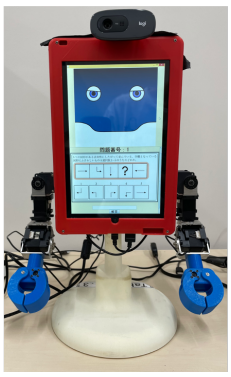


Figure 3: Tabot and camera

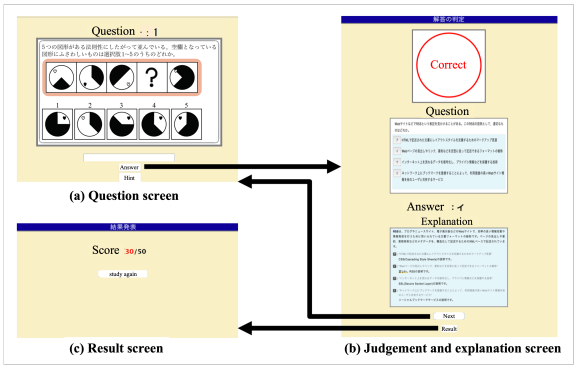


Figure 4: Learning system

These challenges consist of mathematical figure problems and Computer Aptitude Battery (CAB) problems provided by SHL Japan [15]. The mathematical figure problems can be effectively tackled by utilizing auxiliary lines in the drawing process, while the CAB problems involve identifying patterns and regularities. By presenting these challenges that demand both inspiration and puzzle-solving skills, we created an environment in which participants were consistently perplexed.

Table 3: Experimental information

Item	Contents
Property	Undergraduate and graduate students
The number of participants	34
Male:Female	20:14
Period	2020/9/14 ~1/30 2021/5/6 ~6/30 2022/2/17 ~3/17

3.3 Experimental procedure

Experiments were conducted in which undergraduate and graduate students and robots learned together. Experimental information is summarized in the table3 below.

In the experiment, we compare two groups of teacher-type robots equipped with different emotion recognition models. The groups to be compared are the eight emotion group equipped with the eight emotion estimation method and the proposed group equipped with the perplexion estimation method.

The information for the groups is shown in Table 4. For the proposed group, the estimated start time is set to 106 seconds after the first press of the hint button. This is because the average time elapsed until the first press of the hint button is 106 seconds during the experimental period up to 2021. This experiment is divided into three-time periods due to the number of participants. The proposal group was conducted in 2022, and the eight emotion group was conducted in the other dates.

The experimental procedure is shown below.

Step 1 (Pre-learning) Participants learn with a computer-based learning system. The learning time was about 60 minutes.

Table 4: Group information

	Eight emotion group	Proposed group
emotion recognition model	Eight emotion estimation method	Perplexion estimation method
hint provide condition	When estimating the perplexion	
Support frequency limitation	No limit	two times
hint button	nothing	
Estimated start time	Immediately after the problem is presented	After 106s

Step 2 To reduce the learning effect, we allowed a one-week interval. In addition, the groups were assigned so that the number of times, the participants pressed the hint button in the pre-learning period would be as equal as possible.

Step 3 (Collaborative learning) Participants learned with a robot from the group to which they were assigned. The learning time was about 60 minutes.

Step4(Survey) Immediately after the collaborative learning the participants answer a survey.

3.4 Evaluation index

We employed the ”Godspeed Questionnaire [17, 18],” a questionnaire methodology designed for the subjective evaluation of human-robot interaction. One of the questionnaire items utilized in our study was focused on assessing ”Animacy.” Additionally, we used the questionnaire to evaluate the timing of the robot's assistance. We selected these evaluation indices based on their perceived significance in the context of teacher-type robots. Previous research suggests that users tend to be more emotionally engaged and influenced by objects exhibiting animacy [19]. Therefore, a higher level of animacy in the robot may lead to a greater receptiveness to advise provided by the teacher-type robot. Furthermore, the autonomous timing of the robot's assistance is crucial for creating a more human-like learning environment. Each adjective pair in Godspeed Questionnaire was rated on a 5-point scale, and we quantified each pair on a scale of 1 to 5, with the positive adjective side receiving a higher score. We defined the average score of six animacy items as ”animacy” and compared it across different groups. Figure 5 illustrates the questionnaire pertaining to support timing.

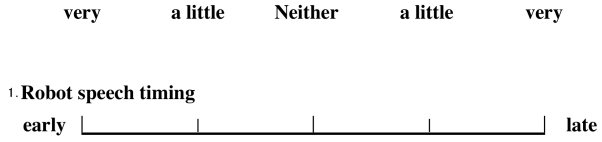


Figure 5: Questionnaire on support timing

Table 5: Scoring Procedure

survey	score
3	3
2 or 4	2
1 or 5	1

To analyze the evaluation results, we employed a Student's t-test for animacy, assuming equal population variance across the groups. Regarding the timing of support, two types of analyses were performed. The first analysis involved testing the population mean when the population variance was unknown, using ”3” for ”neither” as the optimal timing criterion. The second analysis utilized a Student's t-test with the timing questionnaire as the scoring criterion. Details of the scoring procedure can be found in Table 5. The significance level was adjusted to $p < 0.0125$ using the Bonferroni correction to account for multiple comparisons, as four tests were conducted. The tests included comparisons of animacy, the average support timing between the eight emotion group and the proposed group, and comparison of support timing scores.

3.5 Result

Figure 6, 7, and 8 show the mean values of lifeliness and support timing, the mean values when scoring, and the test results. The proposed group had higher animacy than the eight emotion group. The test results showed no significant differences. These results suggest that the animacy that participants felt toward the robot was the same in each group of robots.

The support timing results for the “eight emotion group” and the “proposed group” were both found to be early and close to the criterion value of “3”. The statistical analysis revealed a significant difference between the eight emotion group and an insignificant difference between the proposed group. This indicates that the eight emotion group was perceived as providing support too early, while the proposed group offered support at an appropriate timing. Moreover, employing the perplexion estimation method for support timing could lead to optimal timing of assistance. The results of scoring the timing of support were higher in the suggestion group than in the eight-emotion group.

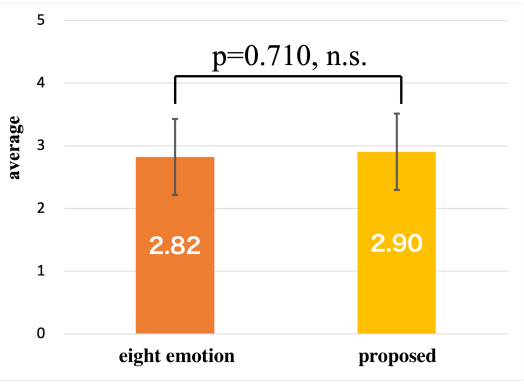


Figure 6: Result of Animacy

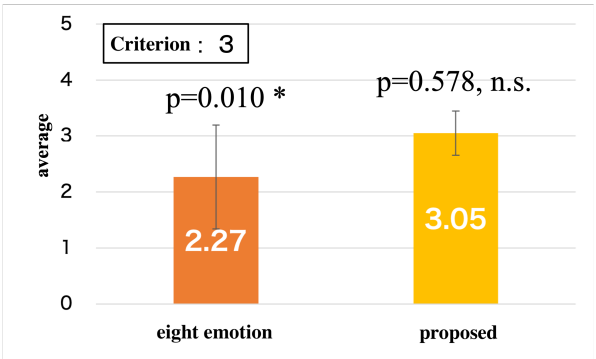


Figure 7: Result of Support timing

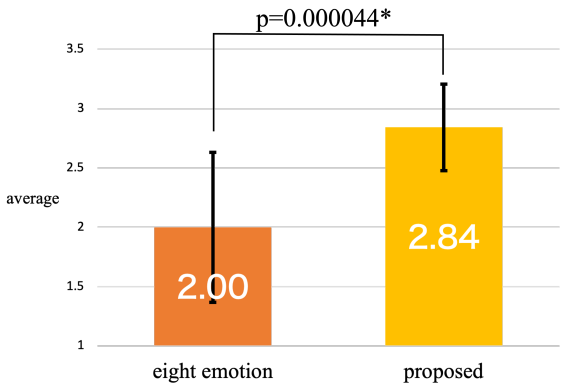


Figure 8: Result of support timing score

The results of the test showed that there was a significant difference. Therefore, the results indicated that the proposed group was better in support timing.

4 discussion

The experimental results indicated that there was no significant difference in animacy, suggesting that the learners had similar impressions regardless of the group they were in. However, a significant difference was found in the support timing between the eight emotion group and the proposed group. In the one-group test based on the criterion value of "3", the eight emotion group showed a timing that was too early for the learners, while the proposed group provided support at the optimal timing. Furthermore, a significant difference was also found in the test when the questionnaires were converted into scores. These differences will be discussed in the following sections, along with the results of each questionnaire.

Animacy was higher in the proposed group than in the eight emotion group. We believe that this may be related to the number of speech utterances during collaborative learning. To investigate this, we examined the number of speech utterances for six participants in the eight emotion group and nineteen participants in the proposed group. The average number of utterances per participant was 29 for the eight emotion group and 0.2 for the proposed group. These results suggest that too many speech utterances may impair Animacy.

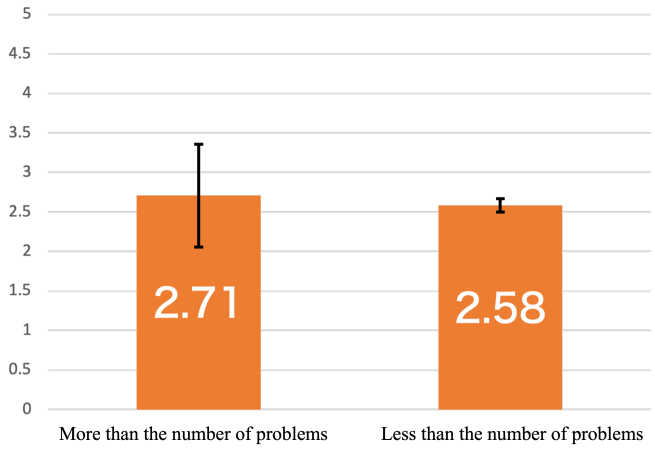


Figure 9: Animacy when divided by the number of support (The eight emotion group)

To further explore this, we investigated the Animacy scores of participants in the eight emotion group whose the number of speech utterances exceeded the number of questions. Four out of the six participants exceeded the number of questions, and we calculated the average Animacy scores of these four participants and the other two participants (Figure 9). Contrary to our expectations, Animacy was higher for those who received more support. However, due to the small number of participants, we cannot conclude with certainty that there is a clear trend. One possible explanation is that repeated utterances made the participants feel that there was a response.

In the Animacy section, there is a "responsiveness" column. The average score of the four respondents who received more support was 3.5, while that of the two respondents who received less support was 2. This result also suggests that a higher number of speech utterances may influence Animacy, as the respondents feel more responsive. However, the average response to the questionnaire regarding the timing of support was 1.75 when the number of times support was high and 2 when the number was low. Based on the open-ended responses from participants who had a high frequency of speech utterances, some of them expressed that there were "excessive number of hints" and they felt "confused by the continuous stream of advice". Therefore, we believe that an excessive number of speech utterances does not have a positive effect.

Similarly, in the proposed group, the mean score of Animacy was calculated for four of the 19 participants for whom learning support was confirmed and for 15 participants for whom it was not confirmed (see Figure 10). Animacy scores were higher for the four participants for whom learning support was confirmed. Additionally, as in the eight emotion group, we calculated the mean score for each "responsive" item. The mean score for the four participants for whom learning support was confirmed was 4, and for the 15 participants for whom learning support was not confirmed, it was 3.3. Based on the analysis of the eight emotion group, we believe that having learning support is better than not having it, but that frequent learning support, such as continuous speech, may not give a good impression.

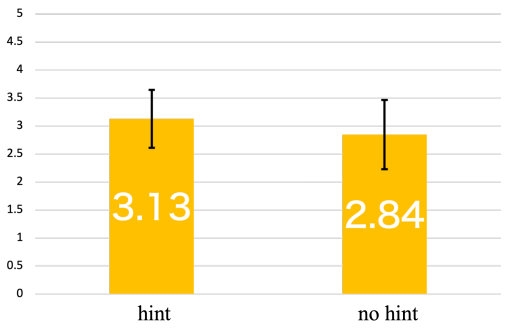


Figure 10: Animacy when divided by the number of support (The proposed group)

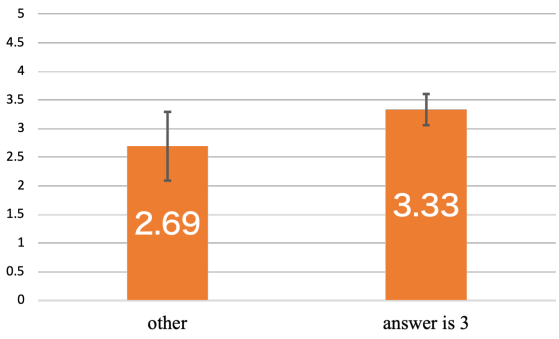


Figure 11: Animacy by support timing answers (eight emotion group)

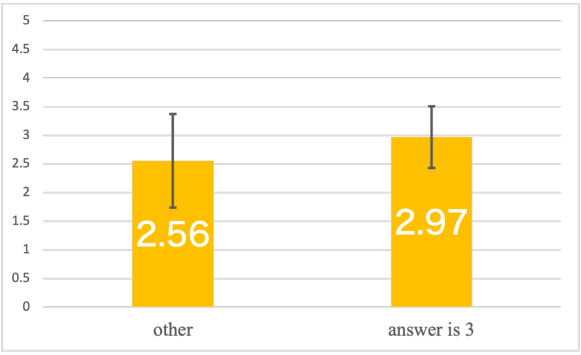


Figure 12: Animacy by support timing answers (proposed group)

We believe that the timing of support is an essential factor in autonomous support provided by a teacher-type robot. If support is given too early, learners may feel frustrated that it is offered at a time when they do not need it, whereas if it is provided too late, it may cause learners to feel frustration at not receiving support and not being able to solve problems. The experimental results showed a significant difference between the eight emotion group and the ideal timing of "3". Therefore, it is thought that the participants felt that the support was premature. However, no significant difference was observed in the proposed group. This suggests that the support timing used in the proposed group may be the optimal timing. The test of scoring the timing of support showed a significant difference, with the proposed group having a higher score than the eight emotion group, indicating that the proposed group was superior. However, as seen in the previous data, the eight emotion group tends to talk too much, resulting in a lower score. Similarly, the proposed group may not have been judged as early or late because they provided support significantly less frequently. Therefore, the results suggest that the perplexion estimation method used to solve this problem led to good results, as the evaluation of too many utterances was low.

Animacy and timing results suggest that teacher-type robots that provide appropriate timing and frequency of support can improve Animacy. Figures 11 and 12 show the average Animacy scores of participants who answered "3" in the questionnaire about support timing, and those who answered "other." Three participants in the eight emotion group and 16 participants in the proposed group answered "3" for support timing, and in both cases, those who answered "3" had higher scores. These results indicate that refining support timing is effective in improving Animacy. To achieve this, we believe that the accuracy of the perplexity estimation method needs improvement. The current method has a reproducibility of 88%, but its accuracy is 67%, which is inferior to the eight emotion estimation methods. We believe that improving accuracy is necessary to provide an environment that can estimate the learner's perplexity more accurately without missing it.

5 Conclusion

Our study developed a deep learning-based method to estimate the learner's state of perplexion. We conducted experiments to compare the performance of a robot equipped with our proposed method to a robot equipped with the conventional eight emotion estimation methods. The results of the participant experiment revealed that there was no significant difference in terms of animacy between the two groups. This indicates that our proposed method provided a similar level of animacy to the learners compared to the conventional robot. However, when it comes to support timing, the robot equipped with our proposed method demonstrated the ability to provide support at the optimal timing. Furthermore, when the support timing was scored, a significant difference was observed, indicating that the support timing of the robot equipped with the proposed method was better. In other words, the proposed method can potentially enable the realization of a collaborative learning environment between a teacher-type robot and a learner without the need for buttons. Overall, we believe that this study has demonstrated the potential for realizing a collaborative learning environment between a robot and a learner using the proposed perplexion estimation method.

In the future, our plan is to develop a robot suitable for actual learning environments for junior high school and high school students, and to evaluate the impression the proposed robot leaves on learners as well as its effectiveness in supporting learning. To achieve this, we intend to conduct participant experiments that evaluate both the impression of the robot and its impact on learning. Moreover, to create a more effective learning environment for the learners, it is essential to enhance the accuracy of the perplexion estimation method. To achieve this, we are considering two approaches. The first approach is to fine-tune the method by training it on the data collected from the planned experiments, so that it can accommodate a wide range of age groups. This will enable us to effectively support learners with youthful facial features, such

as middle and high school students, who are expected to be actual users of the system. The second approach is to treat the learning data as time series data, including changes in facial expressions. Currently, we only focus on the moments of perplexion. However, by capturing the trends of perplexion onset, we anticipate that we can improve the estimation accuracy to a greater extent.

Conflict of Interest The authors declare no conflict of interest.

References

[1] K. Okawa, F. Jimenez, S. Akizuki, T. Yoshikawa, "Effects of a Teacher-type Robot Using a Perplexion Estimation Method on College Students," in 2022 Joint 12th International Conference on Soft Computing and Intelligent Systems and 23rd International Symposium on Advanced Intelligent Systems (SCIS&ISIS), 2022, doi:10.1109/SCISISIS55246.2022.10001988.

[2] Ministry of Education, Culture, Sports, Science and Technology, "Measure to Promote the Use of Adbanced Technology to Support Learning in the New Era (Final Summary)," 2019, doi:https://www.mext.go.jp/a_menu/other/1411332.htm.

[3] T. Kanda, "How a communication robot can contribute to education?" Journal of Japanese Society for Artificial Intelligence : AI, **23**(2), 229–236, 2008, doi:10.11517/jjsai.23.2_229.

[4] K. Shinozawa, F. Naya, J. Yamato, K. Kogure, "Differences in Effect of Robot and Screen Agent Recommendations on Human Decision-Making," International Journal of Human-Computer Studies, **62**(2), 267–279, 2005, doi:10.1016/j.ijhcs.2004.11.003.

[5] T. Belpaeme, J. Kennedy, A. Ramachandran, B. Scassellati, F. Tanaka, "Social robots for education: A review," Science Robotics, **3**, 2018, doi:https://doi.org/10.1126/scirobotics.aat5954.

[6] R. Yoshizawa, F. Jimenez, K. Murakami, "Proposal of a Behavioral Model for Robots Supporting Learning According to Learners'Learning Performance," Journal of Robotics and Mechatronics, **32**(4), 769–779, 2020, doi:https://doi.org/10.20965/jrm.2020.p0769.

[7] V. Aleven, K. R. Koedinger, "Limitations of student control:Do students know when they need help?" in Proceedings of the 5th International Conference on Intelligent Tutoring Systems, 2000, doi:10.1007/3-540-45108-0_33.

[8] J. A. Walonoski, N. T. Heffernan, "Detection and Analysis of Off-Task Gaming Behavior in Intelligent Tutoring Systems," in Proceedings of the 8th International Conference on Intelligent Tutoring Systems, 2006, doi:10.1007/11774303_38.

[9] I. Roll, V. Aleven, B. M. McLearen, K. R. Koedinger, "Improving students' help-seeking skills using metacognitive feedback in an intelligent tutoring system," Learning and Instruction, **21**(2), 267–280, 2011, doi:10.1016/j.learninstruc.2010.07.004.

[10] T. Matsui, "Estimation of Learners' Physiological Information and Learners' Mental States by Machine Learning and Its Application for Learning Support," Transactions of Japanese Society for Information and Systems in Education, **36**(2), 76–83, 2019, doi:https://doi.org/10.14926/jsise.36.76.

[11] M. Wimmer, B. A. MacDonald, D. Jayamuni, A. Yadav, "Facial Expression Recognition for Human- Robot Interaction - A Prototype," in In Proceedings of the 2nd International Workshop on Robot Vision, 2008, doi:10.1007/978-3-540-78157-8_11.

[12] T. Nishime, S. Endo, K. Yamada, N. Toma, Y. Akamine, "Feature Acquisition From Facial Expression Image Using Convolutional Neural Networks," Journal of Robotics, Networking and Artificial Life, **32**(5), 1–8, 2017, doi:http://dx.doi.org/10.2991/jrnal.2016.3.1.3.

[13] O. Arriaga, P. G. Ploger, M. Valdenegro, "Real-time Convolutional Neural Networks for Emotion and Gender Classification," 2017, doi:10.48550/arXiv.1710.07557.

[14] K. Okawa, F. Jimenez, K. Murakami, S. Akizuki, T. Yoshikawa, "A Survey on Impressions of Collaborative Learning with an Educational Support Robot Estimating Learners' Perplexion," in Proceedings of the 37th Fuzzy Systems Symposium, 171–176, 2021, doi:10.14864/fss.37.0_171.

[15] SHL Japan products, 2020/12/06 accessed, doi:http://www2.shl.ne.jp/product/index.asp?view=recruit.

[16] A. Gulli, S. Pal, Chokkan Deep learning: Python×Keras de aidea o katachi ni suru reshipi, O'Reilly Japan, 2018, doi:https://www.ohmsha.co.jp/book/9784873118260/.

[17] C. Bartneck, D. Kulic, E. Croft, S. Zoghbi, "Measurement Instruments for the Anthropomorphism, Animacy, Likeability, Perceived, Intelligence, and Perceived Safety of Robots," International Journal of Social Robotics, **1**(1), 71–81, 2009, doi:https://psycnet.apa.org/doi/10.1007/s12369-008-0001-3.

[18] T. Nomura, "Humans' Subjective Evaluation in Human-Agent Interaction(HAI)," Journal of the Japanese Society for Artificial Intelligence, **31**(2), 224–229, 2016, doi:https://doi.org/10.11517/jjsai.31.2_224.

[19] B. J. Fogg, Persuasive Technology: Using Computers to Change what We Think and Do, Morgan Kaufmann, 2002, doi:10.1016/B978-1-55860-643-2.X5000-8.

Advances in Science, Technology and Engineering Systems Journal Vol. 8, No. 4, 36-40 (2023)

www.astesj.com

Special Issue on Computing, Engineering and Multidisciplinary Sciences

ASTESJ

ISSN: 2415-6698

Fuzzy MPPT for PV System Based on Custom Defuzzification

Abdelmadjid Allaoui*, Mohamed Nacer Tandjoui, Chellali Benachaiba

Electrical Engineering Department, Faculty of Science and Technology, Tahri Mohamed University, Bechar, 08000, Algeria

ARTICLE INFO

Article history:
Received: 20 January, 2023
Accepted: 07 May, 2023
Online: 25 July, 2023

Keywords:
PV system
Fuzzy MPPT
Defuzzification

ABSTRACT

Due to the variations in weather conditions, photovoltaic systems adopt a technique based on maximum power point tracking to extract the maximal power of the solar module. In the literature, there are many different methods classical and intelligent of maximum power point tracking (MPPT). But, due to the semiconductor effect, the current-voltage characteristics of the solar module is nonlinear. This affects its efficiency and make its control not easy. In this contribution, we present a new fuzzy PV MPPT based on custom defuzzification. The obtained power using the proposed fuzzy PV MPPT based on custom defuzzification is significant compared to Pertub & observe and fuzzy PV MPPT in term of performances indices such as: Rise time and overshoot.

1. Introduction

Generally, energy is an important development factor in any economy. Also, energy consumption is a progress indicator. The energy crisis due to the drop of conventional energy sources and the rise in CO_2 emission and environmental pollution has imposed the search for other solutions which are renewables and cleans. As renewable source, photovoltaic power is a very powerful and promising energy potential. The solar energy is converted into electrical energy by solar PV panel. Each type of PV panel has its own specific characteristic according to local conditions such as irradiation, and temperature and this makes the tracking of maximum power point (MPP) a complicated problem. To remedy this problem, many MPPTs algorithms have been presented [1-6]. Conventional MPPTs have proven to be less efficient because of the functioning principle of photovoltaic system which depends on weather.

This, made extracting the maximum power point a difficult task. But, with the development of semiconductor switches which work with high switching frequencies; new MPPT controllers have been developed. Among the intelligent MPPTs, there is the logic-based MPPT which has interested several researchers. In the literature, there are several publications on this fuzzy MPPT controller with more approaches [7-13]. In this context, we will present an intelligent MPPT based on fuzzy logic but which is different with a custom defuzzification function because most of the papers dedicated to this field use default and predefined defuzzification functions.

2. Modeling of photovoltaic system

Photovoltaic system works on the photovoltaic effect to convert solar radiation into direct current. When the sun attacks the panel, the energy is absorbed by a semiconductor. This energy will release the electron-hole pairs from their binding state to supply the load of the photovoltaic system. The output power of the photovoltaic panel depends on environmental variables such as irradiation and temperature. Therefore, to operate the PV system at its maximum power point; the MPPT mechanism is very important and useful. Many MPPT mechanisms have been introduced in the literature by many researchers since the year 1960. Some well-known conventional MPPT methods are incremental conductance, perturbation and observation and constant voltage.

However, this type of method presents classical and limited algorithms. Their implementation requires a good, accurate sensor to measure either voltage or current. Recently, the MPPT based on artificial intelligence is widely used in PV system. These intelligent MPPTs are dynamic with high efficiency and have made the PV system interesting and competitive. Figure 1 shows the block diagram of the proposed standalone PV system. The system consists of a PV array, a MPPT controller combined to a DC- DC converter and a load.

The irradiation (G) and temperature (T) are in charge of the working point of PV system at the maximum power point (MPP) [14]. The cell current, I, which represents the mathematical model of the PV cell can be express as:

*Corresponding Author Abdelmadjid Allaoui, allaoui.abdelmadjid1978@gmail.com

https://dx.doi.org/10.25046/aj080405

$$I = I_{ph} - I_0\left(e^{\left(\frac{q(V+IR_s)}{A.K_cT}\right)} - 1\right)\frac{V+IR_s}{R_{sh}} \quad (1)$$

where, I_{ph} is light-generated cell current (A), I_0 is cell reverse saturation current (A), q is electronic charge, A is ideality factor, K_c is Boltzmann's constant, and T is cell temperature (K).

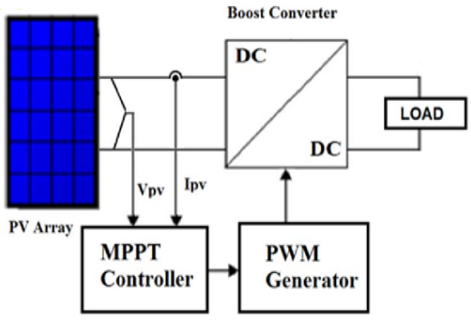


Figure 1: Block diagram of the PV system

3. Perturb & Observe MPPT

Perturb & Observe algorithm is a conventional method. It is used in photovoltaic systems because of its simple implementation. Also, it needs a few measured parameters. It is based on the measure of the PV current and voltage. From these values the power is calculated at each time to find out the maximum power point (MPP).

The principle of this algorithm is based on the operating voltage of the PV module which is perturbed by a small increment and the change of power is observed. If the change of power is positive, then it is supposed that it has moved the operating point closer to the MPP. So, the voltage disturbed in the same track should move the operating point toward the MPP. If the change of power is negative, the operating point has moved away from the MPP. In this case, the direction of perturbation should be reversed.

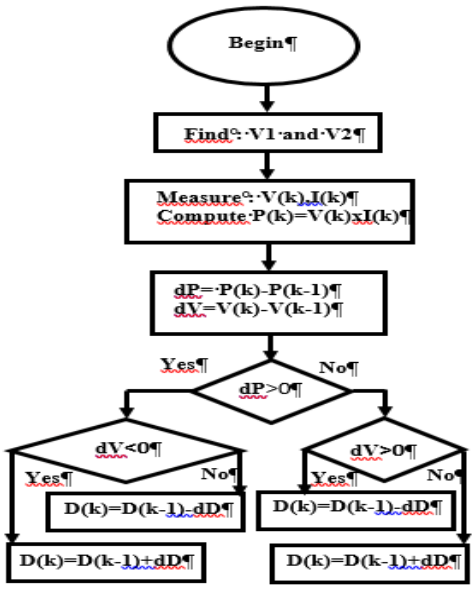


Figure 2: Flowchart of P&O method

4. Fuzzy MPPT based custom defuzzification

The MPPT allows the PV system to work at the maximum despite the variation of its parameters, irradiation, temperature and load. Conventional MPPT methods are limited, however, MPPT based on fuzzy logic fuzzy offers the advantage of being robust, efficient and works to the PPM. The implement of the fuzzy MPPT has three steps: the fuzzification, inference engine and defuzzification (Figure 3).

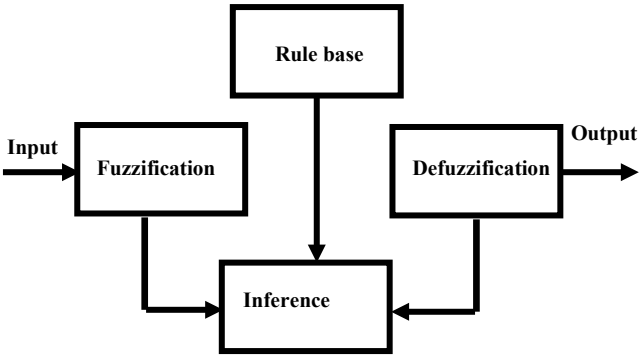


Figure 3: Fuzzy MPPT block diagram

4.1 Fuzzification method

Fuzzification is a method by which sharp values are blurred. To do this, the linguistic variables and the membership functions that will be implemented to model the system must be defined. The principle of fuzzification consists in the decomposition of the universe of discourse of linguistic input and output variables into a number of membership functions.

4.2 Inference engine

Inference is a method by which new information is deduced from the information of premises. Inference in fuzzy logic control systems is a method by which the result of each rule is deduced from the results of each activated rule.

In the literature, dedicated to fuzzy logic, different methods that can be applied to establish an inference engine. The most popular are Mamdani and Takagi-Sugeno-Kang. The Mamdani inference was developed by Ebrahim H. Mamdani in 1975. It was used to modify the behavior of a steam engine. Mamdani's inference was inspired from Lofti Zadeh's paper describing fuzzy sets for systems. For Mamdani max-min inference, the minimal operation is used at implication stage, while the max-min operator is employed to the premises.

4.3 Defuzzification

$$z_{COA} = \frac{\int_z \mu_A(z)zdz}{\int_z \mu_A(z)dz} \quad (2)$$

where, z_{COA} is the control output.

The fuzzy MPPT has two inputs: error (err) and the variation of the error (derr) which are defined by the equations (3) and (4).

$$err = \frac{P_{pv}(\mathrm{k}) - P_{pv}(\mathrm{k-1})}{V_{pv}(\mathrm{k}) - V_{pv}(\mathrm{k-1})} \quad (3)$$

$$derr(k) = err(k) - err(k-1) \quad (4)$$

where, $P_{pv}(k)$ and $V_{pv}(k)$ are respectively the output power and instantaneous voltage of the photovoltaic source.

The inference rules are used to evaluate the linguistic values of the activated rules according to their membership degrees. For Mamdani inference, the inference of each activated rule gives a surface. The aggregation of these surfaces gives a final surface which by the defuzzification generates the value of the modulator. The defuzzification allows the conversion of membership degree in crisp value.

Each linguistic variables of the input and the output of the proposed fuzzy MPPT based on custom defuzzification has five membership functions. The used linguistic values are: NB (Negative Big), NS (Negative Small), ZE (Zero), PS (Positive Small) and PB (Positive Big).

The majority of researchs uses Matlab's predefined defuzzification functions for fuzzy MPPT controller. Through this contribution, we will propose a custom function of defuzzification which will boost the action of the fuzzy MPPT controller.

$$Z_{Custom} = \frac{\sum_{i=0}^{n} k\ ymf_i\ x_i}{\sum_{i=0}^{n} ymf_i^{\ 2}} \quad (4)$$

where, k is a gain that regulates the rise time.

5. Simulation and discussion

The proposed fuzzy PV MPPT based on custom defuzzification is compared to P&O MPPT. The simulation system is presented below.

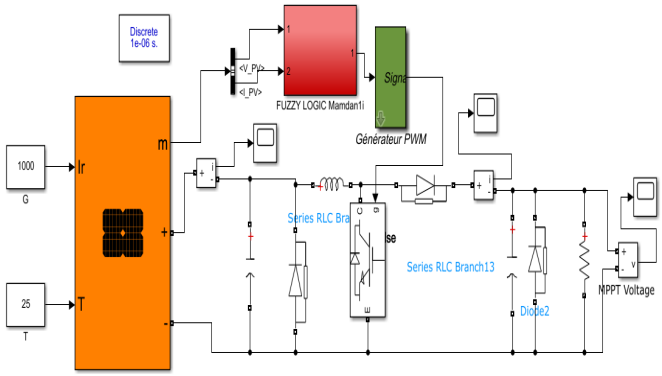


Figure 4: Simulation system under Matlab Simulink

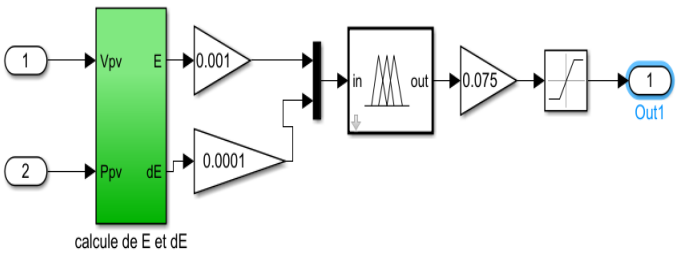


Figure 5: Fuzzy MPPT controller design

Figure 5 shows the design of fuzzy PV MPPT controller which has two inputs: Error and change of error according to "(2)" and "(3)". Table 1 presents the electrical characteristic values of the PV module.

Table 1: Electrical characteristics of the PV module

Electrical characteristics	values
Maximum power (W)	85.383
Cells per module (Ncell)	36
Open circuit voltage Voc (V)	22
Short circuit current Isc (A)	5.2
Voltage at maximum power point Vmp (V)	17.9
Current at maximum power point Imp (A)	4.77

Figures 6-9 show simulation steps of fuzzy controller for PV MPPT with custom defuzzification function. The inputs to the fuzzy controller are error and error variation of the PV system. The error is the ratio between the variation of the power on the variation of the PV voltage. The output of the fuzzy controller represents the modulator which will modulate the carrier of the PWM generator to produce the PWM signal which will drive the switch of the boost converter according to the inferences of fuzzy controller for PV MPPT.

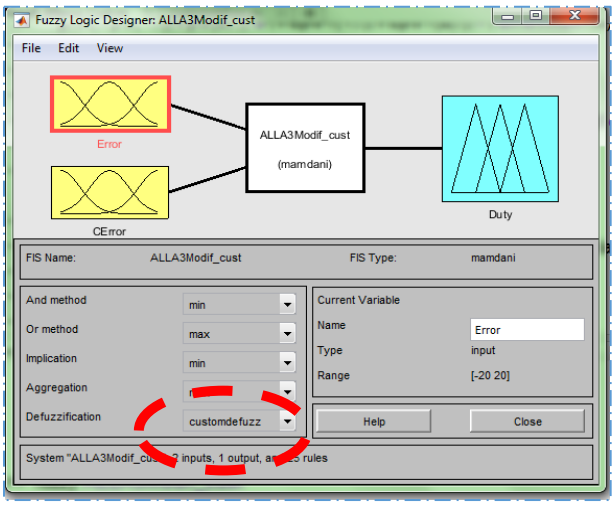


Figure 6: Fuzzy logic designer

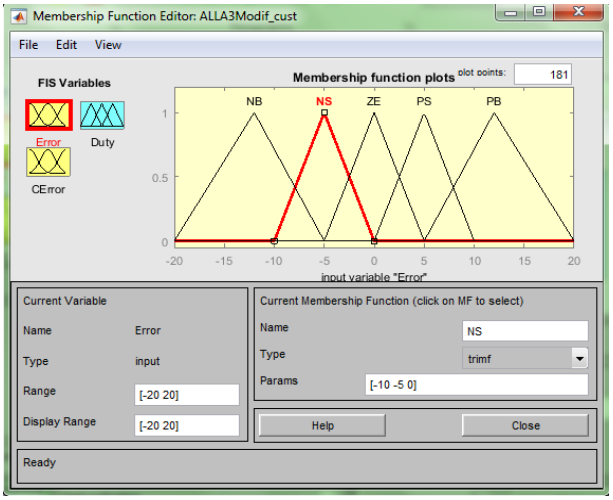


Figure 7: Error membership functions

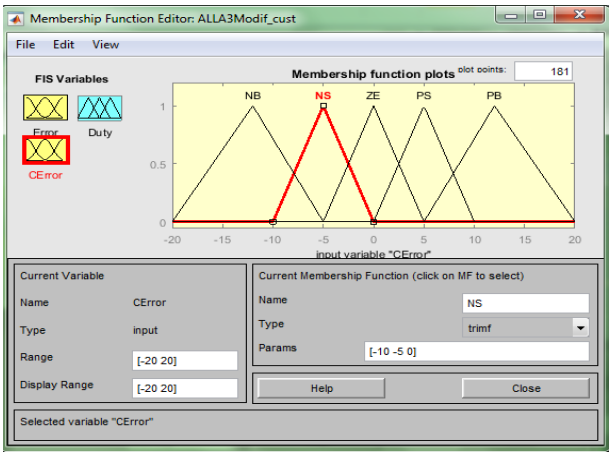


Figure 8: Change of error membership functions

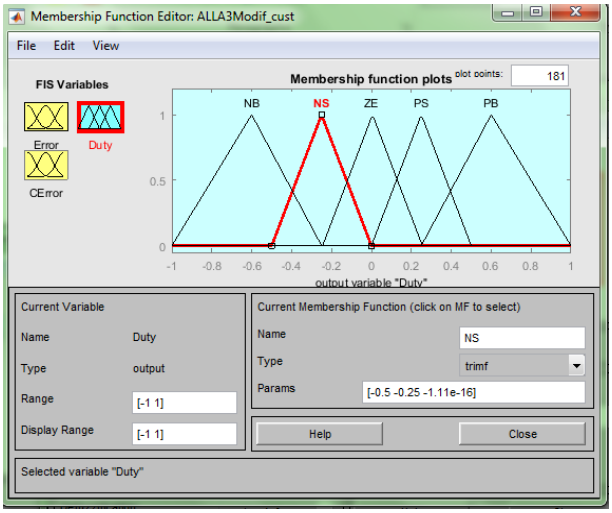


Figure 9: Output membership functions

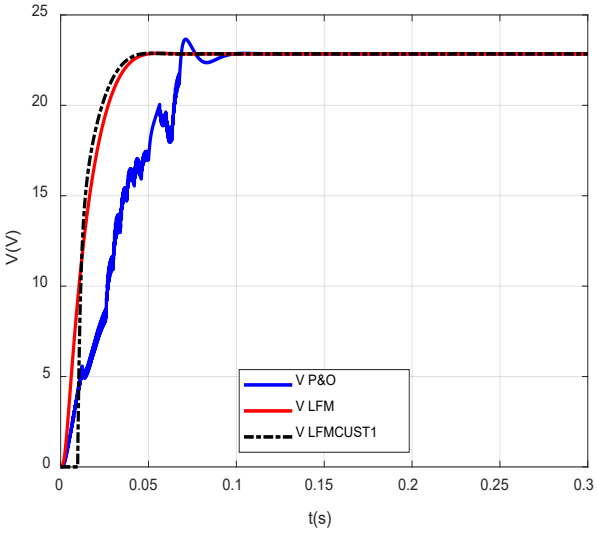


Figure 10: Load voltages

Figures 10-15 present the results obtained from the simulated system which are the voltage, the current and the power of the load. Each figure presents the comparison of the three PV MPPTs which are P&O, fuzzy with Mamdani inference and centroid defuzzification and the response of the proposed fuzzy PV MPPT with Mamdani inference and custom defuzzification. The contribution of fuzzy PV MPPT is interesting. The response of the proposed fuzzy PV MPPT with Mamdani inference and custom defuzzification is better than the other MPPTs. Table 2 shows the load power performance of the three MPPTs, and the response of the proposed fuzzy MPPT for PV system with Mamdani inference and custom defuzzification presents less rise time and overshoot than other MPPTs.

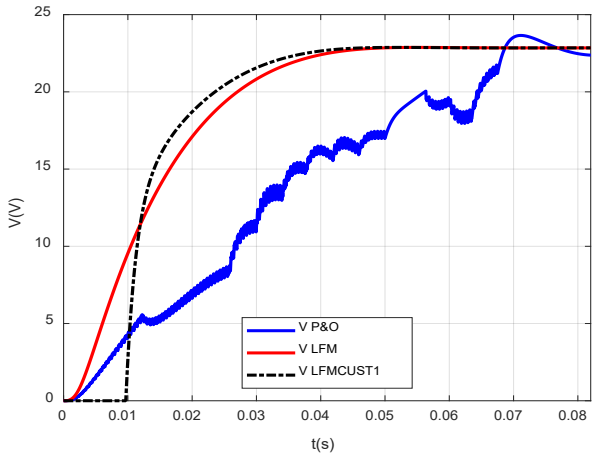


Figure 11: Zoom of load voltages

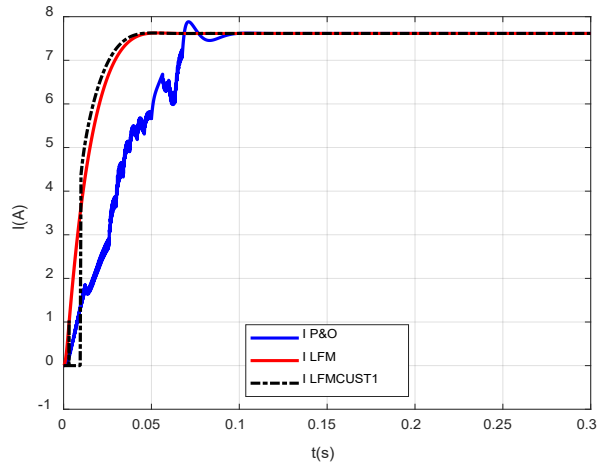


Figure 12: Load currents

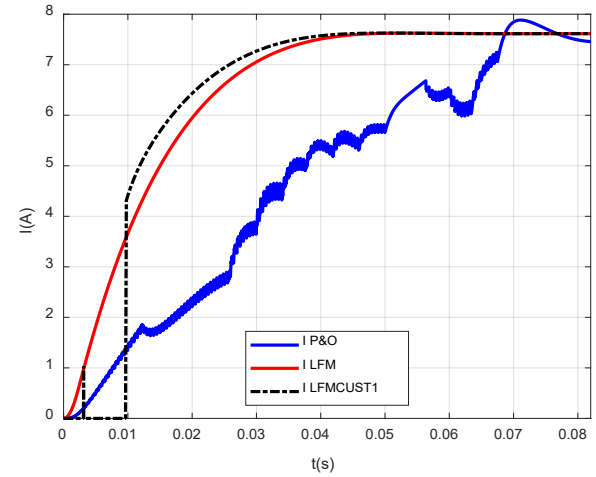


Figure 13: Zoom of load currents

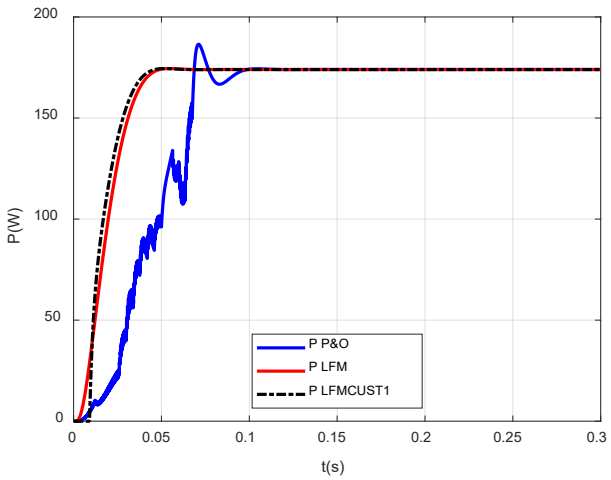


Figure 14: Load powers

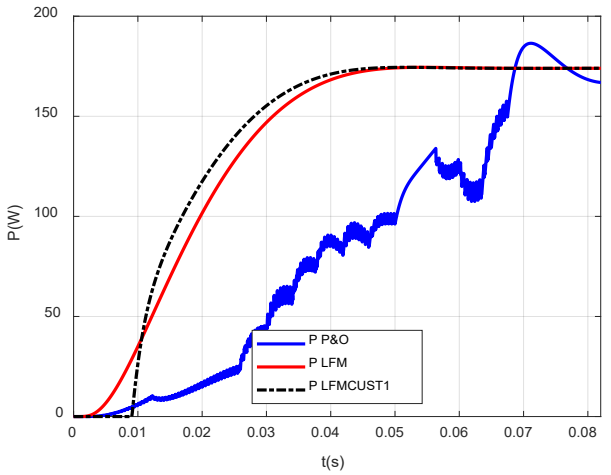


Figure 15: Zoom of load powers

Table 2: Performance of load power responses

Performance indices	P&O	Fuzzy MPPT with Centroid defuzz.	Fuzzy MPPT with Custom defuzz.
RiseTime	0.0470	0.0262	0.0209
SettlingTime	0.0902	0.0422	0.0393
SettlingMin	149.7038	156.5957	156.5957
SettlingMax	186.4765	174.4732	174.4730
Overshoot	7.1669	0.2762	0.2760
Peak	186.4765	174.4732	174.4730
PeakTime	0.0712	0.0545	0.0516

6. Conclusion

Conventional MPPTs have proven to be less efficient because of the functioning principle of photovoltaic system which depends on weather conditions. The principle of fuzzy logic based on the degree of membership has made it possible to implement a more efficient and more robust MPPT controller than conventional ones.

In order to make the MPPT controller faster and at the same time enrich the literature in this domain, we have proposed a new fuzzy PV MPPT controller based on custom defuzzification function. The results obtained show the action of defuzzification is important in the system response.

The proposed fuzzy PV MPPT controller based on custom defuzzification has confirmed this action and has given better results than both MPPT controllers P&O and Fuzzy MPPT for PV system according to of the performance indices of the responses.

References

[1] J. M. Riquelme-Dominguez, S. Martinez, "Systematic Evaluation of Photovoltaic MPPT Algorithms Using State-Space Models Under Different Dynamic Test Procedures," IEEE POWER & ENERGY SOCIETY SECTION, **10**, 45772–45783, 2022, doi: 10.1109/ACCESS.2022.3170714.

[2] X. Li, Q. Wang, H. Wen, and W. Xiao, "Comprehensive studies on operational principles for maximum power point tracking in photovoltaic systems," IEEE Access, **7**, 121407–121420, 2019, doi: 10.1109/ACCESS.2019.2937100.

[3] R. Dutta, R. P. Gupta , "Performance analysis of MPPT based PV system: A case study," 2nd International Conference on Emerging Frontiers in Electrical and Electronic Technologies, Patna, India, doi: 10.1109/ICEFEET51821.2022.9847729.

[4] M. Etezadinejad, B. Asaei, S. Farhangi, A. Anvari-Moghaddam, "An Improved and Fast MPPT Algorithm for PV Systems Under Partially Shaded Conditions," IEEE Transactions on Sustainable Energy, **13**(2), 732–742, 2022, doi:10.1109/TSTE.2021.3130827.

[5] X. Li, H. Wen, Y. Hu, L. Jiang, "A novel beta parameter based fuzzy-logic controller for photovoltaic MPPT application',Renewable Energy, **130**, 416-427, 2019, doi:10.1016/j.renene.2018.06.071.

[6] J. S. Ko, J. H. Huh, J. C. Kim, "Overview of maximum power point tracking methods for PV system in micro grid," Electronics, **9**(5), 816, 1–22, 2020, doi.org/10.3390/electronics9050816.

[7] R.B. Bollipo, S. Mikkili, P. K. Bonthagorla, "Hybrid, optimal, intelligent and classical PV MPPT techniques: a review," CSEE Journal of Power and Energy Systems, **7**(1), 9–33, 2021, doi: 10.17775/CSEEJPES.2019.02720.

[8] Tao Hai, Jincheng Zhoua, Kengo Muranak, "An efficient fuzzy-logic based MPPT controller for grid-connected PV systems by farmland fertility optimization algorithm," Optik, **267**, 2022, doi.org/10.1016/j.ijleo.2022.169636

[9] P. Boonraksa, T. Chaisa-Ard, S. Sommat, P. Pimpru, T. Booraksa, B. Marungsri, "Design and Simulation of Fuzzy logic controller based MPPT of PV module using Matlab Simulink," International Electrical Engineering Congress, 2022, doi: 10.1109/iEECON53204.2022.9741641.

[10] F. Mehazzem, M. André, R. Calif, "Efficient Output Photovoltaic Power Prediction Based on MPPT Fuzzy Logic Technique and Solar Spatio-Temporal Forecasting Approach in a Tropical Insular Region," Energies, **15**, 1–21, 2022, doi.org/10.3390/en15228671.

[11] C. R. Algarín, J. T. Giraldo, O. R. Álvarez, "Fuzzy Logic Based MPPT Controller for a PV System," Energies, **10**(12), 2–18, 2017, doi: 10.3390/en10122036.

[12] G.F.T. Kebir, C. Larbes, A. Ilinca, T. Obeidi, S T. Kebir, "Study of the Intelligent Behavior of a Maximum Photovoltaic Energy Tracking Fuzzy Controller," Energies, **11**, 2–20, 2018, doi: 10.3390/en11123263.

[13] Tehzeeb-ul Hassan, R. Abbassi, H. Jerbi, K. Mehmood, M.F. Tahir, K. M. Cheema, R. M. Elavarasan, F. Ali, I. A. Khan, "A Novel Algorithm for MPPT of an Isolated PV System Using Push Pull Converter with Fuzzy Logic Controller," Energies, **13**, 2–21, 2020, doi: 10.3390/en13154007.

[14] M. Seyedmahmoudian, R. Rahmani, S. Mekhilef, A.M.T. Oo, A. Stojcevski, T. K. Soon, A. S. Ghandhari, "Simulation and Hardware Implementation of New Maximum Power Point Tracking Technique for Partially Shaded PV System Using Hybrid DEPSO Method," IEEE Trans. Sustain. Energy, **6**(3), 850–862, 2015, doi:10.1109/TSTE.2015.2413359.

Advances in Science, Technology and Engineering Systems Journal Vol. 8, No. 4, 41-59 (2023)

www.astesj.com

Special Issue on Computing, Engineering and Multidisciplinary Sciences

ASTESJ
ISSN: 2415-6698

A Circuit Designer's Perspective to MOSFET Behaviour: Common Questions and Practical Insights

Ralf Sommer[*,1,2], Carsten Thomas Gatermann[3], Felix Vierling[1]

[1]*Technische Universität Ilmenau, Electrical Engineering and Information Technology, Electronic Circuits and Systems Group, Ilmenau, 98693, Germany*

[3]*IMMS GmbH, Ilmenau, 98693, Germany*

[2]*Technische Universität Ilmenau, Electrical Engineering and Information Technology, Power Systems Group, Ilmenau, 98693, Germany*

ARTICLE INFO

Article history:
Received: 28 February, 2023
Accepted: 14 May, 2023
Online: 25 July, 2023

Keywords:
MOSFET
Equations
Teaching
Symbolic Circuit Analysis

ABSTRACT

Metal Oxide Semiconductor Field-Effect Transistors are commonly taught in courses for electrical engineers as they are the most common components within integrated circuits. However, despite numerous papers and books on MOSFETs, students still struggle with understanding their behaviour, particularly in the saturation region. This paper presents an expanded explanation of MOSFET behaviour, with a consistent and causal derivation of Level 1 MOSFET behaviour from a few equations, aimed at students without an extensive technological background. The paper provides illustrative explanations to help them understand MOSFET behaviour and addresses common students' questions, such as why the current is limited by charge carriers in the semiconductor substrate and why characteristic curves do not follow a parabolic curve in saturation. In addition to providing a comprehensive introduction to MOSFET behaviour from a circuit designer's perspective, this paper also offers valuable insights into interpreting AC parameters in modern MOSFET models. These parameters are often key to understanding and solving circuit problems related to small signal behaviour and frequency response, as demonstrated through various industrial application examples. These examples highlight how to bridge modern MOS models, such as the BSIM model, with MOS-Level 2 modelling, which is easily interpreted by users. By presenting these real-world examples, analysed by a symbolic analysis tool incorporating the BSIM to Level 2 AC model, this paper provides a practical and accessible approach to teaching MOSFETs and their applications in industry.

1. Introduction

This paper is an extension of the contribution presented at the International Conference on Synthesis, Modelling, Analysis and Simulation Methods, and Applications to Circuit Design (SMACD) conference [1]. The extension includes a comprehensive introduction and technological considerations before deriving the current equations consistently, making them easily understandable for undergraduate students and addressing frequently asked questions. The motivation for this paper were the discussions with students in a Metal Oxide Semiconductor Field-Effect Transistors (MOSFET) fundamentals course in 2019 – it became clear that they often struggle with understanding the behaviour of MOSFETs in the saturation region. Specifically, they frequently ask for a simple explanation of why the current is limited by charge carriers in the semiconductor substrate and why the characteristic curves do not follow a parabolic course in saturation. To aid in their understanding, many analogies have been used, such as "students as electrons" rushing in and out of a lecture hall through doors or buses transporting people from one location to another. However, these analogies have failed to adequately explain the behaviour, and often, lecturers resort to citing additional effects that are not described or are parasitic in nature. In fact, a simple explanation of the Gate-channel capacitance fixing the number of charge carriers by $Q = CV$ and the inability of charge carriers to switch polarity for not further following the parabola can logically answer both questions.

[*]Corresponding Author: Ralf Sommer, ralf.sommer@imms.de

Surprisingly, students told that even lectures held by technologists failed to provide satisfactory answers to these questions.

The industrial application examples have also been expanded, demonstrating how to bridge the gap between modern MOS models such as the Berkeley Short-channel Insulated Gate Field-Effect Transistor Model (BSIM) and MOS-Level 2 modelling that is easily interpretable for the user. This approach in combination with symbolic circuit analysis based on computer algebra [2] can solve even industrial-related circuit problems related to small-signal and frequency response behaviour.

Field-Effect Transistors (FETs) are one of today's fundamental electronic devices whose basic idea is to change the conductivity of a system by the influence of an electric field. The fundamental operating principle of a surface field-effect transistor was first proposed by Lilienfeld ([3], [4] and [5]) and Heil [6] in the early 1930s. Subsequently, Shockley and Pearson [7] studied this idea in the late 1940s. After the first device-grade Si-SiO_2 system was realized by Ligenza and Spitzer in 1960 by thermal oxidation [8], Atalla proposed the basic MOSFET structure based on the Si-SiO_2 system [9]. As a result, the first MOSFET was reported by Kahng and Atalla in 1960 [10]. A complete breakdown of the historical development of the MOSFET can be found in [11] and [12]. For the technology, application and device physics, reference can be made to [13] - [16].

The MOSFETs are themselves subdivided into a wide range of sub-categories. In the following, only the most widespread semiconductor technology for integrated circuits, the MOS technology (based on silicon), will be discussed. Hierarchically, the MOSFET is found as a subgroup of the MISFET, which in turn can be classified under the IG-FET. The abbreviations stand for:

- MOS ... Metal / Oxide / semiconductor,
- MIS ... Metal / insulator / semiconductor and
- IG ... insulated Gate.

In general, FETs are used with a MIS-structure. Replacing the insulator by silicon dioxide (SiO_2) one obtains the MOS-structure with SiO_2 being the Oxide. This type of insulation is preferred because oxides are characterized by the fact that they have a high dielectric strength, which should ideally be infinite in an insulator and also because they increase the Gate capacitance (high relative permittivity κ or ε_r), the importance of which will be discussed later in this paper, without increasing the leakage currents. The insulated Gate significantly reduces the power consumption of the FET (only leakage current flows into the Gate). In addition to oxides, the Gate insulation can also be technologically realized using other materials such as silicon nitride, polymers or a combination of different materials, as in the Metal Nitride Oxide semiconductor FET (MNOSFET). Therefore, MOSFETs are to be considered as a subgroup of MISFETs.

MOSFETs can be used in a wide range of applications due to their favourable characteristics. These include applications such as: analogue switching, high impedance amplifiers, microwave amplifiers and digital circuits (complementary MOS, abbrv. CMOS). Among the most appealing features of a MOSFET are significantly higher input impedances (compared to a bipolar junction Transistor, abbrv. BJT), negative temperature coefficient at high current levels, and the absence of forward-biased pn-junctions. The high input impedances are a result of the Gate insulation, which make them particularly suitable for standard microwave systems. Because of the negative temperature coefficient, the current drops as the temperature rises, resulting in a uniform temperature distribution across the device and preventing thermal runaway and breakdown. A MOSFET is thus thermally stable even in a parallel connection. Due to the lack of forward-biased pn-junctions, minority charge carrier storages cannot form. Between two switching states, unlike BJTs, MOSFETs do not have to compensate for stored charges resulting from diffusion tails, the cause of which is the average lifetime of minority carriers, allowing them to achieve much higher large-signal switching speeds. In addition, MOSFETs can be operated as a constant current source due to their nearly constant current-voltage characteristics in the saturation region. This property is particularly exploited in integrated circuit (IC) technology in the form of diode-connected MOSFETs, e. g. for operating point adjustment or current mirrors.

In the following, this work focuses only on the conventional Bulk (long-channel) MOSFET, i. e., no Silicon on insulator (SOI) MOSFETs or MOSFETs with multiple Gate electrodes are considered.

2. Ease of Use

First, the general structure of a MOSFET must be clarified. For this purpose, Figure 1 shows the typical structure of an n-channel MOSFET, which has three or four terminals: Gate (G), Source (S), Drain (D) and Bulk (B). The controllable current flow occurs between the Source and Drain by manipulating the channel by varying the Gate-Source voltage, which allows the current through the transistor to be selectively influenced according to the current-voltage characteristic. The individual terminals are contacted with regions or semiconductive layers that are diffused into the substrate. Starting from the bottom terminal, the MOSFET Bulk consists of a p-type (PMOSFET: n-type) silicon single crystal, and Source and Drain are formed by two n^+ regions (PMOSFET: p^+). The latter represent reservoirs for the charge carriers that would be minority charge carriers in the substrate. For this reason, like a BJT, a MOSFET can also be modelled by two antiserially connected diodes.

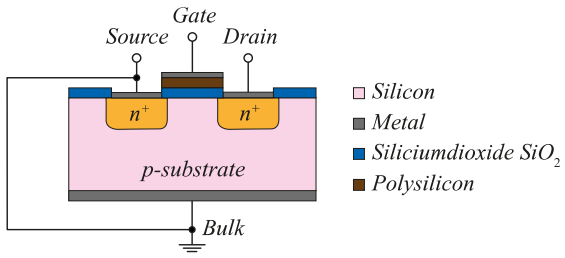


Figure 1: Conventional Bulk NMOSFET – General Structure

In contrast to the BJT, the Gate (analogue BJT: base) of a MOSFET is deliberately so wide that the space charge zones of the Source-substrate- and substrate-Drain- pn-junctions do not overlap. Thus, there is no transistor effect as with a BJT. For a voltage $V_{GS} < V_{th}$, one of the two pn-junctions (Source-substrate or substrate-Drain) on the Source-Drain path is always reverse biased, i. e., one of the two diodes is reverse biased regardless of the polarity of V_{DS}.

The "+" sign of the n^+ respectively p^+-areas means that these are heavily doped with impurity dopants. In most cases, the Gate in NMOSFETs is n^+-type polysilicon. As the name implies, it is not a Si-single-crystal but rather an imperfect crystal with a large number of grain boundaries. The Gate insulator is typically made of a thin SiO_2 layer. In stand-alone MOSFETs the Bulk is internally connected to Source to generate a common reference potential. Current can flow between Source and Drain, when a voltage with the correct polarity (i. e., + at the Gate pin) is applied between Gate and Source.

3. MOS-Structure

The heart of a MOSFET is formed by the MOS structure, which is obtained by omitting the Source and Drain regions of the conventional Bulk MOSFET. The substrate is often, as within this paper, given by silicon. Silicon is in the 4th main group in the periodic table of the elements proposed by Dmitri Mendeleev. According to [17], a chemical bond is particularly stable when the nearest electron gas configuration is reached (noble gas rule). With an electron configuration of: [Ne] $3s^2$ $3p^2$, silicon is still missing four electrons in the 3p-orbital to the next noble gas: Argon [Ne] $3s^2$ $3p^6$ (octet rule). Consequently, a Si atom in an (infinitely extended, i.e. no edges) silicon crystal forms four covalent bonds to the neighbouring Si atoms. Thus, next noble gas configuration is achieved, and the Si crystal is chemically stable. As mentioned before, the p-Bulk MOS structure, as used in NMOSFETs, consists of a Gate electrode (typically n^+ poly-Si), the Gate-Oxide as insulator (or dielectric as explained later) and a p-type substrate (Si Bulk) with a doping concentration N_A. The subscript "A" stands for "acceptor", since p-type doping means the intentional implantation of atoms with less valence electrons than Si into the Si single crystal of the substrate (e. g. boron, B, 3rd main group in the periodic table). Accordingly, the full four bonds to the neighbouring Si atoms in the crystal cannot be formed. As a consequence, the dopant atom is ionized (atomic body is negatively charged), and it provides a free-moving (diffusion processes in the semiconductor crystal) hole to the semiconductor structure. An explanation of the concept "hole" is given in Section 5.

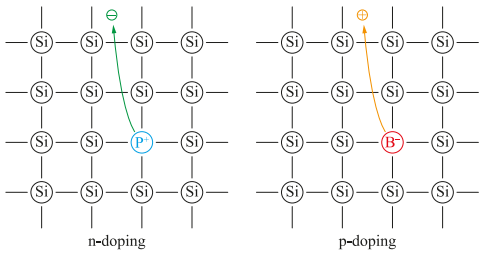


Figure 2: Silicon Crystal Structure with n and p Doping

In the ionized state the dopant atom is called an "acceptor ion" because the underlying atomic body has an affinity for electrons, i. e., the atomic body strives to compensate for the hole and achieve charge neutrality towards the outside. In contrast, an n-Bulk MOS structure of PMOSFETs consists of the Gate (typically p^+ poly-Si), the Gate-Oxide, and the n-type Si substrate with a doping concentration N_D, where the "D" stands for "donor", since atoms with more electrons than Si (like phosphorus, P, 5th main group in the periodic table) are incorporated into the Si single-crystal upon n-doping. With five valence electrons, phosphorus can form five covalent single bonds, unlike silicon. However, since only four neighbours are available for a potential bond in the semiconductor crystal lattice, the pentavalent phosphorus is also ionized (atomic body is positively charged). As a result, the crystal lattice has one free-moving electron (diffusion processes through the crystal lattice) available for charge transport. This is also the reason why these ions are called "donors". Figure 2 is intended to illustrate these relationships.

Because the Bulk contact is assumed to be grounded, the voltage drop across the MOS structure V_{GB} is equal to the applied potential at the Gate V_G, as shown in Figure 3. In the following, electrical potentials shall be labelled with only one letter in the subscript, e.g., V_G for the Gate potential, and voltages (potential differences) with two letters, for example V_{GS} for the Gate-Source voltage. Important for the voltages is the arrangement of the letters. The first letter indicates the start potential, and the second letter indicates the end potential, which makes the voltage arrow V_{XY} go from potential V_X to potential V_Y (according to the way of speaking). Therefore, the following relationship between voltage and potentials applies: $V_{XY} = V_X - V_Y$.

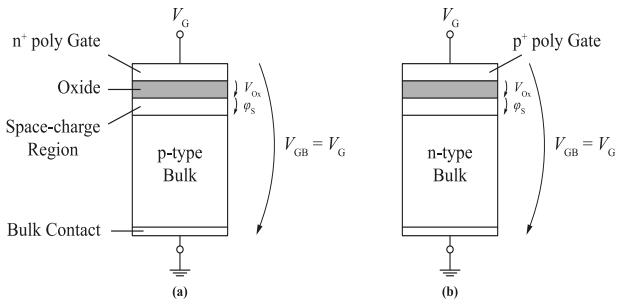


Figure 3: Two-Terminal MOS Structures. (a) p-Bulk MOS. (b) n-Bulk MOS. (adapted from [18]).

To derive the equations and gain a deeper understanding of the operating principle of the MOSFET, the type of transistor used must be specified. In this work, an n-channel transistor is taken as a model, since the NMOS is the standard model in lectures. The derivations and analysis can be performed analogously for a p-type, although some properties must be considered vice versa. This is due to the analogue but inverse MOS structure of the PMOSFET. This means for a PMOSFET, all voltages enter the equations with reversed polarity. In addition, the electrons contribute significantly to the charge transport (current flow) in an NMOSFET and the holes in a PMOSFET.

From a technological point of view, further insight into the physics of the MOS structure can be obtained by means of the energy band diagram. Figure 4 shows the energy band diagrams of the separated components of a NMOS structure under equilibrium conditions, i. e., Gate potential $V_G = 0$. The vacuum energy E_{vac} was chosen as the reference energy level. Each of the three regions has its own electron affinity χ, which is defined by the difference between E_{vac} and the conduction band lower edge E_C, and a work function ϕ, which results from the separation between E_{vac} and the Fermi energy E_F. Due to of the high n-doping of the Gate, the Fermi level is slightly above the conduction band edge E_C. The electron affinity is the same for both the Gate and the Bulk since both are Si, χ_{Si}.

The work function in the Gate ϕ_G corresponds approximately to the electron affinity χ_{Si}, which can be explained by the high doping. In the Bulk, the work function ϕ_B depends strongly on the Bulk doping N_A. The insulator (SiO_2) is characterized by the high band gap E_{Gox} (gap energy E_G = difference between conduction band lower edge E_C and valence band upper edge E_V).

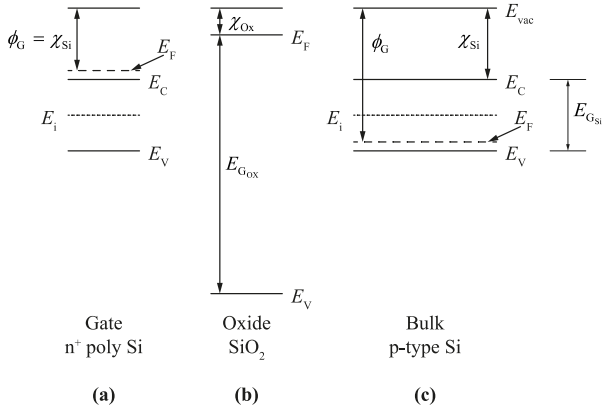


Figure 4: Energy Band Diagrams of the individual (separated) Components of a p-Bulk MOS Structure. (a) Gate (n^+-poly Si). (b) Insulator respectively Oxide (SiO_2). (c) Substrate (p-Bulk). (adapted from [18]).

When the three individual components are brought together (galvanic contact) to form a MOS structure and the equilibrium state is considered with no externally applied voltage ($V_G = 0$), the Fermi level of all three regions must be aligned in the same vertical position (same energy level) throughout the entire structure. To achieve this condition, the band diagram of the Bulk material can be held fixed at the contact points to the Oxide, while the rest of the structure can be pulled down until the Fermi energies E_F in the Bulk material and poly-Si match. This process results in band bending of the conduction and valence bands and of the intrinsic Fermi level E_i. To obtain a correct energy band diagram, the following points should be noted:

- In the substrate, the bands are bent only near the surface, i. e., the Si / SiO_2 interface, while far away from it the band relations remain unchanged. This is also the reason why the contact points to the Oxide should be held fixed during the band displacement.
- The relations between valence and conduction bands remain unchanged compared to the case where all three regions were separated.
- The intensity of the bending of the valence and conduction bands as well as the intrinsic Fermi level E_i are identical.
- Since the doping level of the poly-Si is so high, the bands at the Gate-Oxide interface bend only to a very small extent, so they can be neglected in the following.

The final band diagram of a p-Bulk MOS structure is shown in Figure 5 (a). The reason for the band bending is the work function difference between Gate and substrate (Bulk). To eliminate band bending, i. e., to make the bands flat, a certain voltage must be applied to the Gate. This specific voltage is called flat-band voltage V_{FB}. Note that for a p-Bulk MOS structure, V_{FB} is negative and for n-Bulk structures it is positive.

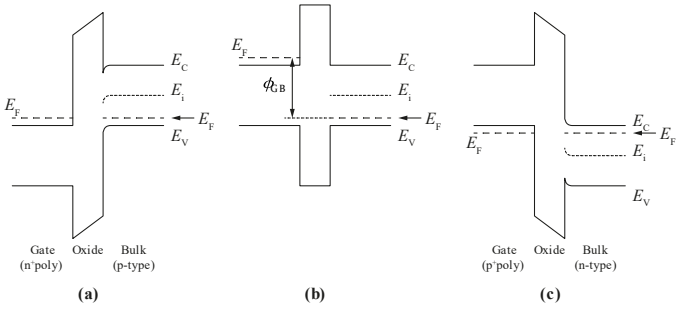


Figure 5: Band Diagrams of different MOS Structures. (a) p-Bulk MOS at $V_G = 0$.(b) p-Bulk MOS at $V_G = V_{FB}$ (c) n-Bulk MOS at $V_G = 0$ (adapted from [18]).

Another important parameter in MOS theory is the surface potential φ_S, which is related to band bending.

$$\varphi_S = \varphi_i(y=0) - \varphi_i(y \to \infty) = -\frac{E_i(y=0) - E_i(y \to \infty)}{e} \quad (1)$$

Here φ_i is the electrostatic potential introduced in Appendix B (cf. (31)) and y is the spatial coordinate in the direction of the perpendicular of the semiconductor surface (i. e. into the depth of the substrate). Therefore, the origin of the y-axis lies in the Oxide / substrate interface and is often referred to as the surface.

4. States of the (N)MOS Structure

To gain a deeper understanding of the operation of an NMOS structure (p-Bulk), the effects of different applied Gate voltages are first investigated. Throughout this analysis, the term "surface" (subscript: "S") is used to reference the interface between the Oxide (insulator) and the substrate (strong band bending). In contrast, the term "Bulk" (subscript: "B") is used to indicate that the analysis is performed far away from the interface, i. e., deep inside the substrate (little to no band bending). Additionally, all the different cases will be considered with a Source-Drain voltage equal to zero, $V_{DS} = 0$. In total, four relevant states can be identified:

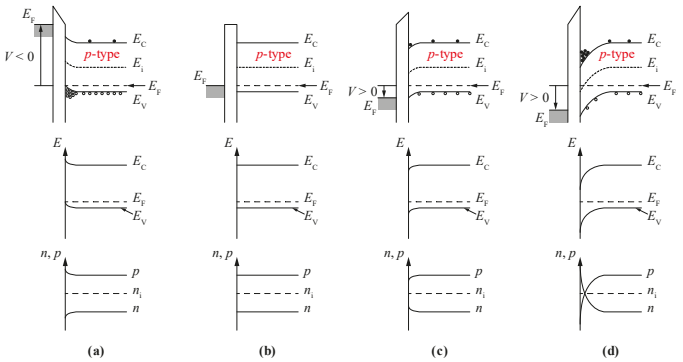


Figure 6: Band Diagram (top), Carrier Concentration (middle, Ordinate has log-scale) of Si Substrate for a p-Bulk MOSFET. (a) Accumulation, $V_G < 0$. (b) Flat-band Case, $V_G = V_{FB}$. (c) Depletion, $V_G > V_{FB}$. (d) Onset of strong Inversion, $V_G > V_{th}$. (modified taken from [18] and [19]).

Note, that from now on only the interior of the substrate is considered, with $x = 0$ being the interface between Oxide / substrate resp. Oxide / Bulk.

(1) $V_G < 0$ resp. $\varphi_S < 0$ (Accumulation case)

A sufficiently large negative voltage at the Gate (in relation to the Source) causes the energy bands to bend upwards near the substrate surface. This leads to a negative surface potential, $\varphi_S < 0$. Now the surface majority carrier (in case of a p-substrate (NMOS): holes) concentration p_S is larger than the Bulk majority carrier concentration p_B, and the surface minority carrier concentration n_S is smaller than its corresponding Bulk value n_B.

Within circuit technology and digital computing technology (CMOS circuit technology) this condition has no further meaning: In analogue circuits, MOSFETs are used primarily as amplifiers, and their behaviour is largely determined by the relationship between the input and output voltages and currents. Hence, most circuits operate in the saturation region, only a few in the linear region, but both require inversion. In digital circuits, MOSFETs are used primarily as switches, and the most important characteristics are their on / off states and the speed with which they can switch between these states. So, the accumulation region has no direct impact on these characteristics, and hence, it is also not relevant to digital circuit design.

(2) $V_{GS} = V_{FB}$ resp. $\varphi_S = 0$ (Flat-band case)

As the name suggests, the energy bands become flat at a less negative voltage, the flat-band voltage V_{FB}, i. e., the band bendings disappear.

(3) $V_{GS} > V_{FB}$ resp. $\varphi_S > 0$ (Depletion case)

When the Gate voltage is more positive than V_{FB}, the energy bands near the surface bend downward and the surface potential becomes positive, $\varphi_S > 0$. Thus, the holes are repelled from the surface (diffusion processes), which leads to the fact that only the uncompensated acceptor atomic hulls (ions) remain (firmly bound to the lattice structure of the substrate) and a space charge region is formed.

(4) $V_{GS} \geq V_{th}$ (Inversion)

Further increase of the Gate voltage enhances the depletion, i. e. n_S increases, p_S decreases and the thickness of the space charge region increases further. Once n_S equals p_S, the type of conductivity near the surface goes inverted. Consequently, in an NMOS with a p-substrate, the depletion region changes from p-conducting to n-conducting. This is called the onset of weak inversion. With the Gate voltage so large that the surface electron concentration is as high as the Bulk hole concentration, $n_S = p_B$, the onset of strong inversion is initiated. Now a channel (for NMOS: n-channel) is formed, which allows an effective current conduction. In other words, with the onset of inversion, the resistance of the channel is reduced. In addition, due to the local charge reversal, the effect of the antiserially connected pn-junctions (diodes; Source-substrate and substrate-Drain) is suppressed.

Note that the threshold voltage V_{th} is one of the most important electrical parameters of MOSFETs. It is defined as the Gate voltage that triggers the transition of the transistor from the off-state to the on-state. Because there is no uniform definition of the transition between the off-state and the on-state, several definitions of the threshold voltage are in use. In the Appendix A one can see the different definitions [18]. The authors restrict themselves within this publication to definition “Constant current”.

As soon as the Gate voltage V_{GS} exceeds V_{th}, a conductive channel is established. If the complete MOSFET, as shown in Figure 1, is now considered, a Drain current I_D can flow through the channel region. Therefore, the state of inversion is also called on-state and all other cases with $V_{GS} < V_{th}$ are called off-state. These terms originate from CMOS circuitry (digital logic).

5. Transition Process in the MOS Structure from the OFF-State to the ON-State

Within the off state, according to the mass action law

$$n \cdot p = n_i^2 , \tag{2}$$

with n being the electron concentration, p the hole concentration and n_i the intrinsic charge carrier concentration, a very temperature- and doping-sensitive charge carrier equilibrium is established, which is shown externally by the electrical charge neutrality. If the voltage between Gate and Bulk or Source (internal connection) is now increased so that the Gate receives a positive voltage, the charge carriers in the p-doped silicon shift. Electrons are increasingly drawn from the substrate below the interface between substrate and insulator (Oxide), which in turn causes the positive “charges” (holes) to be displaced from this region via diffusion processes into deeper substrate layers. It is important here that the electrons from the substrate do not leave it. Nor can they do so since the Gate is insulated by the Oxide. The phenomenon of displacement can be attributed to the effect of the electric field that builds up between the substrate and the Gate metallization. The result is the formation of a Gate-channel capacitance, with the Oxide acting as a dielectric.

Looking at (20), which will be explained in more detail in Section 8, it can be seen that the Drain current I_D, i. e. the current to be controlled, also depends on the Gate capacitance C_{ox}. Accordingly, with an increase in C_{ox}, an increase in I_D can be achieved. To enhance C_{ox} further, the insulator layer thickness was reduced in semiconductor technology. However, this caused several new problems with increasing miniaturization. On the one hand, the maximum field strength before an electric breakdown was reduced and on the other hand, the leakage currents increased drastically due to the tunnel effect. To ensure that C_{ox} can still be increased, and the insulator layer does not become too thin, the trend is toward high-k materials. The artificial word “high-k” is composed of the adjective “high” and the letter “k”. The “k” refers to the Greek letter kappa κ, which is the symbol used in English-speaking countries for relative permittivity. High-k materials (e. g. hafnium dioxide) are mainly used because they have a higher dielectric constant (relative permittivity ε_r), allowing for a thicker dielectric resp. insulator (Oxide) layer without compromising the device’s performance. This helps reduce Gate leakage current, improving the device's power efficiency, and Gate tunnelling current, improving the device's switching speed. In contrast, low-k materials are used to reduce parasitic capacitances in interconnects but are not used for the dielectric layer of very small MOSFETs due to their low dielectric constant. This is the reason why the permittivity of the Gate insulator has been increasingly addressed in recent years.

In the following, the Gate capacitance will be modelled as a plate capacitor with negligible edge effects. This insight gives rise to two explanations for the electron accumulation at the surface:

1) Equivalent charge principle:

Based on this principle, there are always the same number of charges (charges are always quantized) with opposite sign on the two electrodes of a capacitor. Therefore, if positive charges accumulate on the Gate electrode due to the positive (NMOS) Gate potential (conception: "suction" of free electrons from the e. g., highly doped poly-Si), negative charges (electrons) must accumulate at the interface between Oxide and substrate due to the Gate channel capacitance. Otherwise, the electrical charge neutrality of a capacitor cannot be preserved. Accordingly, the number of charge carriers which can accumulate at one electrode is limited by the other electrode (in this case the applied electrode voltage potential) and a proportionality constant which is impressed by the capacitance.

2) Influence:

With the "suction" of the electrons from the Gate electrode, the charge balance (neutrality) is disturbed in such a way that a positive centre of charge is formed. Since equal charges always repel each other (keyword: Coulomb or electrostatic forces), the conceptual "positive" charges (holes) are displaced into deeper substrate layers, leaving the negative charges (electrons). In addition, electrons are drawn out of the substrate under the interface via the same forces.

In an NMOS, the electrons in the p-substrate are the minority charge carriers while the holes are the majority charge carriers. Since the accumulation of electrons near the substrate / Oxide interface now occurs due to the above reasons, the holes within the interface are filled up by neighbouring places of the lattice. The consequence of the enrichment with electrons is noticeable by the depletion layer forming at the substrate / Oxide interface.

If there is a further increase in the Gate-Source voltage, the threshold voltage V_{th} is exceeded, and inversion comes into picture. Under these conditions, so many electrons are locally accumulated that the electrons become majority charge carriers. Thus, more electrons accumulate locally than the doping of the p-substrate can compensate. Because of the locally limited accumulation, the inversion is also locally limited. The so-called charge reversal starts. Incorrectly, it is often also referred to as a doping inversion, but this is wrong in the context of doping, since – in the case of an NMOS – there are no donor ions in the substrate. On the contrary, only acceptor ions can still be found in the p-doped substrate. The charge reversal in an NMOS creates a continuous, low-resistance n-conducting channel, which disables the function of the pn-junction between the n-doped islands and the substrate. Consequently, a current can now flow between Drain and Source: I_D. The channel region is only as large as the inversion zone and isolated from the substrate (channel: n-surplus, substrate: p-surplus) due to further pn-junctions forming (i. e. diodes in blocking direction). It follows that the MOSFET is voltage controlled, which makes it fundamentally different from the BJT, which is current controlled. Now it also becomes apparent where the terms NMOS and PMOS come from. The N and P refer to the conductivity of the forming channel. This is characteristic for the respective type of MOSFET, because an NMOS with an N-channel would immediately compensate all holes or positive charges due to the accumulation of electrons within the channel region and would therefore be non-conductive for these "positive" charge carriers. However, the electrons allow other electrons to pass. The opposite is true for the PMOS. Accordingly, the PMOS also requires a negative voltage at the Gate to form a p-type channel. At this point, a further difference can be identified in comparison to the BJT, since only one charge type contributes to the transport of the current depending on the substrate type. Therefore, the MOSFETs belong to the unipolar transistors and not to the bipolar transistors.

At this point it should be additionally noted that this is only a model, i. e., holes are not actual charge carriers, such as electrons or positrons. Holes are crystal lattice defects, where an electron is missing at the respective lattice place (e. g., consequence of doping with trivalent boron). This defect is interpreted as a conceptual positive "charge". Hence, holes cannot move by themselves. Instead, their mobility results from diffusion effects. For example, an electron from a neighbouring lattice place can fill up the vacancy (hole), creating a new hole at the lattice point where the electron came from. The electron has, so to speak, "jumped" one place further in the lattice. One can imagine this process as "wandering of the holes".

6. Deriving the MOSFET Equations (NMOS)

After explaining the individual states of the MOS structure, analytical equations describing the NMOSFET properties in the on-state (i. e., strong inversion case) are derived below. As mentioned above, the Gate-Source voltage causes to form a capacity between the substrate underneath the Gate and the Gate electrode, whereby – due to the positive potential at the Gate – electrons are attracted towards the Gate. The attracted and accumulated charge carriers create an induced n-layer. This yields a complete n-region from Source to Drain, i. e., a channel which forms kind of a low resistance tube where electrons can go from the Source electrode to Drain electrode. Thereby, the resulting electron flow builds up the desired current I_D through the MOSFET. A further increase of the Gate-substrate voltage attracts more electrons underneath the Gate and a higher current can be achieved. But why is the current limited by the Gate-substrate voltage and the number of electrons attracted from the p-substrate to forming the channel?

Looking at the MOSFET from a circuit-theoretic perspective, the MOSFET can be modelled as a voltage controlled current source (VCCS) as the Gate-substrate voltage resp. the Gate-substrate capacity influences the current between Source and Drain.

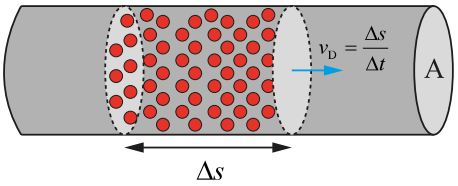


Figure 7: Electrons within a Conductor

The number of charge carriers N being stored in the tube shown in Figure 7 can be calculated using the equality

$$N = n \cdot A \cdot \Delta s = n \cdot V_{\text{Tube}}, \tag{3}$$

where n is the charge carrier density, A the cross-sectional area of the "tube" and Δs describes the way passed by the electrons in each Δt. Multiplying the cross-sectional area A with the way passed Δs gives the volume V_{Tube} in the second part of (3). The current through the MOSFET can be obtained by partially differentiating the charge by the time, i. e.

$$I = \frac{e \cdot n \cdot A \cdot \Delta s}{\Delta t} = e \cdot n \cdot A \cdot v_D \,, \tag{4}$$

In (4) e describes the elementary charge of an electron with $e = 1.6022 \cdot 10^{-19}$ As and v_D the drift velocity given as

$$v_D = \mu \cdot E \,, \tag{5}$$

The drift velocity is composed of the mobility μ and the electric field E applied to the tube.

7. Deriving the Current in the NMOS

The derivation of the current in the channel of a MOSFET is based on the Gate-channel capacity. This capacitance is the linchpin of the explanation of the saturation or limitation of the current: The electrodes of the "capacitor" are formed by the Gate on the one hand and the electrons in the channel underneath the Gate-Insulator on the other hand. It is important to mention that the electron layer, also called inversion layer, is in contact with the Source and the Drain at the same time (cf. Figure 8)

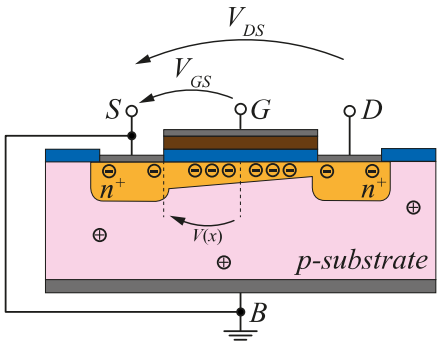


Figure 8: MOSFET with channel (NMOS)

The potential difference within the channel (and therefore the inversion layer) is determined by the potential difference between Source and Drain. The emerging capacity between Gate and channel can be modelled via the equation

$$Q = C \cdot V \,, \tag{6}$$

The real mathematical description of the charge within the Gate-channel capacity has a slightly shifted outcome as the threshold voltage V_{th} needs to be considered, since significant conductivity of the channel only occurs when this voltage is exceeded.

As previously indicated, the Gate-channel capacitance will be modelled as a plate capacitor with negligible boundary effects. The simple equation for the capacitance design applies. Adapting (6) to the current case gives

$$Q = C_{Ox} \cdot \left(V_{GC} - V_{th}\right) = \varepsilon \frac{A_{SiO_2}}{d_{SiO_2}} \left(V_{GC} - V_{th}\right) , \tag{7}$$

where V_{GC} represents the voltage across the Gate channel capacitance and $\varepsilon = \varepsilon_0 \cdot \varepsilon_{SiO2}$ the permittivity of the Oxide (insulator) layer. The area A describes the area covered by the Gate-insulator and can be approximated as

$$A_{SiO_2} = W \cdot L \,, \tag{8}$$

using the Gate width W and its length L, where length refers to the distance between the Source and Drain pn-junctions. We assume, as can be seen in Figure 8, that the channel length is equal to the Gate length. In reality, the effective channel length is assumed to be a little less than the full channel length. This is due, for example, to the fact that the ion implantation of the n^+ regions of the Source and Drain contacts cannot be focused precisely, resulting in a slight underdiffusion under the Gate (but still in the substrate), which leads to a reduction of the actual effective channel length. The thickness of the Oxide (distance between the Gate and substrate) is described by the parameter d_{SiO2} which is called TOX in Simulation Programs with Integrated Circuit Emphasis (SPICE).

As the capacitance is nearly constant, the number of charge carriers is enforced by the Gate-Source voltage (equivalent charge principle) and, hence, limited ($Q = C \cdot V$). Thus, it cannot be further increased by charge carriers from the battery.

To determine the current characteristic, the voltage $V(x)$ underneath the Oxide is introduced that varies between $V_S = 0$ und $V_D = V_{DS}$ as the Source potential is set as the reference potential. Therefore, the charge is also a function of the location x. If the voltage V_{GS} is not considered from the point of view of the Gate, but from the point of view of the channel, i. e., the counter-electrode, then due to the principle of equivalent charge, the voltage is reversed. Due to this, the voltage V_{GS} in the substrate points from Source to Drain and $V(x)$ is oppositely directed. In the next step, to determine the channel voltage at a certain location x, all voltages can be superimposed according to the superposition principle. Thus, the Gate channel voltage V_{GC} is composed of the difference between V_{GS} in the substrate and $V(x)$. Considering the infinitesimally small area

$$\mathrm{d}A_{SiO_2} = W \cdot \mathrm{d}x \,, \tag{9}$$

the charge can be evaluated to

$$\mathrm{d}Q = -\frac{\varepsilon_0 \varepsilon_{SiO_2} W}{d_{SiO_2}} \left[V_{GS} - V_{th} - V(x)\right] \mathrm{d}x \,, \tag{10}$$

The negative sign is due to electrons being the charge carriers in an n-channel MOSFET. As a preparation for the integration to get the Source-Drain current an essential case study must be executed: linear region and saturation region need to be looked at separately. The essential argumentation for these two regions is that the voltage term

$$V_{GS} - V_{th} - V(x) \,, \tag{11}$$

may never get negative or change its sign because otherwise the sign of the charge $\mathrm{d}Q$ would be reversed and the electrons being the charge carriers would suddenly become positrons.

It is also possible to look at the energy of a capacitor which cannot become negative, too. The current

$$I_{DS} = I_D = \frac{\mathrm{d}Q}{\mathrm{d}t} \,, \tag{12}$$

flows against the direction of travel of the electrons. Their speed is described by the drift velocity v_D. Henceforth, it shall be additionally assumed that the electron mobility in the channel is constant and does not depend on the electric field in the x-direction. Mobility reduction due to the vertical electric field will not be

considered here. This assumption works well for long-channel MOSFETs, but accuracy is reduced, especially when considering nanometre MOSFETs, since constant mobility and thus a linear velocity-(electrical)field characteristic is assumed throughout the entire device. This is remedied by the two-region MOSFET model, which will not be discussed here, but can be looked up in [18]. The minus sign is a consequence of the direction of the electric field, which points from Drain (positive potential) to Source (negative potential; grounded), where the electrons flow in the opposite direction. It is additionally noteworthy that the mobility μ_n is not the substrate mobility, but the effective mobility, which is explained in more detail in Appendix C.

$$v_D = -\frac{dx}{dt} = -\mu_n E = -\mu_n \frac{dV(x)}{dx}, \tag{13}$$

Since – as already mentioned in the introduction – this work is restricted to long channel MOSFETs, the longitudinal field along the channel is not sufficient to achieve a charge carrier saturation velocity. Under these circumstances, the velocity is coupled to and limited by the mobility (and thus within this model constant). Otherwise, the short-channel effects explained in more detail in [18] would have to be considered.

Using (12) and differentiating (10) regarding the time the differential equation for the current I_D,

$$I_D = -\frac{\varepsilon \cdot \mu_n W}{d_{SiO_2}}\left[V_{GS} - V_{th} - V(x)\right]\frac{dV(x)}{dx}, \tag{14}$$

can be obtained. Equation (14) can be solved by integrating from 0 to L along the channel and from V_{DS} to 0 because of the voltage directed in reverse to the flow of electrons. Changing the integration variable by multiplying (14) with dx the voltage $V(x)$ can be replaced by V as the integration incorporates solely the voltage difference and no longer the location along the channel.

$$\begin{aligned} \int_0^L I_D dx &= \int_{V_{DS}}^0 \left\{-\frac{\varepsilon \cdot \mu_n W}{d_{SiO_2}}\left[V_{GS} - V_{th} - V\right]\right\} dV \\ I_D \cdot L &= \frac{\varepsilon \cdot \mu_n W}{d_{SiO_2}}\left[\left(V_{GS} - V_{th}\right)V_{DS} - \frac{1}{2}V_{DS}^2\right] \end{aligned} \tag{15}$$

The factor

$$\mu_n \frac{\varepsilon}{d_{SiO_2}} = \mu_n \frac{\varepsilon_0 \varepsilon_{SiO_2}}{d_{SiO_2}}$$

is called K' or KP, whereby the fraction (without the mobility μ_n) is called C'_{ox} (effective oxide capacitance per unit area).

Looking at the Drain current I_D it is clear to see that the characteristic curve has a quadratic behaviour at the beginning (cf. Figure 10). The distribution of the charge carriers in the channel has an approximate triangular shape (at least shown in many textbooks, e. g. [20]), which forms the channel, as the electrons are pulled stronger towards the Drain the higher the Drain-Source voltage V_{DS} is. At the peak point of the characteristic curve, that is when

$$V_{DS} = V_{GS} - V_{th}, \tag{16}$$

the current does not sink but remains nearly constant.

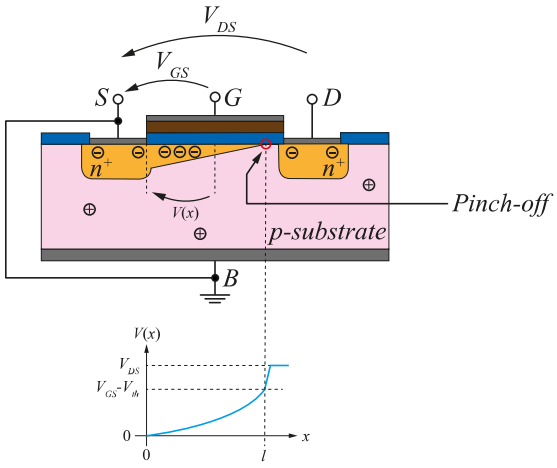


Figure 9: NMOSFET with Pinch-off

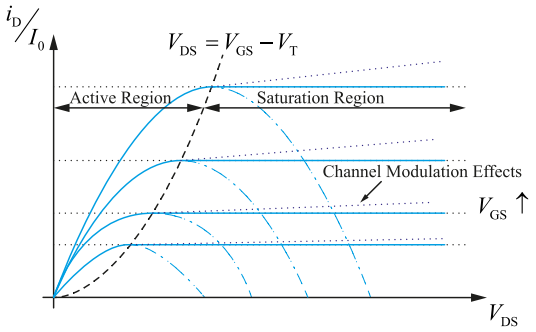


Figure 10: I_D as a Function of V_{GS} and V_{DS}

This peak is the point where the so-called Pinch-off takes place. No more charge carriers can reach for the Drain. If the Drain-Source voltage is further increased the length of the Pinch-off region varies in a way that only a certain amount of charge carriers can reach the Drain. The remaining electrons recombine as minority charge carriers with the holes in the p-substrate. Here one finds the missing arguments in most of the books and lectures: There is no inversion regarding the polarity of the charge carriers nor negative energy can be stored on a capacitor ((10), (11)). The resulting voltage for the current respectively the charge carriers remains constant. This is the explaining argument clarifying the trouble of the students in understanding the saturation region, and why it does not follow the parabola (15) after its peak (cf. Figure 10, blue dashed line).

$$V_{DS} = V_{GS} - V_{th} = V_{DS,Sat} \tag{17}$$

If $V_{DS} \leq V_{GS} - V_{th}$ (15) shows a linear correlation with V_{GS}. This section of the characteristic curve is therefore called linear, ohmic or active region. When $V_{DS} > V_{GS} - V_{th} = V_{DS,Sat}$ the voltage term in (15) needs to be set to $V_{DS} = V_{GS} - V_{th} = V_{DS,Sat}$. The equation for I_D in the saturation region can then be evaluated to

$$I_D = K' \frac{W}{2L}\left(V_{GS} - V_{th}\right)^2, \tag{18}$$

The current I_D stays nearly constant (regarding this model approach) and has a quadratic dependence on the Gate-Source voltage V_{GS}, which can be perfectly modelled by a VCCS. The quadratic dependence on V_{GS} can be seen in Figure 10 as the

distances between each blue curve within the saturation region are not uniform.

Within the device, the Pinch-off effect can be thought of as follows:

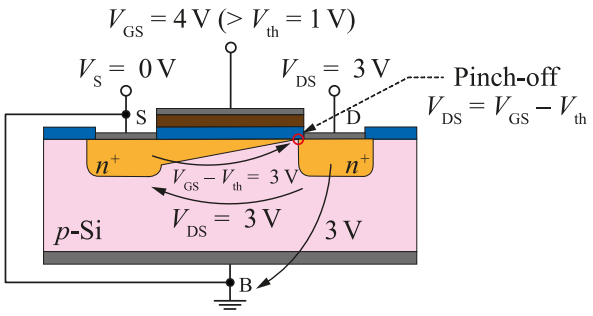


Figure 11: Calculation Example for the Superposition of the Voltage in the Substrate

The Pinch-off effect within the MOSFET device is a result of voltage drops within the substrate. This effect is illustrated by a small calculation example, where the voltage drops superpose within the substrate, thereby affecting the shape of the channel. This calculation example is only intended to illustrate the facts of the voltage drops and is by no means an accurate model. However, as can be seen very well, two voltages overlap within the substrate. Firstly, the one which is impressed by V_{DS} from Drain to Source and secondly, the one which drops due to the principle of equivalent charge between the interface substrate / Oxide and Source. The latter goes from Source to substrate / Oxide interface, since for an NMOS $V_{DS} > 0$ applies and thus the other plate of the plate capacitor (corresponding to the modelling of the MOS structure) must be negatively charged and thus the Source potential $V_S = 0$ V is larger. Consequently, the two voltages are oriented oppositely. They cancel each other out exactly when $V_{DS} = V_{GS} - V_{th}$ is fulfilled. This point is the so called "Pinch-off point", which occurs when there is a reduction in the relative voltage between the Gate and the substrate, as explained and shown in Figure 11. The point marks the x-coordinate at which the channel is pinched off and the strong inversion changes into a depletion zone. If the Drain voltage V_{DS} is now further increased (above $V_{DS,sat}$) at constant Gate voltage (V_{GS} = const.), the voltage drop from the Drain to the Source in the substrate predominates. As a result, the channel tends to be pinched off even sooner with respect to the x-coordinate. For $V_{DS} > V_{DS,sat}$, the Pinch-off point moves closer to the Source, but the voltage at the Pinch-off point remains the same at $V_{DS,sat}$. So, the saturation voltage is the voltage at which just such a channel and thus the strong inversion still exists in the substrate. This voltage cannot change with a fixed Gate voltage V_{GS}, since the charge carriers in the channel (minority charge carriers with respect to the substrate) are determined by the MOS capacitance C_{ox} – as already explained in the derivation. Thus, the total number of charge carriers (proportional to the overall charge) in the channel cannot change (otherwise the charge neutrality would be violated), whereby also according to $Q = C \cdot V$ with $C = C_{ox}$ = const. and Q = const. the voltage drop, that needs to be compensated for the onset of saturation (channel Pinch-off) by the Drain-Source voltage, cannot change.

In summary, the Pinch-off phenomenon in MOSFETs is influenced by the interplay between the Gate-Source and Drain-Source electric fields. The Gate-Source field creates the channel, while the Drain-Source field causes a voltage gradient along the channel that directly affects the electron concentration in the substrate. When V_{DS} is small and $V_G > V_{th}$ is fulfilled, the Drain-Source electric field is relatively weak, and the channel remains uniformly populated with electrons. As V_{DS} increases, the Drain-Source electric field becomes stronger, leading to a reduction in electron concentration near the Drain region (cf. Figure 12). When the Drain-Source field becomes strong enough to oppose the Gate-Source field at the Drain end, the channel is pinched off, and the transistor enters the saturation region. Beyond the Pinch-off voltage, the channel's resistance increases, and the Drain current I_D remains relatively constant despite further raise of V_{DS}.

It is important to note that as the Drain potential V_D increases, the voltage drop from the Drain to the Bulk also increases due to the internal connection of Source and Bulk. Consequently, the space charge region and the area of influence of the depletion induced by the Drain-pn-junction increase on the Drain side. This enlargement of the space charge zone results in more occupiable states for the same number of minority charge carriers, leading to a decrease in charge carrier density on the Drain side. Thus, the area of strong inversion (channel) decreases. Figure 12 shows the described situation, where the wedge-shaped region marks the area of strong inversion, where depletion comprises the area where more minority charge carriers are present than in the Bulk. Sufficient minority carriers are also present outside the strong inversion, i. e., in the depletion region, where most of the substrate majority carriers are compensated. Therefore, in the saturation region, the Drain current can still reach the Drain-side n^+-island. However, the electrical resistance within the depletion zone is higher than that within the local inversion (conductive channel). Thus, within the saturation, the resistance that the electrons "see" increases as they flow through the substrate from the Source to the Drain region (physical current direction). This shows on the one hand, why the current does not abruptly drop to zero and on the other hand, why the current remains almost constant within the saturation. Due to the constant voltage $V_{D,sat}$, the number of charge carriers arriving at the Pinch-off point remains the same. Consequently, despite the reduction of the effective channel length from L to L', approximately the same current I_D flows. Only if the shortened amount of the effective channel length is a substantial fraction of the channel length, an increase in the Drain current can be observed. As the channel length is reduced, the electric field in the channel increases, which causes the depletion region to expand towards the Source. This expansion reduces the effective length of the channel, and the Pinch-off voltage also reduces. As a result, the Drain current increases as the channel length decreases, assuming the voltage between the Source and Drain remains constant. This effect is considered within the channel length modulation in a more detailed model description, as discussed in the next section.

In summary to understand the difference between Pinch-off voltage and threshold voltage, the Pinch-off voltage is the voltage at which the depletion region around the Drain meets the depletion region around the Source, causing the channel to be "pinched off" and the Drain current to saturate. It is also sometimes referred to as the saturation voltage $V_{DS,sat}$. On the other hand, the threshold voltage is the voltage at which the MOSFET just starts to conduct current, with the channel beginning to form under the Gate. It is the minimum voltage required at the Gate to induce a channel and allow current flow between the Source and Drain. Hence, the threshold voltage V_{th} refers to the Gate-Source-voltage and Pinch-off voltage $V_{DS,sat}$ refers to the Drain-Source-voltage. They can be seen in the equations that describe the device's behaviour, such as the Drain current equation. In the saturation region, the Drain

current is essentially independent of the Drain voltage and is limited by the Pinch-off voltage. As mentioned before, this effect is particularly exploited in integrated semiconductor technology, since the MOSFET resembles a current source within saturation, i. e. almost constant current over a wide voltage range. Both voltages are important parameters in the MOSFET model and are influenced by various, such as the Gate-Oxide thickness and the doping concentration of the semiconductor material.

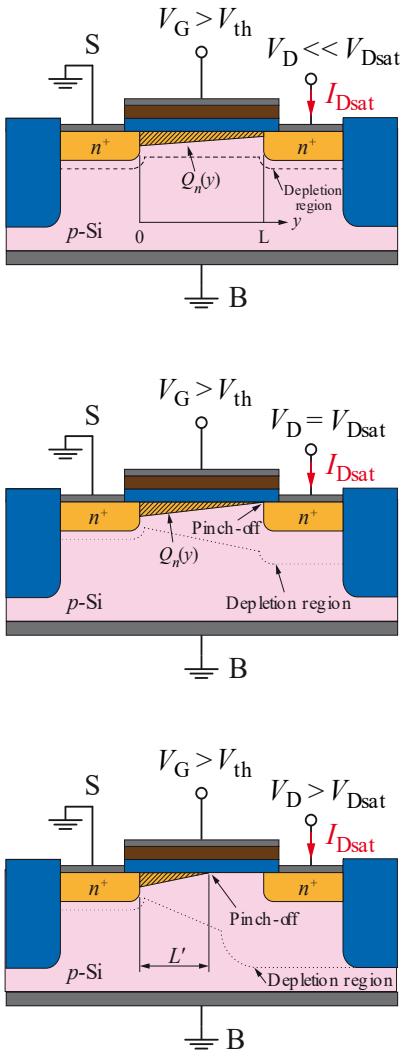


Figure 12: MOSFET operated (a) in the linear Region, (b) at Onset of Saturation, and (c) beyond Saturation (reduced effective Channel Length) with technical Current Direction. (taken from [19])

8. A more accurate Modelling of the Characteristic Curves

The ideas developed beforehand aimed to provide an easy derivation and intuitive understanding of the MOSFET characteristics. Another important parameter that has not been discussed so far is the threshold voltage V_{th}. It describes the Gate-Source voltage from where the inversion layer is built up and a significant current flow starts.

Using the equality (Shichman-Hodges resp. Level 1 model; no consideration of device random noise (thermal and flicker), sub-threshold behaviour, and high frequency effects)

$$V_{th} = V_{th0} + \gamma\left(\sqrt{2\phi_F - V_{BS}} - \sqrt{2\phi_F}\right), \tag{19}$$

With V_{th0} as zero threshold voltage, γ as the body effect parameter, and ϕ_F as the Fermi potential remaining constant, when the semiconductor is in balance. The zero-threshold voltage is the value of the threshold voltage V_{th} when $V_{BS} = 0$, i. e., the substrate has Source potential. The Fermi potential is the energy at which the Fermi-Dirac distribution has the value 1/2. Since this energy corresponds directly to a distribution function, it is (in very simplified terms) a measure of the number of free-moving charge carriers in semiconductor and is strongly influenced by the doping.

Equation (19) reveals an influence of the Bulk-Source voltage V_{BS} on the concrete value of the threshold voltage V_{th}. Consequently, the threshold voltage varies as a function of the Bulk-Source-voltage. This effect is commonly referred to as the "body effect". In the small-signal model (cf. Figure 19, Figure 20, Figure 24), this must be taken into account (Bulk-control). As mentioned before, usually the Bulk-potential is set to a fixed value (e. g. ground) or connected to Source, which reduces or compensates the body effect. It allows further control of the transistor and is especially important in integrated circuit technology. Less commonly, through hole technology (THT) MOSFETs also have the Bulk terminals pulled out of the package as a fourth pin.

Another important effect within the MOSFET and the Shichman-Hodges model is the channel-length modulation (CLM). Increasing the Drain-Source voltage over the value of $V_{DS,sat} = V_{GS} - V_{th}$ causes the Pinch-Off inducing a reduction of the effective channel length and therefore raises the current. Modelling the rising current when the Drain-Source voltage increases can be done with the channel length modulation factor λ. Within an empirical model, λ results from the common intersection point of the voltage axis and the extensions (tangents created in the saturation region) of all branches in the output characteristic field (I_D as a function of V_{GS}).

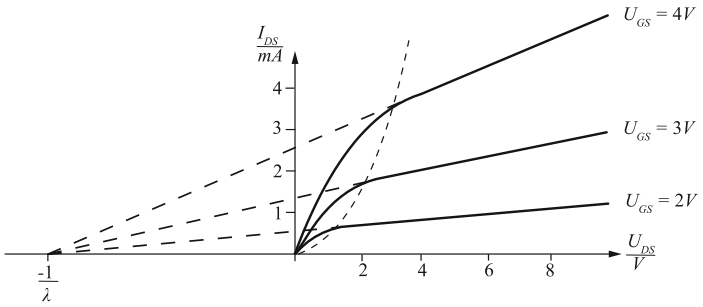


Figure 13: Channel Length Modulation, Asymptotes in Saturation Region meet at the Point of Early Voltage.

Using the intercept theorem, the equation for the characteristic curve of the MOSFET in saturation region can now be written as

$$I_D = \frac{\mu C'_{ox}}{2}\frac{W}{L}\left(V_{GS} - V_{th}\right)^2\left(1 + \lambda \cdot V_{DS}\right), \tag{20}$$

Considering the whole characteristic, the linear region also needs to be modified:

$$I_D = \frac{\mu C'_{ox} W}{L}\left[\left(V_{GS} - V_{th}\right)V_{DS} - \frac{V_{DS}^2}{2}\right]\left(1 + \lambda V_{DS}\right), \tag{21}$$

Typically, λ has a value of 0.05 V^{-1}. λ induces an r_{DS} in the alternating current (AC) model. The effect on the characteristics of the MOSFET resulting from the channel-length modulation can be compared with the Early effect in bipolar transistors. However, their origins are different, which is why they are not identical.

Up to now, only the case where $V_{GS} > V_{th}$ has been considered analytically for NMOSFETs. In this range, the presented model works very well, but it lacks accuracy for Gate-Source voltages smaller than V_{th}. When this condition is met, only a very small Drain current I_D (not zero) flows. This is called the subthreshold regime. For the Drain current in the subthreshold regime of an NMOSFET, a rather long derivation, which will not be explained in detail here but can be found in reference [18], results in the following expression:

$$I_D = \frac{\mu W}{L}\sqrt{\frac{\varepsilon_{Si} e N_A}{4V_T \ln\left(\frac{N_A}{n_i}\right)}}\cdot V_T^{\,2}\cdot e^{\frac{V_{GS}-V_{th}}{m\cdot V_T}}\left(1-e^{-\frac{V_{DS}}{V_T}}\right), \tag{22}$$

with V_T as thermal voltage: $V_T = k \cdot T / e$ and m as body factor:

$$m = 1+\sqrt{\frac{\varepsilon_{Si} e N_A}{4V_T \ln\left(\frac{N_A}{n_i}\right)}}\cdot\frac{1}{C_{ox}'}, \tag{23}$$

Because of the resulting impractical expression, a new widely used parameter has been introduced to characterize subthreshold MOSFET behaviour: sub-threshold slope S. This emerges from (22) assuming $V_{DS} >> V_T$, since under this condition the last exponential term converges to zero. S is then defined as the derivative of the logarithm of I_D with respect to V_{GS}.

$$S = \left(\frac{\mathrm{d}}{\mathrm{d}V_{GS}}\log I_D\right)^{-1}\begin{cases}=\ln 10\cdot m\cdot V_T\\ \approx 2.3\cdot m\cdot V_T\end{cases}, \tag{24}$$

The unit of S is typically given as mV/dec. The subthreshold slope gives an indication of how much the Drain current varies as a function of the Gate-Source voltage. For example, if S is given as 100 mV/dec and the Gate voltage is changed by 100 mV, the subthreshold Drain current will change by a factor of 10, i. e., a decade. The theoretical lower limit of S is 60 mV/dec, since an ideal case of $m = 1$ would then be encountered. In addition, it should be noted in the considerations that the equations only yield ideal values, since, for example, short-channel effects are neglected. These effects will not be discussed further in this paper. However, [18] can be cited as a suitable source for further reading.

What has also not been discussed in the previous MOSFET models are parasitic resistances. So far, it was always assumed that the voltages V_{GS} and V_{DS} applied to the external terminals are equal to those appearing at the intrinsic transistor. However, in reality there are series resistances between the intrinsic transistor and the external terminals, since the Drain current that flows through the channel must also pass through the Source and Drain series resistances R_S and R_D. These are caused, for example, by path resistances of the Drain and Source regions or contacting resistances. This leads to parasitic voltage drops, which is why the voltages across the intrinsic transistor are usually lower than the voltages applied to the terminals. The relationship between the external voltages (subscript: "ext") and the internal voltages (subscript: "int") are given by:

$$V_{DS,ext} = V_{DS,int} + I_D\cdot\left(R_S + R_D\right), \tag{25}$$

$$V_{GS,ext} = V_{GS,int} + I_D\cdot R_S, \tag{26}$$

The resistors have only a negligible effect on the MOSFET current-voltage characterization (I-V-characteristic) below the threshold voltage, i. e., in the subthreshold region. In contrast, in the on-state, i. e., beyond threshold, they lead to reduced Drain currents compared to the ideal case of the intrinsic transistor. This effect of reduced Drain current in the on-state can play a substantial role in nanometre MOSFET. For this reason, one of the design goals is to minimize R_S and R_D. In addition, this reduces the thermal losses of the transistors, which is why this can also be quite relevant for circuit design.

9. Models for Computer-aided and manual Analysis

Analog circuit designers are continuously facing the challenge to squeeze as much performance out of their circuits as the underlying semiconductor technology permits. Pushing a circuit topology to its technological limits requires extensive knowledge of previous designs and deep insight into both the intended functional behaviour of the device under construction and the parasitic effects that degrade its performance. Since conventional numerical circuit simulators cannot provide qualitative insight into the functional dependencies between circuit parameters and behavioural characteristics, it is often necessary to perform a manual analysis of the circuit. At the end of such an analysis the goal is to get an understanding of how the circuit works by interpretation of the mathematical (symbolic) expressions. In this context, the most important application for symbolic (manual) analysis is to gain design knowledge about undesired circuit behaviour observed in numerical simulations, e. g., in the form of resonance effects and instabilities due to parasitic poles or poor power-supply rejection behaviour.

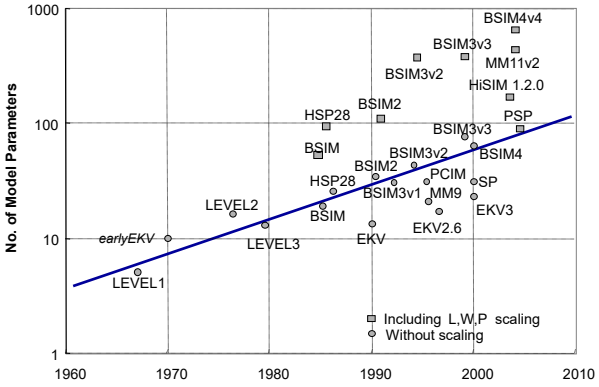


Figure 14: Evolution Model Parameters for MOS Transistors. (taken from [21])

For a fully comprehensive description of the internal operation of a MOSFET to be possible, three-dimensional complete quantum mechanical and atomistic simulations would be required. Since this currently seems unattainable due to a lack of computational resources, various device engineers and researchers have developed simplified abstract MOSFET models with different levels of complexity over the past decades. Because the share of CMOS chips is the largest, a lot of effort has been put into the modelling of MOSFETs, especially pushed by semiconductor manufacturers. Today, models (e. g., BSIM) with several hundred parameters are not uncommon (cf. Figure 14).

On the other hand, designers do not think in the BSIM, especially for the small-signal behaviour as shown in Figure 21. To gain insight into the operation of the circuits, many analogue designers use (15) or (18) for Direct Current (DC), since these equations describe the static behaviour of the MOSFET. Here, the fundamental transistor function of forming and controlling the channel current is modelled. As mentioned above, from a circuit point of view, this model can be represented by a VCCS. However, the dynamic behaviour is neglected and remains unconsidered. This can be remedied by inserting additional internal capacitances. The most important capacitances of a MOSFET and their summary (on the right) are shown in Figure 15.

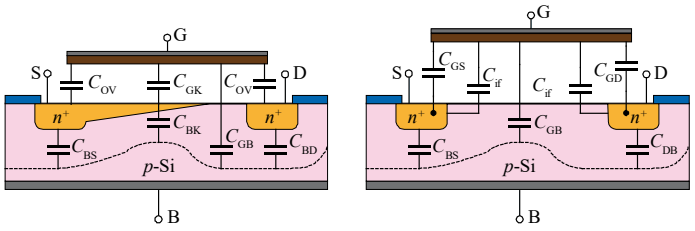


Figure 15: Sectional View of the MOSFET with the most important Capacitances, summarized on the right.

It should be noted that the effect of the individual capacitances is largely dependent on the operating range in which the transistor is located. For example, voltage-dependent capacitances can be defined depending on the Gate-source voltage V_{GS} (cf. Figure 16).

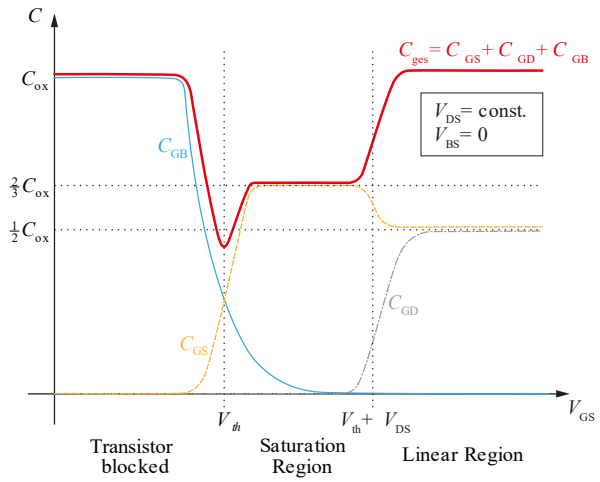


Figure 16: Compilation of the qualitative Capacitance Curves as a Function of the Gate-Source Voltage V_{GS}.

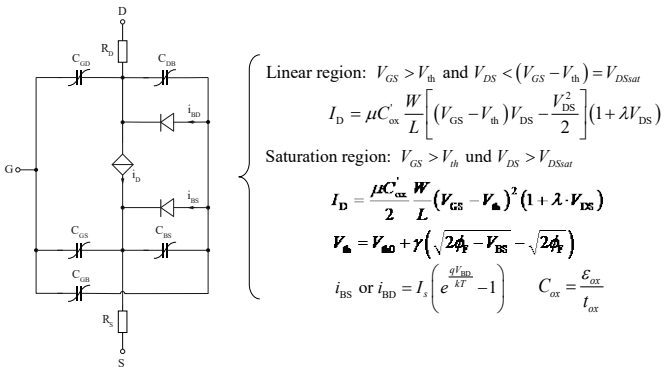


Figure 17: MOSFET Model with the static large-signal Equations.

With additional consideration of the pn-junctions from the Source or Drain to the Bulk, which are abstracted to junction diodes, as well as the terminal resistances, a complete (dynamic) large-signal equivalent circuit can be assembled (cf. Figure 17).

For example, after the operating point has been found using the large-signal equivalent circuit diagram, which roughly corresponds to the SPICE Level 1 model (cf. Figure 18), circuit engineers are then interested in the behaviour of the circuit when excited with signals of small amplitude and power. This is referred to as small-signal behaviour. The corresponding equivalent circuit diagrams can be derived from the large-signal model.

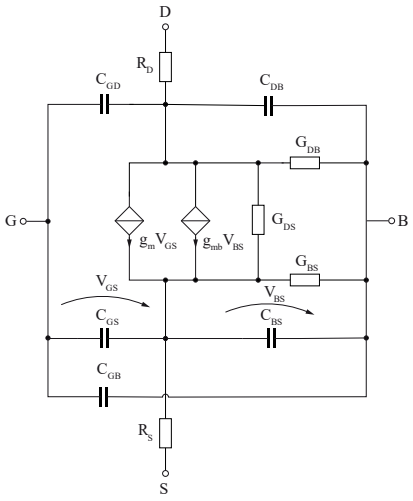


Figure 18: SPICE MOSFET Level 1 small-signal Model.

Using most manual analysis, the goal is to obtain good descriptions of the circuit functional behaviour with as little computational effort as possible. For this reason, the SPICE MOSFET Level 1 model is often already far too complex. A possible simplification for manual analysis is shown in Figure 19. In most cases, however, it will find application without g_{mb}, since Source is mostly connected with Bulk, which makes V_{BS} zero and therefore the corresponding controlled current source can be neglected. The transconductance g_{mb} describes the dependence of the Drain current on the Bulk-Source voltage, i. e., the previously explained body-effect.

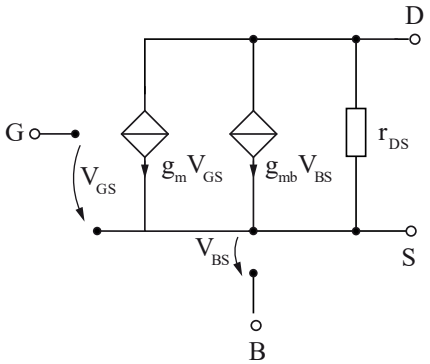


Figure 19: Simple static MOSFET small-signal Model for manual Analysis.

Herby the small-signal quantities are calculated as follows:

Linear / Triode range:

$$\text{NMOS}: V_{GS} > V_{th} \text{ and } |V_{DS}| \le |V_{DS,sat}|$$
$$\text{PMOS}: V_{GS} < V_{th} \text{ and } |V_{DS}| \le |V_{DS,sat}|$$
$$g_m = \frac{\partial I_D}{\partial V_{GS}} = \mu C'_{ox} \frac{W}{L} V_{DS}$$
$$g_{mb} = \frac{\partial I_D}{\partial V_{BS}} = \frac{\mu C'_{ox} \gamma}{2\sqrt{2|\Phi_F| - V_{BS}}} \frac{W}{L} V_{DS}$$
$$r_{DS} = \frac{1}{g_{DS}} = \frac{\partial V_{DS}}{\partial I_D} = \frac{1}{\mu C'_{ox} \frac{W}{L}(V_{GS} - V_{th} - V_{DS})}$$

Saturation range:

$$\text{NMOS}: V_{GS} > V_{th} \text{ and } |V_{DS}| \ge |V_{DS,sat}|$$
$$\text{PMOS}: V_{GS} < V_{th} \text{ and } |V_{DS}| \ge |V_{DS,sat}|$$
$$g_m = \frac{\partial I_D}{\partial V_{GS}} = \mu C'_{ox} \frac{W}{L}(V_{GS} - V_{th}) = \sqrt{2\mu C'_{ox} \frac{W}{L} I_D} = \frac{2I_D}{V_{GS} - V_{th}}$$
$$g_{mb} = \frac{\partial I_D}{\partial V_{BS}} = g_m \frac{\gamma}{2\sqrt{2|\Phi_F| - V_{BS}}} = \eta g_m$$
$$r_{DS} = \frac{1}{g_{DS}} = \frac{\partial V_{DS}}{\partial I_D} = \frac{1}{\frac{1}{2}\mu C'_{ox} \frac{W}{L}(V_{GS} - V_{th})^2 \lambda} \approx \frac{1}{\lambda I_D}$$

It is important to note again that most analogue circuit applications are operated in the saturation region because that is where the transistor acts like an almost ideal controlled current source.

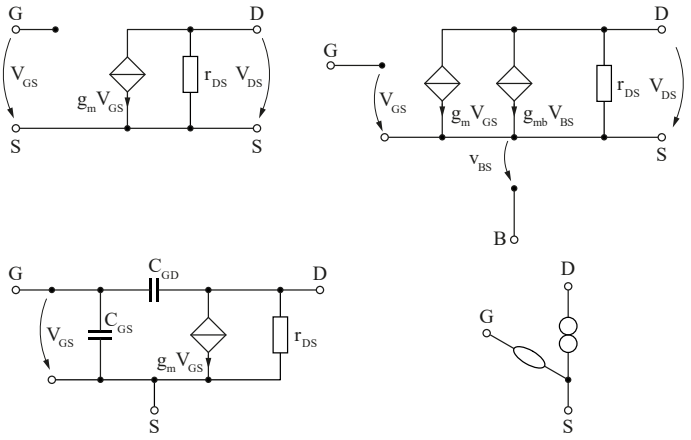


Figure 20: Small-Signal Models for hand calculations.

Other typical small-signal equivalent circuit diagrams may include those shown in Figure 20, including the g_m-only, g_m-r_{DS} and those extended by parasitic capacitances and/or the Bulk-Source induced g_{mb} if necessary and / or needed. Not well known is the Nullor model in the right lower corner, which is often beneficial and very efficient if dynamic analysis is to be performed.

10. Overview of MOS Models

Accurate modelling of MOSFET devices is critical for the design and simulation of electronic circuits, as well as for the optimization of device performance. Over the years, various MOSFET models have been developed to address the evolving requirements of semiconductor technologies. In this overview, we will discuss the development of MOSFET models, including Level 3, BSIM3, BSIM4, BSIM6, Enz-Krummenacher-Vittoz (EKV), and Penn-State Philips (PSP), highlighting their advantages and disadvantages. [22], [23]

MOS Level 1 to Level 3 are the earliest MOSFET models, utilizing simple equations to describe the basic operation of MOSFETs. The advantages of this model include simple calculations and low computational effort. In this article, we utilize the simpler Level 1 to Level 3 MOS models for a specific purpose: to obtain a straightforward understanding of MOSFET functionality and easily interpretable formula expressions for circuits. This is particularly important in the context of education, where students need to grasp fundamental concepts and relationships. Furthermore, circuit designers also rely on simple and easily interpretable relationships for qualitative circuit explanations, making the use of these basic models valuable in both teaching and practical applications. This is especially the case for small signal and frequency behaviour. However, the model suffers from inaccuracies for modern semiconductor devices, e. g., for short channel lengths and modern process technologies. It should be noted that these inaccuracies mainly refer to large signal behaviour and parasitic effects.

The BSIM3v3 model, specifically version 3.3.0, became an industry standard for accurately describing short-channel MOSFETs down to 180 nm. It accounts for various effects, such as Drain Induced Barrier Lowering (DIBL) that reduces threshold voltage with increasing V_{DS}, short-channel and narrow-channel effects impacting threshold voltage variations, mobility reduction due to vertical electric fields, and velocity saturation. Furthermore, the model considers channel length modulation, weak inversion conduction, parasitic resistances in Source and Drain regions, and "hot electron" effects that influence output resistance and threshold voltage over time.

The BSIM4 model significantly enhances the BSIM3v3 model, offering improvements in various aspects, such as better DC modelling accuracy, improved noise modelling crucial for radio frequency (RF) design, an enhanced capacitance-voltage (C-V) model for a wider range of operating conditions, and a new material model accounting for non-SiO_2 insulators, non-poly-Si Gates, and non-Si channels. These advancements make the BSIM4 model more versatile and suitable for modern process technologies and a broader array of applications compared to its predecessor.

BSIM6 is an extension of the BSIM4 MOSFET model and is a charge based symmetric MOSFET model with a charge-based core. BSIM6 has been designed to improve the accuracy of transistor simulations at nanometre scales and to model advanced device structures like FinFETs, nanowire FETs, and double-Gate MOSFETs. Some of the key differences between BSIM4 and BSIM6 include the modelling of carrier-induced voltage effects, addition of symmetry-breaking in mobility models, improved modelling of weak inversion region, and better modelling of back bias dependence. BSIM6 also provides more accurate modelling of sub-threshold slope dependence on Gate length, and improved noise models. Additionally, BSIM6 includes new parameters to

capture short-channel effects and improved models for DIBL and threshold voltage roll-off.

The EKV Model was developed for use in analogue and mixed-signal circuits, particularly in submicron CMOS technology. The model accounts for effects such as CLM, substrate resistance, and subthreshold conduction. The advantages of the EKV model include a good compromise between accuracy and computational effort, simpler mathematical formalism compared to BSIM models, and suitability for low-power applications. The model's disadvantages include not being as detailed as the newer BSIM models, especially for very short channels and cutting-edge process technologies.

The PSP Model is a compact MOSFET model intended for digital, analogue, and RF design, which is jointly developed in the early 2000s by NXP Semiconductors (formerly part of Philips) and Arizona State University (formerly at The Pennsylvania State University). [24] It was designed for improved accuracy, scalability, and predictability for advanced process technologies. The advantages of the PSP model include its physically based nature, good scalability, relatively compact structure, and less complexity than some newer BSIM models. However, the model may not provide the same level of detail as the latest BSIM models.

11. The Crux with the Small-Signal Models

As today's device models are very complex (cf. Figure 14), such as in the case of BSIM3 or higher, their structure is often based entirely on mathematical considerations instead of the underlying geometrical properties of the device. This makes interpretation of the resulting expressions more difficult, as the BSIM AC models involves transcapacitances (cf. Figure 21), i. e., differentiating VCCS instead of capacitances like C_{GS}.

These transcapacitances, which are based on charge derivatives on the various terminal voltages of the transistor may also become negative. This is difficult to be interpreted and, hence, might be confusing especially if stability problems are investigated being caused by parasitic capacitances.

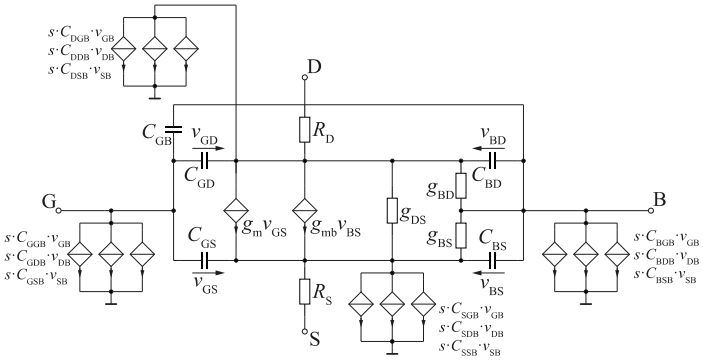


Figure 21: BSIM3 Small-Signal Model with Transcapacitances.

The operational amplifier (artificial term: OpAmp) depicted in Figure 22 is implemented in a 250 nm technology node and BSIM3v3 models were used to analyse its behaviour. The focus of this analysis was to study the sizing of the transistors for obtaining a good operating point and to investigate the AC response based on the small-signal parameters of the circuit.

In Figure 23, the output file provides information about the transcapacitances, which may sometimes exhibit negative values. While in Cadence Spectre, these values are denoted as Cxxxx (e. g., Cbgb), they are represented in Infineon's in-house simulator Titan as DQxDVyy depicted in the simulator output file in Figure 23. This naming convention represents the charge derivative at x with respect to the voltage yy. This naming is, thus, more informative, as it clearly identifies the transcapacitances as charge derivatives, and reduces confusion for those who may not be familiar with the concept of transcapacitances. It is worth mentioning that the BSIM small-signal models used to extract the transcapacitances are – in contrast to SPICE Level 1-3 – not publicly published, and different simulator providers may have different implementations. For example, some providers use branch-based derivatives, while others use node potential-based derivatives. This leads to the possibility of incompatible parameters between different simulators, adding to the difficulty of physically interpreting the impact of the transcapacitances on the frequency response of the circuit.

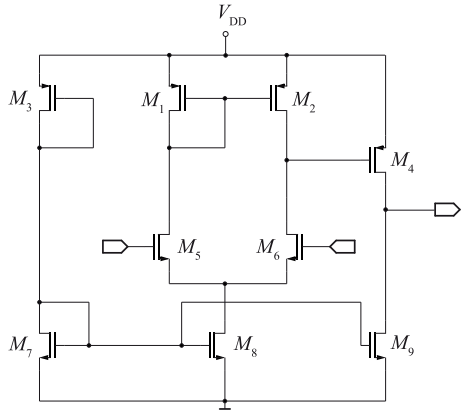


Figure 22: Operational Amplifier in 250nm Technology Node incorporating BSIM3v3 Models.

Despite these challenges, an AC analysis that provides interpretable information about the involved transistor capacitances is important for understanding the performance of OpAmps, especially in advanced technology. This information can help designers understand and adjust the frequency behaviour of their circuits, leading to more optimal designs.

```
**** MOSFETS

M_M1   MODEL: PMOD
      CBD          CBS          CBTOT        CDTOT        CGBOVL       CGD          CGDOVL       CGS          CGSOVL       CGTOT
      8.602E-015   6.737E-015   1.494E-013   2.194E-013   5.911E-026   2.593E-013   4.071E-015   4.322E-015   4.071E-015   3.197E-013
      CSTOT        DQBDVDB      DQBDVGB      DQBDVSB      DQDDVDB      DQDDVGB      DQDDVSB      DQGDVDB      DQGDVGB      DQGDVSB
      1.109E-014   -8.973E-014  -4.421E-014  -1.133E-016  1.382E-013   -1.072E-013  2.869E-016   -2.552E-013  3.115E-013   -2.511E-016
      DQSDVDB      DQSDVGB      DQSDVSB      GDS     *)   GM      *)   GMB     *)   GSTDS        GSTM         GSTMBS       IBD
      2.067E-013   -1.601E-013  7.745E-017   2.640E-008   1.716E-005   4.927E-006   -3.146E-021  -7.001E-021  -1.708E-021  0.000E+000
      IBS          ID           ISTRAT       VBS          VDS          VDSAT        VGS          VTH
      5.475E-015   1.742E-006   -1.601E-022  5.474E-001   5.474E-001   -1.685E-001  0.000E+000   -3.759E-001
...,
 M_M5   MODEL: NMOD
      CBD          CBS          CBTOT        CDTOT        CGBOVL       CGD          CGDOVL       CGS          CGSOVL       CGTOT
      6.213E-015   8.105E-015   1.507E-013   1.350E-014   5.927E-026   7.285E-015   7.282E-015   4.382E-013   7.282E-015   5.165E-013
      CSTOT        DQBDVDB      DQBDVGB      DQBDVSB      DQDDVDB      DQDDVGB      DQDDVSB      DQGDVDB      DQGDVGB      DQGDVSB
      3.016E-013   -9.090E-019  -9.025E-014  -4.611E-014  2.707E-018   -1.647E-013  1.908E-013   -2.490E-018  5.020E-013   -4.309E-013
      DQSDVDB      DQSDVGB      DQSDVSB      GDS          GM           GMB          GSTDS        GSTM         GSTMBS       IBD
      6.914E-019   -2.470E-013  2.862E-013   5.099E-008   3.753E-005   5.700E-006   5.470E-011   1.156E-010   1.785E-011   -2.753E-014
      IBS          ID           ISTRAT       VBS          VDS          VDSAT        VGS          VTH
      -1.010E-014  1.742E-006   6.676E-012   -1.010E+000  1.742E+000   7.587E-002   6.399E-001   6.087E-001
...,
```

Figure 23: Operating Point and small-signal Parameter Information for the BSIM3v3 MOSFETs (Titan Simulator).

To solve the problem, one can combine the formulas for calculating the SPICE 5-capacitance MOSFET AC equivalent circuit (cf. Figure 24) with the operating points calculated from the BSIM models to obtain reinterpretable capacitances.

However, for a hand analysis, the calculations are too complicated, so the use of a computer program is useful. Thus, such a conversion was implemented for the symbolic analysis tool Analog Insydes [2], which is based on the computer algebra program Mathematica.

Interestingly, it turned out that the SPICE Level 3 equations showed inaccuracies especially in the subthreshold area, so the formulas for the capacitances were revised using the fringing capacitances from Tsividis [16], among others. Fringing capacitances in a MOSFET are parasitic capacitances that occur at the edges of the Gate electrode, where the electric field extends beyond the edge of the physical Gate structure. These capacitances can affect the performance of the MOSFET by increasing the total Gate capacitance, which can impact the speed and power consumption of the device. An important update was the introduction of a partitioning factor (XPART) to distribute C_{ox} between C_{GS} and C_{GD}.

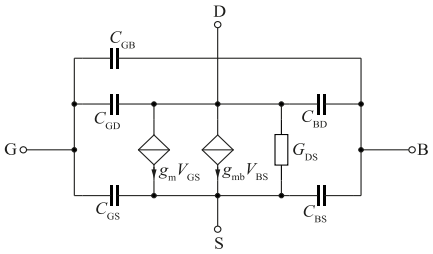


Figure 24: Simplified Level 2 SPICE 5-Capacitance MOSFET AC Model.

To show the steps involved in modifying the SPICE Level 2 5-capacitance MOSFET small-signal model (cf. Figure 24) to match the results of the BSIM small-signal model (cf. Figure 21), an industrial CMOS folded-cascode OpAmp (180 nm technology) is shown in Figure 25, and Figure 26 displays the frequency response of the OpAmp's open-loop differential-mode voltage gain. Here, the red curve shows the original simulation with the BSIM model, while the green curve shows the AC simulation performed with the SPICE Level 2 AC model, where the parameters were determined from the operating point using the full BSIM model. The agreement is very good for low frequencies but shows a significant deviation for higher frequencies.

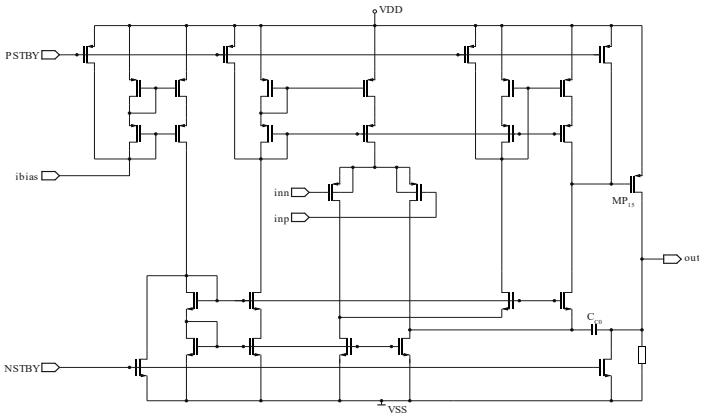


Figure 25: CMOS folded-Cascode Operational Amplifier.

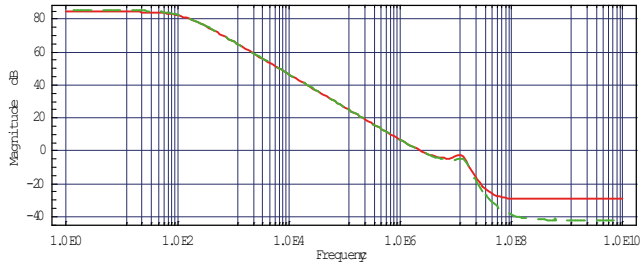


Figure 26: Frequency Response of the OpAmp's open-loop differential-mode Voltage Gain with BSIM (red) and Level 3-AC Model (green).

The deviation between BSIM model AC simulation and SPICE Level 2 AC analysis becomes even more significant for the power-supply feedthrough (PSF) characteristic in Figure 29 of the amplifier shown in Figure 27 (top-level circuit) and Figure 28 (transistor-level circuit).

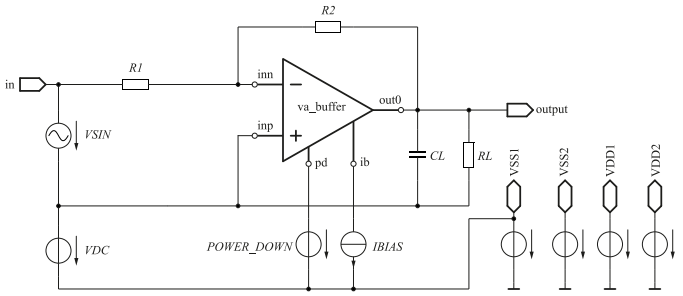


Figure 27: Top-level Circuit Schematic.

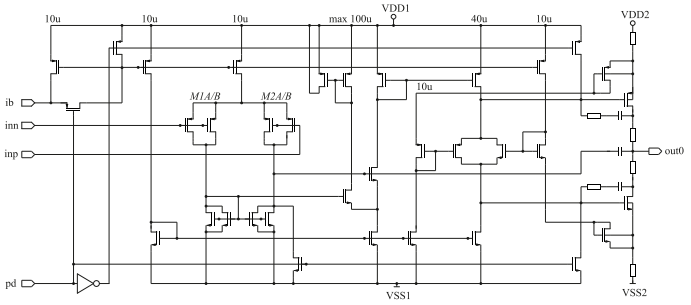


Figure 28: Transistor-level Circuit of the OpAmp with the Power supply rejection ratio (PSRR) Problem.

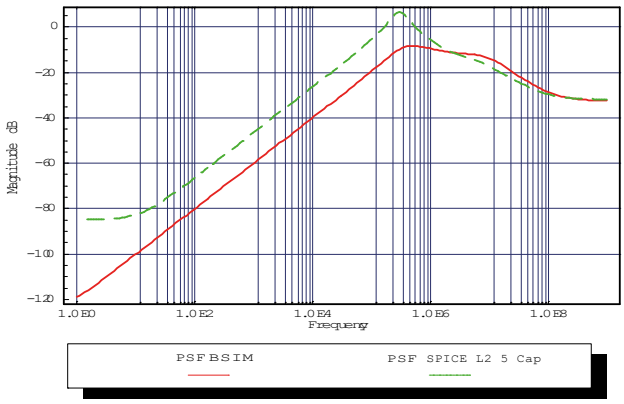


Figure 29: Frequency response of the OpAmp's power-supply feedthrough (PSF) with BSIM (red) and Level 3-AC model (green).

An analysis of the deviations led to a modification of the underlying SPICE level 2 AC model for the intrinsic Gate-Source capacitance C_{GS} and intrinsic Gate-Drain capacitance C_{GD} for the saturated MOSFET by introducing a partitioning factor XPART for C_{ox} between both capacitances.

Another problem is rooted back to the fact that SPICE Level 2 assumes perfectly pinched-off channel and no fringing capacitances. Hence, fringing capacitances C_{if} we added to the model in the subthreshold region (cf. Figure 15). Summing up all modifications the following small-signal equations were

implemented in Analog Insydes allowing symbolic analysis with interpretable results with an acceptable error.

Sub-threshold region: $V_{GS} \le V_{th}$

$$C_{GB} = C_{OX}\frac{0.5\gamma}{\sqrt{0.25\gamma^2 + V_{GB} - V_{GB}}} + CGBOVL$$
$$C_{GS} = C_{if} + CGSOVL;$$
$$C_{GD} = C_{if} + CGDOVL$$
$$C_{if} = W\frac{2}{\pi}\cdot\varepsilon_{GB}\ln\left[1+\frac{XJ}{TOX}\sin\left(\frac{2}{\pi}\frac{\varepsilon_{OX}}{\varepsilon_{Si}}\right)\right]$$

Linear region: $V_{GS} > V_{th} + V_{DS}$

$$C_{GB} = CGBOVL$$
$$C_{GS} = CGSOVL + \frac{2}{3}C_{OX}\left(1-\frac{(V_{GS}-V_{ON}-V_{DS})^2\left(4-\frac{3}{xpart}\right)}{(2(V_{GS}-V_{ON})-V_{DS})^2}\right)xpart$$
$$C_{GD} = CGDOVL + C_{if}\left(1-\frac{2C_{GD}}{C_{OX}}\right)$$
$$+\frac{2}{3}C_{OX}\left(1+\left(1-\frac{(V_{GS}-V_{ON})^2}{(2(V_{GS}-V_{ON})-V_{DS})^2}\right)\left(1+\frac{4}{3}(xpart-1)\right)-xpart\right)$$

Saturation region: $V_{th} < V_{GS}$ and $V_{GS} \le V_{th} + V_{DS}$

$$C_{GB} = CGBOVL$$
$$C_{GS} = CGSOVL + \frac{2}{3}C_{OX}\cdot xpart$$
$$C_{GD} = \frac{2}{3}C_{OX}(1-xpart) + CGDOVL + C_{if}\left(1-\frac{2C_{GD}}{C_{OX}}\right)$$

12. Application Examples using the modified SPICE Level 2 Model to solve industrial Circuit Problems in a CMOS Technology requiring BSIM MOS Models

The root cause of many industrial circuit problems is often traced back to their dynamic behaviour, particularly regarding frequency compensation and stability issues. Hence, symbolic extraction of poles and zeros is a crucial aspect of using symbolic analysis in the design of industrial integrated circuits. As designers wanted to understand the cause of the circuit problems it is essential for them to get interpretable results. Resulting in interpretable formulas, which can then be utilized to identify the appropriate frequency compensations for the circuits being analysed. This is achieved by adjusting certain circuit parameters that cause the poles to move in a manner that guarantees stability and eliminates peaking effects in the frequency response, maximizing bandwidth.

The Bode diagram (cf. Figure 26) of the OpAmp shown in Figure 25 displays a prominent peak at approximately 10 MHz, resulting from a pair of parasitic complex poles located close to the imaginary axis. Note, that in the frequency domain complex pole pairs can cause resonance peaks and phase shifts, which can affect the bandwidth and stability of the system. In the time domain, complex pole pairs can lead to oscillatory and damped responses, which can affect the settling time and overshoot of the system. The objective is to extract a simplified symbolic formula for these poles, to identify the components that significantly contribute to the peak. Therefore, it is necessary to shift the conjugate complex pole pair away from the imaginary axis, i. e., either by decreasing the imaginary part or by increasing the real part. The goal is to push the pole into the region bounded by the 45° axes in quadrants 2 and 3 of the root locus plot, cf. Figure 32 (dashed grey lines). The complexity of this issue is evident from the fact that after expanding the model, the netlist comprises of 321 primitive components, leading to a system of 29×29 modified nodal equations. Instead of the BSIM3 AC model, the modified SPICE Level 2 AC model was utilized.

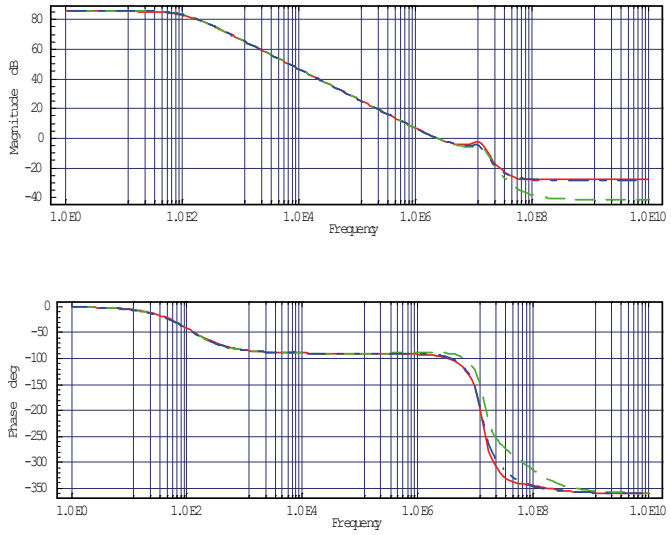


Figure 30: Frequency Response of the OpAmp's open-loop differential-mode Voltage Gain with BSIM (red) and Level 2-AC Model (green) and modified Level 2-AC Model (blue).

The differential-mode voltage transfer function, shown in Figure 30, has 19 poles and 19 zeros. In its fully expanded symbolic form, it would comprise more than a multitude of $5 \cdot 10^{19}$ product terms. Hence, to understand the cause of the peak, the symbolic extraction of the pole pair at $s = (-2.1 \pm 8.3j) \cdot 10^7$ (with j as the imaginary unit) is essential. To symbolically determine the pole pair, a symbolic approximation algorithm [25] was applied, yielding the following formula:

$$-\frac{(C_{C0}+C_L)g_{m\$MN6}}{2C_{C0}C_L} \pm \frac{\sqrt{C_{gs\$MP15}g_{m\$MN6}\left(C_{gs\$MP15}(C_{C0}+C_L)^2 g_{m\$MN6} - 4C_{C0}^2 C_L g_{m\$MP15}\right)}}{2C_{C0}C_L C_{gs\$MP15}}$$

The formula shows that, given a fixed load C_L and operating conditions, an increase in the Gate-Source capacitance of PMOS transistor MP15 $C_{gs\$MP15}$ will result in a reduction of the pole pair's imaginary components. It should be noted that altering the compensation capacitance C_{CO} will not affect the resonance peak, as its contribution is in the same order of magnitude in the numerator and the dominator in the square-root expression that yields the imaginary part of the pole pair.

Hence, reducing the imaginary part can be achieved by adding a shunt capacitor between the Gate and Source terminals of MP15 (cf. Figure 31). Figure 32 presents a root locus plot of the amplifier, calculated from the original (unsimplified) system with BSIM models as is varied from 1 pF to 10 pF. The plot confirms the

validity of the conclusion drawn from the approximated symbolic pole expression with physical interpretable small-signal capacitance.

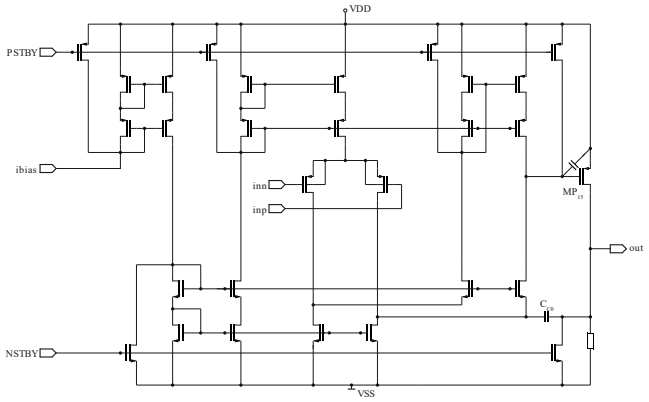


Figure 31: CMOS folded-Cascode OpAmp with Compensation Capacitor derived from symbolic Analysis.

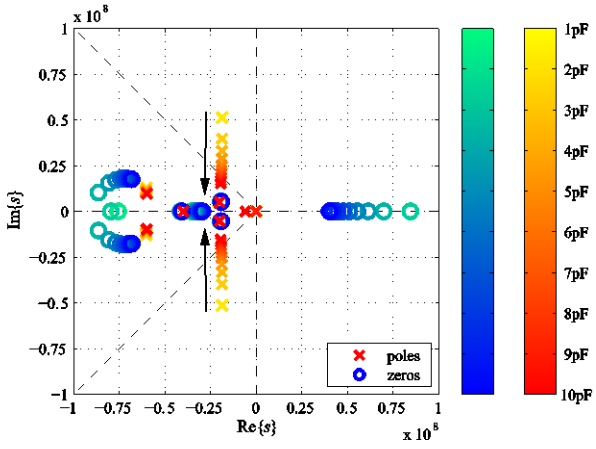


Figure 32: Root locus Analysis of the Voltage Transfer Function as $C_{GS,MP15}$ is swept from 1 pF to 10 pF.

The second example for the application of Analog Insydes [2] in industrial circuit design using the modified SPICE Level 2 model is the CMOS OpAmp, shown in Figure 25. The circuit operates as an inverting amplifier (in Figure 27).

The AC PSF characteristic of the circuit, depicted in Figure 29, demonstrates an unexpectedly large PSF value of over at 1 kHz frequency, which was predicted to be better than based on rough manual calculations by the designer. PSF occurs when unwanted signals from the power supply source leak into the output of an electronic circuit, hence it is the transfer function from the voltage supply to the output. Power supply rejection ratio (PSRR) is a measure of how well a circuit rejects ripple coming from the input power supply at various frequencies. It is defined as the ratio of the change in supply voltage to the equivalent (differential) output voltage it produces. To identify the reasons for the discrepancy and enhance power-supply rejection, Analog Insydes was again utilized for a symbolic analysis of the circuit behaviour. Solving the unapproximated equations symbolically for the transfer function from V_{DD} to the output node would result in a complex expression with over 10^{27} terms. Hence, Analog Insydes' symbolic approximation algorithms are utilized to estimate the PSF behaviour at low frequencies. With an error limit of 10 % at $f = 1$ kHz, we quickly obtained the following straightforward description of the amplifier's PSF characteristic, using only a few seconds of central processing unit (CPU) time.

$$V_{\text{outbyVDD}} = -2C_{\text{gb1}}R_2 \cdot s \cdot V_{\text{DD}}, \tag{27}$$

Additionally, this formula of the PSF can avoid misinterpretation of the real signal flow. The signal path for the PSF is not through the amplifier, but instead, it goes in reverse over the outer feedback resistance R_2 to the Bulk of the input pair. The designer neglected these Bulk capacitances in his manual calculations, as he believed them to be insignificant. That also explains that optimizing the OpAmp does not solve the PSF or PSRR problem, and oversimplifying the problem beforehand prevented the true cause of the unwanted circuit behaviour from being discovered. Once the true reason due to (27) was found, introducing a separated well for the input pair, which reduces the influence of the Gate-Bulk capacitance, solved the issue so that the PSF could be reduced. More details on the methodology especially on the generation of approximated symbolic formulas can be found in [25] and [26].

13. Conclusion

The study of MOSFETs is an important part of electrical engineering education as they are the most used components in integrated circuits. Therefore, a systematic and logical explanation of the behaviour of MOSFETs is necessary to facilitate comprehension among students in the context of circuit design. Despite the availability of substantial resources on MOSFETs, students continue to face difficulties in comprehending the concepts of charge carrier limitation and saturation. To address this challenge, a systematic and logical derivation of the Level 1 behaviour of MOSFETs is presented, incorporating simple equations and clear illustrations, to clarify the questions commonly raised by students without adding complexity through the introduction of additional semiconductor effects. Additionally, the use of modern standard MOSFET models, such as BSIM, may also pose difficulties in interpretation. A special problem is rooted back to the fact that – in contrast to the well-known AC models of SPICE Level 1 to 3 – BSIM and PSP small signal models are not published and that they include transcapacitances due to their charge-based modelling instead of physical capacitances. The acquisition of knowledge and insight regarding the behaviour of circuits is imperative for both designing circuits and resolving issues such as instability, ringing and other related phenomena. Symbolic analysis is a useful tool to facilitate this understanding, however, it is important to ensure that the formulas are comprehensible. The utilization of BSIM model and its parameters in conjunction with a modified SPICE Level 3 AC model can help in bridging the divide between technical formulas and their practical interpretation and applications. This methodology can enhance the understanding of MOSFET behaviour and its influence within the circuit design tasks for both students and experienced designers alike.

14. Acknowledgement

We would like to thank Dominik Krauße and Wladek Grabinski for providing valuable material for this contribution as well as Marius Steindel for his support in creating the illustrations and diagrams that helped to clarify and enhance our explanations of MOSFET behaviour.

15. Appendix A

Definitions of the threshold voltage V_{th} [18]:

Classical V_{th}

Conventionally, the threshold voltage is defined as the Gate voltage that must be applied to the MOS structure to reach a carrier concentration near the substrate-insulator (Si / SiO_2) interface that is equal to the majority carrier concentration in the substrate (i. e., far from the interface). In the case of an NMOS with applied threshold voltage and a p-doped substrate, this means that the electron concentration near the interface would be equal to the hole concentration in the Bulk. This condition marks the onset of the strong inversion. In particular, the band bending in the energy band model reaches a value twice as large as the Bulk potential at the location of the interface.

However, this definition is problematic due to the experimental measurement uncertainties of carrier concentration and band bending at the interface.

Extrapolated V_{th}

As can be seen in the transfer characteristic of a MOSFET, i. e., Drain current I_D plotted as a function of Gate voltage V_{GS}, the characteristic behaves almost linearly over a wide Gate voltage range in the on-state. Extrapolating the linear part of the curve to a current $I_D = 0$, the threshold voltage V_{th} can be obtained. Because the slope is not constant everywhere, consequently the threshold voltage also depends on V_{GS}. The point at which the extrapolation was applied decides the intersection location with the x-axis. To avoid this ambiguity, the point at which the slope is maximum is often selected for the linear extrapolation.

Constant current

A more pragmatic approach to defining the threshold voltage V_{th} would be to have a certain predefined current flow through the transistor at the Gate voltage $V_{GS} = V_{th}$. During this, two variants can be formulated:

The first formulation is independent of the Gate length:

$$I_D(V_{GS} = V_{th}, V_{DS}) = C_1 \cdot W , \tag{28}$$

where C_1 is the specified current threshold per unit Gate length (values around 10^{-7} A / µm) and W is the Gate width.

A second variant additionally considers the Gate length L and is given by:

$$I_D(V_{GS} = V_{th}, V_{DS}) = C_2 \cdot \frac{W}{L} , \tag{29}$$

with C_2 as a constant current specification (values around 10^{-7} A).

Constant sheet concentration

Due to the proportional dependence of the Drain current in a MOSFET on the inversion layer sheet concentration, the constant C_1 from (28) can be converted to such, as an inversion layer sheet concentration threshold.

16. Appendix B

The relation between energy E and potential φ is given by:

$$\varphi = -\frac{E}{e} , \tag{30}$$

where e is the elementary charge. Thus, the potentials corresponding to the intrinsic Fermi level E_i and the Fermi level E_F can be defined as the electrostatic potential φ_i and the Fermi potential φ_F, respectively.

$$\varphi_i = -\frac{E_i}{e} \text{ and } \varphi_F = -\frac{E_F}{e} , \tag{31}$$

17. Appendix C

The charge carrier mobility within the Si MOSFET channel can be significantly smaller than the Bulk mobility. The transport properties of the carriers are strongly influenced by the surface, in particular scattering mechanisms play a role. These can be, for example, surface acoustic phonon scattering or surface roughness scattering. Furthermore, the mobility shows a dependence on the electric field perpendicular to the surface. This can be attributed to inversion, since at higher Gate-Source voltages the electric field perpendicular to the surface becomes larger and more electrons are drawn below the surface. Under these circumstances, the electrons interact more and more with each other, which only further enhances the scattering effects at higher Gate-Source voltages. Due to this, an effective mobility is introduced which takes exactly these effects into account.

18. References

[1] C. Gatermann, R. Sommer, "Teaching the MOSFET: A Circuit Designer's View," in Proc. 18th International Conference on Synthesis, Modelling, Analysis and Simulation Methods and Applications to Circuit Design (SMACD), 2022, doi:10.1109/SMACD55068.2022.9816264.

[2] Fraunhofer ITWM, Analog Insydes, Accessed: Feb. 07, 2022, [online] available: https://www.itwm.fraunhofer.de/en/departments/sys/products-and-services/analog-insydes.html.

[3] J. E. Lilienfeld, "Method and Apparatus for Controlling Electric Currents," U.S. Patent 1 745 175, filed 1926, granted 1930.

[4] J. E. Lilienfeld, "Amplifier for Electric Currents," U.S. Patent 1 877 140, filed 1928, granted 1932.

[5] J. E. Lilienfeld, "Device for Controlling Electric Currents," U.S. Patent 1 900 018, filed 1928, granted 1933.

[6] O. Heil, "Improvements in or Relating to Electrical Amplifiers and other Control Arrangements and Devices," British Patent 439 457, filed and granted 1935.

[7] W. Shockley, G. L. Pearson, "Modulation of Conductance of Thin Films of Semiconductors by Surface Charges," Physical Review, **74**, 232, 1948, doi: 10.1103/PhysRev.74.232.

[8] J. R. Ligenza, W.G. Spitzer, "The Mechanisms for Silicon Oxidation in Steam and Oxygen," Journal of Physics and Chemistry of Solids, **14**, 131-136, 1960, doi:10.1016/0022-3697(60)90219-5.

[9] M. M. Atalla, "Semiconductor Devices Having Dielectric Coatings," U.S. Patent 3 206 670, filed 1960, granted 1965.

[10] D. Khang, M. M. Antalla, "Silicon-Silicon Dioxide Field Induced Surface Devices," in 1960 IRE-AIEE Solid-State Device Research Conference, 1960.

[11] D. Khang, "A Historical Perspective on the Development of MOS Transistors and Related Devices," IEEE Transactions Electron Devices, **23**(7), 655-657, 1976, doi:10.1109/T-ED.1976.18468.

[12] C. T. Sah, "Evolution of the MOS Transistor – From Conception to VLSI", in 1988 Proceedings of the IEEE, 1280-1326, 1988, doi: 10.1109/5.16328.

[13] D. Kahng, Applied Solid State Science Supplement 2: Silicon Integrated Circuits Part A, Academic Press, 1981.

[14] Y. Taur, T. H. Ning, Fundamentals of Modern VLSI Devices, Cambridge University Press, 2009.

[15] R. M. Warner, B. L. Grung, MOSFET Theory and Design, Oxford University Press, 1999.
[16] Y. Tsividis, Operation and Modelling of the MOS Transistor, Oxford University Press, 2012.
[17] G. N. Lewis, "The Atom and the Molecule," Journal of the American Chemical Society, **38**(4), 762-785, 1916, doi:10.1021/ja02261a002.
[18] F. Schwierz, H. Wong, J. J. Liou, Nanometer CMOS, Pan Standford Publishing, 2010.
[19] S. M. Sze, Kwok K. Ng, Physics of Semiconductor Devices, John Wiley & Sons, 2006.
[20] L. Stiny, Grundwissen Elektrotechnik und Elektronik, Springer Vieweg Wiesbaden, 2018.
[21] W. Grabinski, "EKV v2.6 Parameter Extraction Tutorial," in 2001 ICCAP Users' Web Conference, 2001.
[22] P. Antognetti, G. Massobrio, Semiconductor Device Modeling with Spice, McGraw-Hill, 1993.
[23] S. K. Saha, Compact Models for Integrated Circuit Design, CRC Press, 2016.
[24] X. Li, W. Wu, G. Gildenblat, G.D.J. Smit, A.J. Scholten, D.B.M. Klaassen, R. van Langevelde, NXP Semiconductors PSP 102.3, 2008, Technical Note: NXP-R-TN-2008/00162.
[25] R. Sommer, D. Krauße, E. Schäfer, E. Hennig, "Application of Symbolic Circuit Analysis for Failure Detection and Optimization of Industrial Integrated Circuits," in Design of Analog Circuits through Symbolic Analysis, pp. 445-477, Bentham Science Publishers, 2012, doi:10.2174/978160805095611201010445.
[26] E Hennig, R Sommer, "A reliable iterative error tracking method for approximate symbolic pole/zero analysis", in 2001 European Conference on Circuit Theory and Design (ECCTD'01), 2001, lib.tkk.fi.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 8, No. 4, 60-65 (2023)

www.astesj.com

Special Issue on Computing, Engineering and Multidisciplinary Sciences

ASTESJ

ISSN:

Proportional Derivative and Proportional Integral Derivative Controllers for Frequency Support of a Wind Turbine Generator in a Diesel Generation Mix

Abdul Ahad Jhumka[1,2], Robert Tat Fung Ah King[*,1], Chandana Ramasawmy[2], Abdel Khoodaruth[3]

[1]*Department of Electrical and Electronic Engineering, University of Mauritius, Reduit 80837, Mauritius*

[2]*Advanced Mechanical and Electrical Services Ltd., Rose-Hill 71364, Mauritius*

[3]*Department of Mechanical and Production Engineering, University of Mauritius, Reduit, 80837, Mauritius*

ARTICLE INFO

Article history:
Received: 28 February, 2023
Accepted: 28 May, 2023
Online: 25 July, 2023

Keywords:
Wind Turbine Generator
Stability
Proportional Derivative Controller
Proportional Integral Derivative Controller
Inertia

ABSTRACT

The levelized cost of electricity production is highly dependent on the cost of fuel oil on the world market. In order to reduce the dependency on the fuel oil, many countries are adopting an energy transition towards distributed generation. Distributed generation can be described as various means of generating electricity at or near where it will be used. Such generating mode can be a solar PV system, wind turbine generator and other renewable energy sources. However, it entails lots of challenges as it uses power electronics devices as the power grid interface, which causes a reduction in the system inertia and at the same time affecting the frequency, thereby affecting the stability. To enhance this stability, appropriate control measures need to be adopted. This paper brings forward a novel approach for frequency control support of a wind turbine generator (WTG) in a diesel generation mix. The novelty of this research paper explained on the concurrent application of a Proportional derivative (PD) and a Proportional Integral Derivative (PID) for speed and frequency control in a WTG. The analysis of this experimental research was carried out through the modelling of the rate of change of frequency (RoCoF) using MATLAB / Simulink software. The results showed that the use of these controllers in presence of WTG provide frequency support to the system as the frequency varied within the acceptable limit of ±0.5Hz. Additionally, this experimental research work also proved that the use of speed / governor control in form of the PID improved the RoCoF and provided an enhancement in the stability of the test system. Finally, this paper confirmed that the integration of WTG to the grid required the use of appropriate control algorithm for an efficient exploitation of this kind of renewable energy source.

1. Introduction

COP 26 conference on the global climate established the importance on the reduction of the global temperature increase to 1.5°C as a mitigating action against the greenhouse gases [1]. In line with this agreement, many countries are adopting a distributed generation policy as means of reduction of greenhouses gases. Distributed generation (DG) covers the whole spectrum of different power generating technology such as solar PV, wind energy, biomass, etc. [2]. According to [3], the exploitation of distributed generation system is emerging in the global energy market. A direct impact of this alternative form of power generation results in a lower cost of power production from renewable energy sources. In this respect, countries such as Germany and Denmark are making enormous progress in promoting the distributed renewable energy system in their generation mix [4]. Globally, solar PV and wind energy power generation are the most preferred technologies coming out in the light, with wind energy considered as the leading renewable energy source [5]. The increasing use of wind energy in the generation mix brings along a shift towards using power electronics devices as grid interface. The power electronics interface devices have undergone rapid development with semiconductor switches such as insulated gate bipolar transistor (IGBT) are now being used. A direct impact of this transition will

*Corresponding Author: Robert Tat Fung Ah King, , r.ahking@uom.ac.mu

result in reducing the inertia of the grid, which plays a primordial role in the stability of the grid. It is expected that the total inertia of the National Grid in UK will be reduced by up to 70% by 2033/34 [6], [7]. To maintain a stable power grid with wind energy in the generation mix, it is required to avoid unnecessary frequency dip owing to the stochastic nature of the wind energy source. Therefore, the frequency stability of wind turbine provides a hot topic for research.

Previous research works carried on the subject show a replete of control mechanism on the frequency support. In [8], the authors introduced a low order system frequency model with high penetration of wind power plant. This method studies the power system frequency changes during the most critical time, which is ≤ 30 s [9]. It was observed that initial conditions do not have a significant impact on the frequency response. However, the research paper in [8] highlighted on some major limitations of this experiment with regards to the effect of variable speed on the wind turbine generation system (WTGS). In [10], a comparative analysis between wind turbine generation (WTG) and a solar PV system was established. It was observed that wind turbine (WT) requires an extensive control mechanism to be able to provide a stable power as stability was not attained within the first swing. To further substantiate on the research for frequency control support for WT, an efficient control of inertia emulation and frequency support in presence of WTG was proposed in [11]. The experiment proposed a model free control (MFC) strategy for inertia emulation and frequency support of a diesel wind grid system. The MFC employs an approximation-based intelligent proportional integral derivative (PID) controller experiment. The online estimation technique [12] from input-output measurements is a key concept of MFC to approximate the complex system. It was concluded from this experiment that the use of MFC provided a precise inertia emulation and necessary frequency support. Other works on inertia emulation were carried out by [13]-[16].

In [17], the authors adopted a novel approach of using small signal analysis for frequency response of WT. The method depicts a Klein Rogers Kundur (KRK) two area, four generator system. The model was modified to accommodate an additional generator G5 on bus 13. It was concluded from this experiment that integration of wind energy without any frequency control will deteriorate the frequency response of the system. A proper control of the WTG provides necessary frequency support of the system. In [18], the feasibility of using a double fed induction generator (DFIG) to implement frequency regulation was investigated. It was observed that adjustable frequency wind turbines can undertake the frequency regulation responsibility of the power grid. A coordinated primary frequency regulation was considered as essential between the diesel generator (DG) and the WTG. In [19], an assessment of the impact of wind generation on system frequency control was made, where a time series sampling methodology was proposed over a timeframe period for assessing the impact of increased penetration of wind energy. It was concluded that future power system with an increasing penetration level of DFIG and greater levels of High Voltage DC (HVDC) interconnection will present significant frequency control challenges to system operators.

The above expose clearly shows that integrating wind energy to the grid is indeed very challenging due to a reduction of system inertia. It gives a broader picture of the research carried out in the development of the frequency support. Inspired by these obstacles, the rationale of this work was to develop a novel approach of frequency support algorithm, through the concurrent use of a Proportional Derivative (PD) controller and Proportional Integral Derivative (PID) controller.

The remaining of this research paper is structured as follows. Section 2 describes an overview of the system frequency response mechanism, while the methodology is detailed in Section 3. Section 4 showcases the simulations and results based on the principle laid down in Section 3. Section 5 treats about the discussion of the results obtained, whereas Section 6 concludes this research paper.

2. Overview

This section explains the vital role played by the frequency to maintain a stable power with the grid. As per [20], frequency stability is the ability of the power system to maintain a steady frequency following a transient occurrence, which leads to a power mismatch between generation and load. Therefore, it is required to avoid large rate of change of frequency (RoCoF) through application of necessary frequency control. The RoCoF is one of the indicators, which gives the soundness of the system frequency response (SFR).

A frequency control can act in three level steps namely [21]

1. Primary Response
2. Secondary Response
3. Tertiary Response.

Figure 1 shows the frequency response steps in the event of a loss of generation.

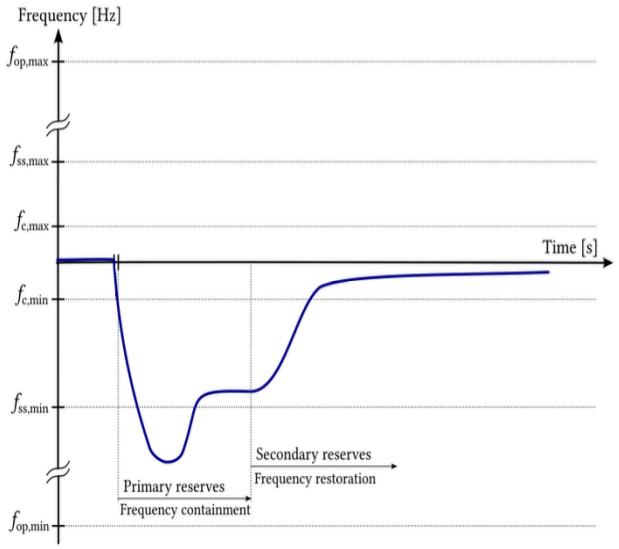


Figure 1: Frequency response in the event of a loss of generation

Another indicator, which can evaluate the system frequency response (SFR) is the frequency nadir. The frequency nadir is the minimum frequency reached during a transient period. It is therefore primordial to understand the role of frequency in maintaining a sound network particularly in presence of

renewable energy sources like WTG. The next section explains the relation between the power imbalance brought in by the frequency disturbance and the system inertia. It also shows that the RoCoF is dependent on the speed of the generator. Hence the importance of having a governor and speed control in the integration of a wind farm.

3. Methodology

Stability of power system is deeply impacted by large frequency deviation. In order to understand the impact of the frequency deviation on the power system, it is required to understand the swing equation, which gives the variation of the system inertia (H) with respect to the change in power ΔP. Equation (1) gives the swing equation.

$$\Delta P = P_m - P_e = \frac{2H}{\omega_s}\frac{d^2\delta}{dt^2} \tag{1}$$

where

ΔP: Accelerating power (pu)

δ: Rotor angle (rad)

ω_S: Angular speed (rad/s).

H: Inertia (MJ/MVA)

P_m: Mechanical power (pu)

P_e: Electrical power (pu)

According to [22], the acceleration of the prime mover caused by the unbalanced torque is governed by the equation of motion as in (2)

$$J\frac{d\omega_m}{dt} = T_a = T_m - T_e \tag{2}$$

where

J: Combined moment of inertia of generator and turbine (kg/m^2)

ω_m: Angular velocity of the rotor (mech. rad/s)

t: Time (s)

Given that kinetic energy of the rotating masses is represented by

$$E_k = \frac{1}{2}J{\omega_m}^2 \tag{3}$$

And that power is the rate of change of energy

$$\Delta P = \frac{dE}{dt} = J\frac{d\omega_m}{dt} \tag{4}$$

As

$$H = \frac{\frac{1}{2}J{\omega_m}^2}{VA_{base}} \tag{5}$$

Substituting for J in (5) into (4)

$$\Delta P = \frac{dE}{dt} = \frac{2HxVA_{base}}{{\omega_m}^2} \text{ x } (d\omega_m)/\text{ dt.} \tag{6}$$

As ω = 2πf

$$\Delta P = \frac{dE}{dt} = \frac{2HxVA_{base}}{{\omega_m}^2} \text{ x } 2\pi \text{ (df/dt).} \tag{7}$$

Hence based on (7), the power imbalance brought in by the nature of wind directly impact on the system inertia (H) and the RoCoF (df/dt). Therefore, it is required to have a low RoCoF and additional synthetic inertia to counter the change in power.

As per [23], the maximum power that can be harnessed from a WTG is governed by (8) below

$$P = 0.5\rho AV^3 \tag{8}$$

where

P: Power reaped from the WTG (W)

ρ: Density of air (kg/m^3)

A: Area of blade (m^2)

V: Velocity of wind (m/s)

Equating (8) and (4), the following expression can be obtained

$$J\frac{d\omega}{dt} = 0.5\rho AV^3 \tag{9}$$

Since ω = 2πf

$$\frac{df}{dt} = \frac{1}{4\pi}\frac{\rho AV^3}{J} \tag{10}$$

Based on (10), the wind speed directly affects the rate of change of frequency, RoCoF. Therefore, in order to exploit maximum power from the WTG, it will be necessary to provide a stable power to the grid by developing a governor and inertia control.

4. Simulations & Results

Having established the mathematical models that govern the frequency stability and power quality of a WTG, it is required to

develop a power grid with integration of a WTG. A prototype grid in form of an IEEE 9-bus, with integration of a 2 MW WTG (G3), was used as a test case with a system frequency of 50 Hz. This is shown in Figure 2.

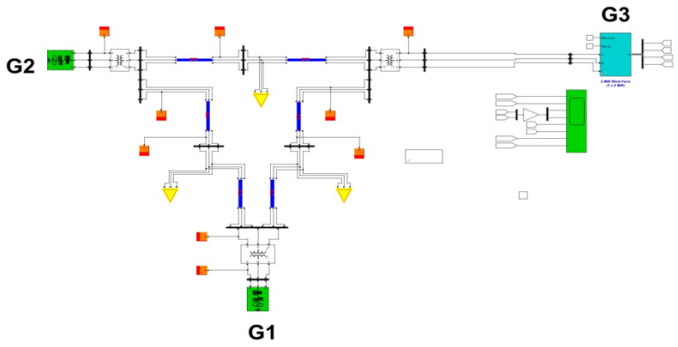


Figure 2: Integration of a Wind Turbine Generator in a generation mix

An initial experiment to assess the impact of integrating WTG, showed that large RoCoF occurred in the mix, which brought instability in the system. Figure 3 shows the unstable characteristics of the generation mix with presence of WTG.

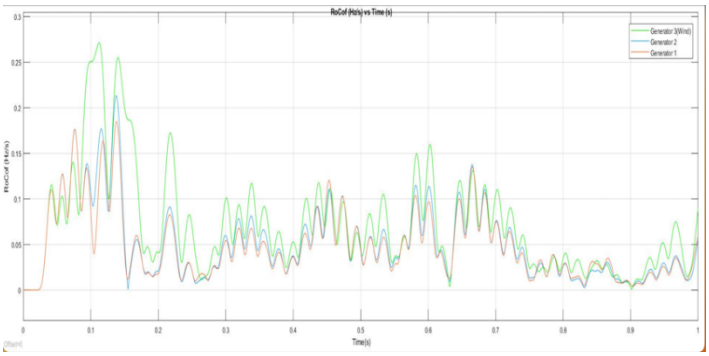


Figure 3: Large RoCoF in presence of WTG

In order to enhance the unstable nature of the system, Proportional Derivative (PD) and Proportional Integral Derivative (PID) controllers were connected in the WTG arrangement for rotor angle and speed control respectively. Such arrangement is shown in Figure 4.

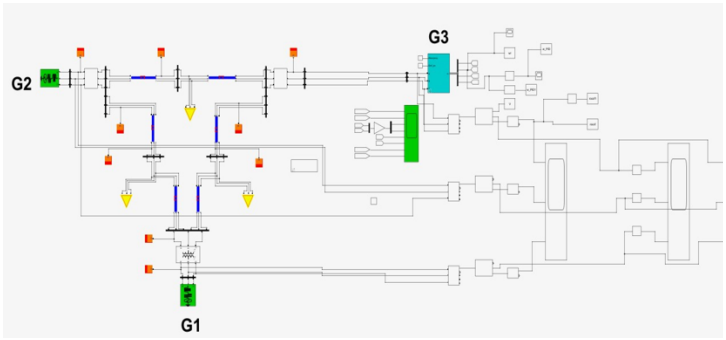


Figure 4: Arrangement of WTG with PD and PID Controllers

Proportional Derivative (PD) Controller

According to [24], a proportional derivative controller is a type of controller used where the output varies in proportion with the input signal. The block diagram of the PD controller used for the rotor angle controller is shown in Figure 5.

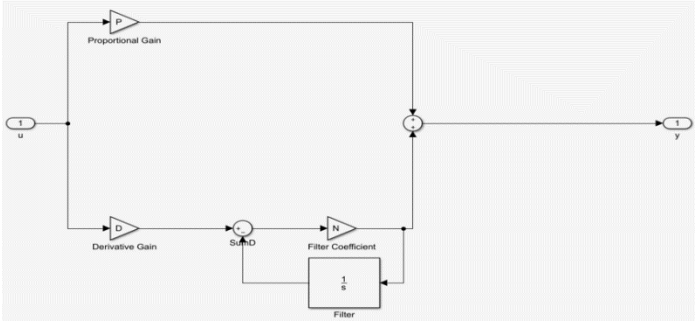


Figure 5: Block diagram of a PD controller used in a WTG

The output of the PD controller can be determined from the transfer function of the controller. The transfer function or gain of the controller gives the ratio of the output signal to the input signal. The gain of the PD controller can be calculated based on the following equation

$$G_s(s) = K_p(1 + T_d(s)). \quad (11)$$

where

K_p: Proportional gain

T_d: Time derivative

Proportional Integral Derivative (PID) Controller

A proportional integral and derivative controller (PID) is a type of controller that uses three different controller types namely the proportional, integral and derivative controllers to set the output function. A typical block diagram of a PID controller is shown in Figure 6.

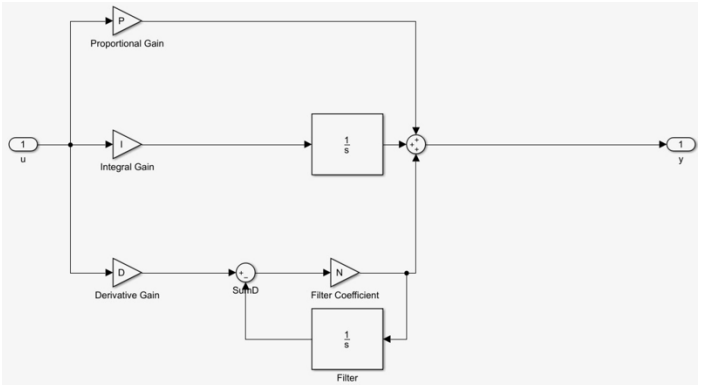


Figure 6: Block diagram of a PID controller

It works by calculating the error margin between the set point and the measured point. Similar to the PD controller, the transfer function of the PID controller will provide the desired output of the controller. The transfer function of the PID controller can be calculated as per (12).

$$G(s) = K_P (1+T_d(s) + 1/(T_d(s)). \quad (12)$$

where

K_p: Proportional gain

T_d: Time derivative

The impact of the PI and the PID controllers was assessed through the temporal variation of the RoCoF over a period of 1s. The result in Figure 7 demonstrates that the contribution of the controllers brought about a reduction in the RoCoF. Hence, an enhancement in the stability of the system.

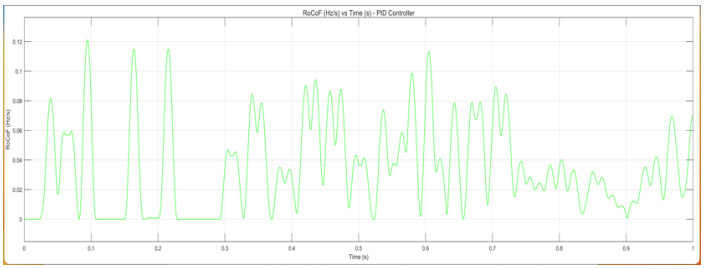


Figure 7: Temporal variation of RoCoF

Additionally, since a PID controller was used for the governor and speed control, the rate of change of frequency with respect to velocity was modelled. The contribution of the PID in the system brought a stabilization in the RoCoF. This is illustrated in Figure 8.

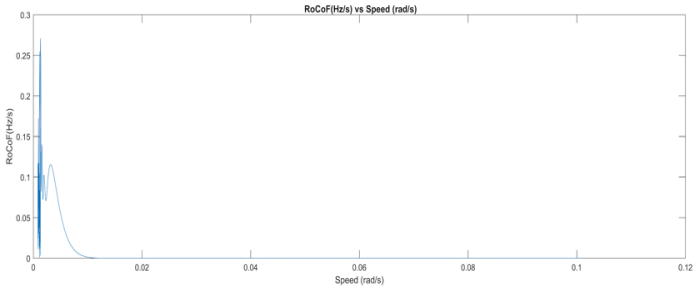


Figure 8: Variation of RoCoF with speed

5. Discussion

The objective of this research work was to develop a novel approach of using concurrently PI and PID controllers for frequency support of a WTG connected in a diesel generation mix. The test case developed in this paper showed that the controllers can provide frequency support to the system by reducing the RoCoF and stabilizing the frequency variation, within ±0.5 Hz as shown in Figure 9.

Additionally, this research work showed that this method proved to be more accurate than [25]. The virtual inertia control developed in [25] as a frequency support resulted in a frequency variation of ±0.8 Hz. Therefore, the concurrent use of PID and PD controllers may be the favoured way in enhancing the stability of a diesel and WTG generation mix.

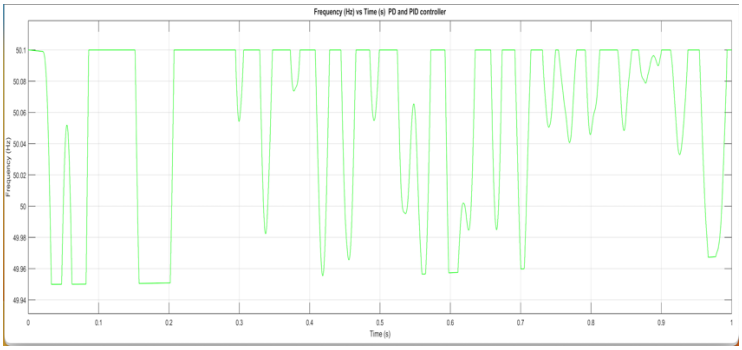


Figure 9: Temporal variation of frequency in presence of PI and PID Controllers

6. Conclusion

This paper discussed about the challenges of integrating a WTG to the grid. It also highlighted the role played by the frequency in arresting the extent of disturbance caused by the penetration of the WTG. Since a WTG is dependent on the external weather factor, any change in speed will result in a high rate of change of frequency, which will lead to power outage.

This research work has also demonstrated that the synchronization of a WTG to a grid is very challenging as it requires extensive control, in the form of PI and PID controllers. A combination of these control strategies constitutes the novelty in the frequency support, through the confirmation of the experimental results obtained. Therefore, it can be concluded that an extensive control concept is mandatory for an efficient exploitation of wind energy. However, the drawback in exploiting such type of renewable energy system will lie in the cost of investment.

This research work can be further extended to the analysis of multi-swing stability.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

This research work was assisted by Advanced Mechanical and Electrical Services Ltd. as part of the company's research policy.

References

[1] M. Denchak, "Paris Climate Agreement: Everything You Need to Know," NRDC, 2021, [Online], https://www.nrdc.org/stories/paris-climate-agreement-everything-you-need-know#sec-summary.

[2]. Environmental Protection Agency, "Distributed Generation of Electricity and its Environmental Impacts," https://www.epa.gov/energy/distributed-generation-electricity-and-its-environmental-impacts (Accessed: December 28, 2022).

[3]. REN21, "Renewable Energy Policy Network for the 21st Century," Renewables—Global Status Report; REN21 Secretariat: Paris, France, 2017.

[4] C.D. Iweh, S. Gyamfi, E. Tanyi, E. Effah-Donyina, "Distributed generation and renewable energy integration into the grid: Prerequisites, push factors, practical options, issues and merits," Energies, **14**(17), 5375, 2021 https://doi.org/10.3390/en14175375.

[5] H. Ibrahim, M. Ghandour, M. Dimitrova, A. Ilinca, J. Perron, "Integration of wind energy into electricity systems: Technical challenges and actual

solutions," Energy Procedia, **6**, 815–824, 2011, https://doi.org/10.1016/j.egypro.2011.05.092.
[6] National Grid, "UK Future Energy Scenarios," 2018, accessed: 2019- 05-15, [Online], http://fes.nationalgrid.com/media/1363/ fes- interactive- version-final.pdf
[7] E. Rakhshani, D. Gusain, V. Sewdien, J. L. Rueda Torres, M. A. M. M. Van Der Meijden, "A key performance indicator to assess the frequency stability of wind generation dominated power system," IEEE Access, **7**, 130 957–130 969, 2019.
[8] M. Krpan, I. Kuzle, "Introducing low-order system frequency response modelling of a future power system with high penetration of wind power plants with frequency support capabilities," IET Renewable Power Generation, **12**(13), 1453–1461, 2018, https://doi.org/10.1049/iet-rpg.2017.0811.
[9] P.M. Anderson, M. Mirheydar, "A low-order system frequency response model," *IEEE Trans. Power Syst.*, **5**(3), 720–729, 1990, doi: 10.1109/59.65898.
[10] A.A. Jhumka, R.T.F. Ah King, A. Khoodaruth, C. Ramasawmy, "Comparative Performance Analysis of Solar Energy and Wind Energy Systems using Rotor Angle Stability," 2022 IEEE 7th International Energy Conference (ENERGYCON), Riga, Latvia, 2022, 1-5, doi: 10.1109/ENERGYCON53164.2022.9830452.
[11] U. Datta, J. Shi, A. Kalam, "Primary frequency control of a microgrid with integrated dynamic sectional droop and fuzzy based pitch angle control," International Journal of Electrical Power & Energy Systems, **111**, 248–259, 2019, https://doi.org/10.1016/j.ijepes.2019.04.001.
[12] M. Fliess, C. Join, H, Sira-Ramirez, "Non-linear estimation is easy", Int. J. Modell. Identification Control, **4**(1), 12-27, 2008, https://doi.org/10.1504/IJMIC.2008.020996.
[13] J Morren, S.W. Haan, L. Kling, J. Ferreira, "Wind turbines emulating inertia and supporting primary frequency control," IEEE Trans. Power Syst. **21**(1), 433-434, 2006, doi: 10.1109/TPWRS.2005.861956.
[14] J.M. Mauricio, A. Marano, A. Gomez-Exposito, J.L. Martinez-Ramos, "Frequency regulation contribution through variable-speed wind energy conversion systems," IEEE Trans. Power Syst. **24**(1), 173-180, 2009, doi: 10.1109/TPWRS.2008.2009398.
[15] B. Wang, Y. Zhang, K. Sun, K. Tomsovic, "Quantifying the synthetic inertia and load-damping effect of a converter-interfaced power source," Proc. IEEE Int. Energy Conf. (ENERGYCON), Limassol, Cyprus, 1-6, 2018, doi: 10.1109/ENERGYCON.2018.8398838.
[16] S. Wang, K. Tomsovic, "A novel active power control framework for wind turbine generators to improve frequency response," IEEE Trans. Power Syst. **33**(6), 6579-6589, 2018, doi: 10.1109/TPWRS.2018.2829748.
[17] B. Park, Y. Zhang, M. Olama, T. Kuruganti, "Model-free control for frequency response support in microgrids utilizing wind turbines," Electric Power Systems Research, **194**, 107080, 2021, https://doi.org/10.1016/j.epsr.2021.107080.
[18] C. Guo, D. Wang, "Frequency regulation and coordinated control for Complex Wind Power Systems," Complexity, 1–12, 2019, https://doi.org/10.1155/2019/8525397.
[19] R. Doherty, A. Mullane, G. Nolan, D. J. Burke, A. Bryson, M. O'Malley, "An assessment of the impact of wind generation on system frequency control," IEEE Transactions on Power Systems, **25**(1), 452–460, 2010, doi: 10.1109/TPWRS.2009.2030348.
[20] P. Kundur, J. Paserba, V. Ajjarapu, G. Andersson, A. Bose, C. Canizares, N. Hatziargyriou, D. Hill, A. Stankovic, C. Taylor, T. Van Cutsem, V. Vittal, "Definition and Classification of Power System Stability," IEEE Trans. Power Syst., **19**(3), 1387–1401, 2004, doi: 10.1109/TPWRS.2004.825981.
[21] P. Fernández-Bustamante, O. Barambones, I. Calvo, C. Napole, M. Derbeli, "Provision of frequency response from Wind Farms: A Review," Energies, **14**(20), 6689, 2021, https://doi.org/10.3390/en14206689.
[22] P. Kundur, N.J. Balu, M.G. Lauby, Power system stability and control, McGraw-Hill. 1994.
[23] J.F. Manwell, J.G. McGowan, A.L. Rogers, Wind Energy Explained: Theory, Design and Application, Wiley: Hoboken, NJ, USA, 2010.
[24] S. Bashetty, J.I. Guillamon, S.S. Mutnuri, S. Ozcelik, "Design of a robust adaptive controller for the pitch and torque control of wind turbines," Energies, **13**(5), 1195, 2020, https://doi.org/10.3390/en13051195.
[25] D. Yang, E. Jin, J. You, L. Hua, "Dynamic Frequency Support from a DFIG-Based Wind Turbine Generator via Virtual Inertia Control," Appl. Sci., **10**(10), 3376, 2020, https://doi.org/10.3390/app10103376.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 8, No. 4, 66-72 (2023)

www.astesj.com

Special Issue on Computing, Engineering and Multidisciplinary Sciences

ASTESJ

ISSN: 2415-6698

Distribution Management Problem: Heuristic Solution for Vehicle Routing Problem with Time Windows (VRPTW) in the Moroccan Petroleum Sector

Younes Fakhradine El Bahi[1], Latifa Ezzine[2], Zineb Aman[*,1], Imane Moussaoui[3], Miloud Rahmoune[3], Haj El Moussami[1]

[1]*Mechanics & Integrated Engineering, ENSAM School, Meknes, 50000, Morocco*

[2]*Modeling, Control Systems and Telecommunications, EST-Meknes, Moulay Ismail University, Meknes, 50000, Morocco*

[3]*Advanced Materials Studies and Applications, EST-Meknes, Moulay Ismail University, Meknes, 50000, Morocco*

ARTICLE INFO

Article history:
Received:26 February, 2023
Accepted:4 June, 2023
Online: 25 July, 2023

Keywords:
Heuristic
VRPTW
Petroleum Sector

ABSTRACT

The attributes of the vehicle routing problem (VRP) are as many additional characteristics or constraints which aim to better take into account the specificities of real application cs. The variants of the VRP thus formed are the support of an extremely rich literature, comprising an immense variety of heuristics. This article constitutes an industrial application and an objective synthesis of successful and challenging heuristic concepts for time-windowed VRP problems. The purpose will be to minimize transport costs and determining the optimal number of trucks by applying a transport algorithm. The results show that the solution method should help to increase the competitiveness of transportation operations in this important economic sector.

1. Introduction

Oil is one of the most important raw materials in the world. It has been the primary energy source since the mid-1950s. The oil and gas industry plays a vital role as the engine of the global economy. The products produced by this industry support many other vital industries like automotive industry and manufacturing industry.

Changing technologies, markets and customer needs affect the competitiveness of companies, which requires a continuous restructuring of the strategy and positioning tactics of the oil industry. Currently, the main problem facing the oil industry is to minimize transportation costs and production costs. Effective supply chain management can increase overall efficiency, competitiveness, and good sourcing.

The economic importance of the transport sector explains the interest of researchers in the problems of vehicle routing Laporte [1-3]. Much attention has been paid to the development of route definition techniques as they have the potential for major cost savings.

The gas station supply problem consists in determining how to optimally distribute several products (mainly gasoline and diesel fuel) to customers (gas stations) from a depot (refinery or regional depot) according to a planning horizon chosen. It is therefore necessary to determine for each period of the planning horizon, the stations to be visited, the routes and the trucks to be used, the quantity to be delivered of each product and their assignment to the trucks, and this, in ensuring that no client station is out of stock.

The Vehicle Routing Problem with Time Windows (VRPTW) is a major operational research problem. It consists of determining a set of routes to collect or deliver goods to customers within time windows and capacity constraints. Customers are usually visited once and only once by a vehicle; their request must be satisfied by this single visit.

The article is divided into four sections. After the introduction, section 2 presents a review of the literature on vehicle routing problems especially VRPTW. Section 3 offers a presentation of our case study. And Section 4 presents our conclusions and directions for future research.

conformity of style throughout a conference proceedings. Margins, column widths, line spacing, and type styles are built-in; examples of the type styles are provided throughout this document and are identified in italic type, within parentheses, following the example. Some components, such as multi-leveled equations, graphics, and tables are not prescribed, although the various table text styles are provided. The formatter will need to create these components, incorporating the applicable criteria that follow.

[*]Corresponding Author: Zineb Aman, ENSAM School, zineb.aman@gmail.com

2. Literature review

Vehicle routing problems are much studied because of the growing importance of passenger and freight transportation today. The simplest and best-known problem of this type is the Traveling Salesman Problem (TSP). It consists of determining a route routed by a vehicle, so as to serve a set of customers distributed in a network at minimal cost. This basic model can be enriched by various constraints relating to the number of vehicles, to their loads, or to constraints relating to the nodes, to their time windows, or to dependencies between them.

Several researches are mainly oriented towards solving the vehicle routing problem VRP (Vehicle Routing Problem). The latter is a problem of optimizing vehicle routes to satisfy transport demands. These vehicle routing problems are generally subject to several types of constraints.

The general vehicle routing problem is known as the Vehicle Routing Problem (VRP) and represents a multi-objective combinatorial optimization problem that has been the subject of many works and many variants in the literature. It belongs to the NP-hard category [4-6].

The VRPTW constitutes a generalization of the VRP insofar as we also introduce a temporal constraint on the requested service. Each customer has a window of time within which he wishes to be served. The central depot also has a time window which we commonly refer to as the service horizon or open time of the day. Its role is to set a time slot during which vehicles can make their rounds.

These time constraints will make it necessary to use several vehicles to satisfy all customers over the service horizon. We may want to limit the number of vehicles to be used and in this case customers may not be served [7].

Each customer must be served within a defined time interval, known in advance by the delivery person and any violation of this constraint may result in a penalty. When the time window constraint is not satisfied, either we reject the solution if we consider the rigid case or we construct a penalty function which will be added or combined with the objective function for the case released. In fact, this is a very common problem. Distribution of perishable products (milk, meat, etc.), newspapers, outpatient services, . . . are practical examples of the VRPTW. Within this class of problems, there are two subclasses:

- The rigid PTVFT where the service must imperatively be performed within the time window.
- The released PTVFT or the delay or the advance only generates a penalty [7].

A time window (TW - Time Window) imposes that one or more problem requests be processed respecting an interval for the start of vehicle processing (delivery, collection). This interval is defined by an earliest date and a latest date.

This constraint can apply to pickup or delivery requests [8]. This type of constraint most often involves allowing the vehicle to arrive early, in which case it waits on the vertex in order to start its operations at the start of the time window. This waiting can be authorized on all the vertices of the graph or, only on a part of the vertices. However, the vehicle is not allowed to arrive after the end of the time window. In the classical version of the constraint, if one of the time windows is not respected then the solution is not valid. There are variants in the literature, for example the Soft Time Window (STW) introduced for the first time on a TSP problem [9], then on a VRP problem [10], which allows violation of the time window but penalizes the objective function [8].

The vehicle routing problem (with time windows) is one of the routing problems. Lenstra and Rinnooy Kan proved that VRP is an NP-DUR problem. Its resolution by an exact method turns out to be inappropriate for large instances. It is therefore inevitable to proceed to its resolution by heuristic approaches, which provide feasible and appreciable solutions in a reasonable time [11]. Recently, researchers treat VRPTW by different ways [12-18].

3. Case study

This part evaluates the effectiveness of the solutions proposed at the level of our previous work [19], and this is done by adding more constraints on the model which will make it more real and better solve the problem of distribution of petroleum products of the company.

This work makes a contribution in two ways. First, we deal with an actual case that differs slightly from the scenario that was previously addressed. Trucks, for instance, operate in shifts with predetermined loading and distribution window times. In contrast to earlier language, shift start and finish times varied from truck to truck.

The ultimate goal in this situation is to maximize customer order fulfillment while adhering to a specific priority and using a small, heterogeneous fleet. Additionally, because each journey has a distinct duration and cost, the charges are unaffected by the amount of time that passes between the start of the first trip and its conclusion.

By forcing the execution of the orders, the priority of the consumers may have been managed; however, because the orders can only be concentrated for a limited period of time, the issue is insoluble.

Our second contribution consists of a heuristic suggestion for this particular instance, the major goal of which is to quickly arrive at a workable operational solution so that people in charge of supply chain management can assist it.

This heuristic is an adaptation of Solomon's insertion heuristic for problems involving vehicle scheduling or routing with time windows.

Since each customer has value that is not always reflected in the net benefit of a transaction, looking at it from the perspective of customer compliance offers a more comprehensive perspective than merely cost reduction. In the case of fuels, this is crucial because the profit per transaction is generally low and the gain is based on the volume sold. As a result, it could be crucial to prioritize a customer who makes a lot of orders.

In this section, we'll explicitly characterize the issue at hand and provide some premises and traits that will affect how we approach finding a solution.

3.1. Problem definition

The VRPTW used in this real case can be defined as follows:

Let $G = (\{0, n + 1\} \cup C, A)$ be a directed graph where:

$\{0, n + 1\}$ is the deposit,

$C = \{1, \ldots, n\}$ is the set of customers,

$A = \{(i, j) : i, j \in \{0, n + 1\} \cup C, i \neq j\}$ is the set of arcs. Each arc (i, j) is associated with a travel time ti,j. Similarly, the service time si at depot or station i, where $i \in \{0, n + 1\} \cup C$, is known. The set of trucks associated with the depot is denoted K.

Customer i orders are made up of Pi order lines, where each line specifies a different product type, bucket type to use, and ideal time window. Trucks must arrive at customer i within the proposed time window $[ai, bi]$. Also, each truck must respect their team's time window $[\alpha'k, \beta'k]$.

It is crucial to strike a balance between the aspects of reality that we will model and those that we will simplify through assumptions because of the true nature of the problem, which frequently results in instances where theory and reality diverge.

The model's route generating step contains one of its most significant simplifications. Travel time is directly related to the zone a customer is in because customers are organized by zones. Generally speaking, trucks that visit multiple stations should only go to those that are nearby. Orders may be clubbed together for the purpose of assigning them to one and only one feasible route due to this and other limitations. In relation to capacity, the resulting routes will only have one type of vehicle, allowing us to solve the problem for each type of truck separately. Orders that do not comply with these rules are left out of this issue, as the compliance decision is in the hands of the planners.

- It's interesting to see that less than 1% of all orders are frequently refused. Following that, more crucial presumptions for this specific issue are noted:
- Orders are made up of order lines, which each client can create until a truck is full. More than one product type may be present in these orders, but only one may be present in each compartment.
- Customers are permitted to create multiple orders. Each of these is handled independently because it has its own specifications and due dates.
- All deliveries must be made in a single trip.
- More than one consumer may accompany a journey, but no more than two.
- Trucks may operate many shifts throughout the day. They are viewed differently because they are independent.
- There is a relevance associated with the priority of a command, this will be represented by the constant ρ.

3.2. Resolution process

We divided the problem into two phases in order to solve it in a reasonable amount of computation time:

- Phase I: Creation of requests on a monthly basis (in our example, the month of July).

- Phase II: Scheduling and truck assignments based on the VRPTW mathematical model. The following notations will be used in addition to the parameters that have already been defined:

- □ Hints:

t, l: Road index

k: Truck index

v : Trip position index

- □ Sets:

T : Road set

Tk : Set of routes associated with truck k

V : Set of possible path indices. $V = \{1,2, \ldots, N\}$

Vk : Set of possible route clues for truck k

- □ Parameters:

Di : D$emand$ of $station\ i$

L : Number of $stations$

dij : D$istance$

C : Truck $capaci$ty

Ri: Residual demand < 33, which can be delivered in a second time

$Zt = \{i, j\}$ / $t \in [1: n]$: cluster of station i and j that we will deliver them in one truck

(the clusters must group together a maximum of 2 stations) and n is the total number of clusters

T_s: Total demand of day S

$n_{i/s}$: Customer i should be delivered the day s

n_{ij}: Number of full trucks to deliver for a customer i which will define the frequency / duration in which I can deliver a single truck to a station. Example: If nij =2 truck then the frequency is 1 month, nij = 4 , frequency) 1 truck per week. nij> 4, frequency every 2 days..)

$B_s = Ts/I$: Number of trucks to be loaded at a time

N : Maximum number of possible trips during a shift

ρ_t : Value associated with the priority of path t

α_t : Start time of route time window t

β_t : End time of route time window t

λ_t : Route duration t

α_k : Start date of truck time window k

β_k : End date of truck time window k

Next, we describe the two phases of this problem and how we deal with them.

Phase I : Generation of requests according to a monthly schedule (the month of July in our case)

This first phase is considered an initialization process.

It consists of generating the delivery schedule for monthly requests (the month of July in our case) for all the service stations, as well as the total number of trips during this month and subsequently the number of trucks to rent. The aggregation of orders in the trucks must follow certain rules of the company. This allows us to group commands so that they are assigned to one and only one possible route.

- Depot: working hours (from 5 a.m. to 1 p.m. // 2 p.m. to 4 p.m.) 26 days a month:

It is a common depot between 5 companies located in Mohammadia, it contains 17 truck loading stations (full load). The number of trucks to be loaded at a time varies from one company to another with a loading time of 1 hour.

Shell 17 trucks at a time;

OLA ENERGY 10 trucks simultaneously with Winxo 7 Trucks;

Total 12 simultaneous with Petruom 5 Trucks.

The order of passage is done in turn each week, which then defines the maximum number of possible loadings. Let I be the maximum number of trips to be made in one day for our company:

Week 1: Ola is the first: 4 trips *10 = 40 trucks

Week 2: Ola is second: (I = 3 trips) *10 = 30 trucks

Week 3: Ola is third: (I = 3 trips) *10 = 30 trucks

Week 4: Ola is the first: (I = 4 trips) *10 = 40 trucks.

- Customer storage capacity:

Each customer has a monthly demand and storage capacity. If the capacity is less than 33T, we cannot deliver a full truck so we split the complete order into 2 orders or more. Otherwise, we deliver a complete truck.

- Algorithm :

1- $Di\ /C\ =\ Ni$

2- $Ni - E(Ni)\ =\ Ri,\ 26\ /\ E(Ni)\ =\ nij$ and $j\ \in \{1, \dots, 26\}$

3- $Cluster\ stations\ :\{Ri + Rj = 1(1\ camion\ )$ and $dij \leq 30km\ \rightarrow\ \square Zn$

4- $Ts\ =\ \sum\ n_{i/S}$ for $s\ =\ 1:26$

$n_{i/S} = 1\ if\ s\ modulo\ nij = 0\ \&\ n_{i/S} = 0\ else$

$If\ (n_{i/S} \neq 0\ \&\ n_{j/S} \neq 0)\ Avec\ i, j\ \in Zt;\ \ t \in [1:n]$

$Ts = Ts + 1$ (We add the trucks that will deliver the residual of the cluster) $\&[1:n] = [1:n] - 1$ (We subtract the cluster that we will deliver on the day s)

$Break$

5- $Bs\ =\ Ts\ /\ I$

$If\ Bs > 10$ so $As\ =\ I\ *10$

$Xs = Ts\ - As$

(Xs it's the deliveries that I have to deliver but I don't have the right for the day s, so I will memorize these deliveries and satisfy them the day I have more trucks)

6- $Xs\ =\ XX_{S-1} + Xs$

$For\ p\ =\ 1 : Xs$

$If\ Bs <\ 10$

$Bs\ =\ Bs\ + 1$

$Xs\ =\ Xs\ - 1$

$Nc\ =\ max\ Bs\ for\ \{s = 1, \dots, 26\}$

Phase II : Truck assignment and scheduling from the VRPTW mathematical model

The company's primary goal is to maximize the trucks and trips that stop at the most service stations while adhering to their time restrictions, each station's priority weighted accordingly.

The assignment problem's goal function can be stated as follows: Maximize:

$$\sum_{t \in T} \sum_{v \in V} \sum_{k \in K} \rho t x_{tvk}\ , \qquad (1)$$

The binary variable xtvk has a value of 1 if route t corresponds to the v-th trip of truck k and a value of 0 otherwise. We employed a sequential insertion heuristic, which distributes roads to one truck at a time until the truck can no longer accept another one, because addressing this problem using exact approaches proved challenging with a huge number of trucks and roads.

For each route of the truck, the starting position and time must be defined, which leads to the scheduling problem. The following model deals with route scheduling for each truck k:

- Variables :

x_{tv}: binary variable which takes the value 1 if the route t corresponds to the v-th trip of truck k; 0 otherwise.

d_v: departure date of the truck's v-th trip k

Minimize: $\sum_{v \in Vk}\ dv \qquad (2)$

Under constraints of:

$$\sum_{v \in Vk} xt,v = 1, \quad \forall t \in Tk \quad (3)$$

$$\sum_{t \in Tk} xt,v = 1, \forall v \in Vk \quad (4)$$

$$\sum_{t \in Tk} at\, xt,v \leq dv \leq \sum_{t \in Tk} \beta txt,v, \quad (5)$$

$$\forall v \in Vk$$

$$dv \geq dv - 1 + \sum_{t \in Tk} \lambda txt, v - 1, (6)$$

$$\forall v \in Vk \mid v \neq 1$$

$$\alpha' k \leq d1 \, (7)$$

$$dn(Vk) \leq \beta' k - \sum_{t \in Tk} \lambda txt, n(Vk) \, (8)$$

$$xt,v \in \{0,1\}, \quad \forall t \in Tk, \ \forall v \in Vk \, (9)$$

$$dv \in \mathbb{R}, \forall v \in Vk \, (10)$$

The goal function (2), which relates to a supplementary optimization criterion, seeks to decrease departure times while prioritizing the shortest routes before the longest ones. Truck k will only make one of the trips given to it, thanks to constraint (3). Each travel position will only be able to handle one route, according to constraint (4). Each route must adhere to its time frame if it corresponds to the v-th trip, according to constraint (5). The start time of the v-th trip will begin after the previous trip has ended, according to constraint (6). The first journey is required to adhere to the truck's departure time per constraint (7). The last trip must finish before the truck's end time k, according to constraint (8). Lastly, limitations (9) and (10) show the nature of the variables.

The sequential insertion heuristic we propose can be expressed as follows:

For each truck:

1. Populate a list with all unassigned routes.

2. Sort routes in descending order of priority.

3. Filter routes based on truck time window compatibility.

4. Assign the truck a route from the list that meets a certain departure criteria.

5. For each item in the list:

(A) Verify compatibility with the allocated routes as of right now. If the (a) check is positive, solve the scheduling problem including the current route.

If not, proceed to the next thing on the list.

(b) If the scheduling issue has been correctly resolved, assign the current route to the truck, remove it from the list, and determine if any other possible routes are still available. If not, proceed to the next thing on the list. (c) Go to the next item on the list if there are any suitable paths. If not, proceed to the next truck.

Once all vehicles have been moved or all routes have been assigned, the process is complete.

On real data from the month of July, we tested our heuristic.

Each day's orders and truck involvement varied, allowing a wide range of potential scenarios to be covered.

The Python code for our heuristic was created using Jupyter Anaconda Navigator on a Windows 10 PC with an Intel Core i5-4200U processor clocked at 1.6 to 2.3 GHz and 4GB of RAM.

Table 1 and Figure 1 present the findings.

Our heuristic therefore allowed us to plan the path or the circuit of truck 1 for example in an optimized way.

Table 1: Planning extract for Truck 1

	Truck 1					
	ID	**Day**	**Arrival Time**	**Departure Time**	**Delivery Volume**	**Distance traveled**
Deposit	LOMA0662668	Thursday		07:00		0
Station-service 6	LOMA0672999	Thursday	08:00	09:30:00	33	18
Deposit	LOMA0793001	Thursday	10:53	12:00		104
Station-service 11	LOMA0662997	Thursday	12:59	13:29	16	124
Station-service 28	LOMA0662974	Thursday	13:36	14:06	17	127
Deposit	LOMA0662668	Thursday	15:16		0	304

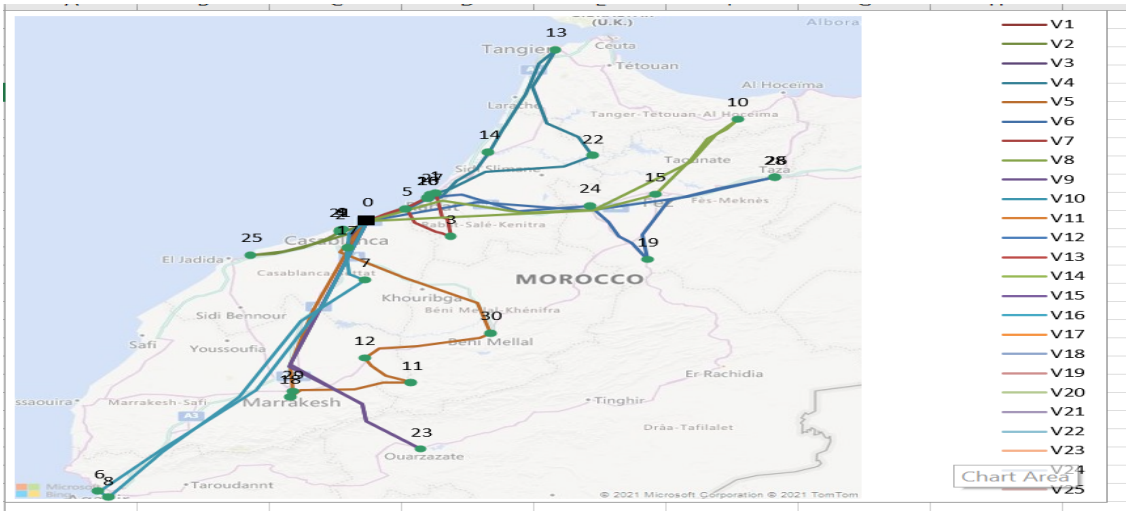


Figure 1: Truck route for the first day of July

4. Conclusion

Our heuristic produces a suggestion that planners typically ask for in the morning or throughout the day, particularly when significant alterations to the current schedule are required. As a result, they must be able to quickly deliver a solid initial plan. The greatest cases of this problem can be handled by the heuristic in a maximum of 10 minutes, which is a respectable amount of time given the compliance attained.

According to the data, it is possible to achieve at least the same level of compliance as the company, thus it is possible to include the truck teams' time slots in the scheduling issue the company confronts.

Future research will concentrate on the impact of more real-world factors in the two phases mentioned, testing the heuristic's various aspects to see if it can solve them, including: the environmental factor, worker safety, road quality, and the efficiency of using secondary optimization criteria on the solution. Finally, testing various route departure criteria may help the heuristic perform better.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

I am grateful to all of those with whom I have had the pleasure to work during this and other related projects.

References

[1] "The vehicle routing problem: An overview of exact and approximate algorithms", European Journal of Operational Research, **59**, 345-358, 1992, doi.org/10.1016/0377-2217(92)90192-C.

[2] G. Laporte, "The traveling salesman problem: An overview of exact and approximate algorithms", European Journal of Operational Research, **59**, 231-247, 1992, doi.org/10.1016/0377-2217(92)90138-Y.

[3] G. Laporte, I. H. Osman, "Routing problems: A Bibliography", Annals of Operations Research, **61**, 227-262, 1995, doi.org/10.1007/BF02098290.

[4] N. Christofides, A. Mingozzi, P. Toth, "The vehicle routing problem", 315-338. Wiley, Chichester, 1979, doi.org/10.1007/978-1-4615-5755-5_1.

[5] M. Desrochers, J.K. Lenstraa et M.W.P. Savelsbergh, "A classication scheme for vehicle routing and scheduling problems", European Journal of Operational Research, **46**(3), 322-332, 1990, doi.org/10.1016/0377-2217(90)90007-X.

[6] M. Sol et M. Savelsbergh, "A branch-and-price algorithm for the pickup and delivery problem with time windows", Memorandum COSOR 94-22, Dept. Of Mathematics and Computin, 1994.

[7] H. Housroum, "Une approche génétique pour la résolution du problème VRPTW dynamique", Thèse Doctorat en informatique, Université d'Artois, 2005.

[8] M. Chassaing, "Problèmes de transport à la demande avec prise en compte de la qualité de service", Thèse Doctorat en informatique, Université Blaise Pascal-Clermont Ferrand II, 2015.

[9] T.R. Sexton, Y.M. Choi, "Pickup and delivery of partial loads with "soft" time windows", American Journal of Mathematical and Management Sciences, **6**(3-4), 369-398, 1986, doi.org/10.1080/01966324.1999.10737484

[10] S.P. Greaves, M.A. Figliozzi, "Collecting commercial vehicle tour data with passive global positioning system technology: Issues and potential applications", Transportation Research Record, **2049**(1), 158-166, 2008, doi.org/10.3141/2049-19.

[11] M. Akil, "Problème de tournées de véhicules avec contraintes et fenêtre de temps", Mémoire Magister en informatique, UMMTO, 2013.

[12] W. Dong, K. Zhou, H. Qi, C. He, J. Zhang, "A tissue P system based evolutionary algorithm for multi-objective VRPTW", Swarm and evolutionary computation, 39, 310-322, 2018, doi.org/10.1016/j.swevo.2017.11.001.

[13] W. Zhang, D. Yang, G. Zhang, M. Gen, "Hybrid multi-objective evolutionary algorithm with fast sampling strategy-based global search and route sequence difference-based local search for VRPTW", Expert Systems with Applications, **145**, 113-151, 2020, doi.org/10.1016/j.eswa.2019.113151.

[14] M. Saint-Guillain, Y. Deville, C. Solnon, "A multistage stochastic programming approach to the dynamic and stochastic VRPTW", In International Conference on Integration of Constraint Programming, Artificial Intelligence, and Operations Research, 357-374, Springer, Cham, 2015, doi.org/10.1007/978-3-319-18008-3_25.

[15] T. Gocken, M. Yaktubay, "Comparison of different clustering algorithms via genetic algorithm for VRPTW", 2019, doi.org/10.2507/IJSIMM18(4)485.

[16] V. Pureza, R. Morabito, M. Reimann, "Vehicle routing with multiple deliverymen: Modeling and heuristic approaches for the VRPTW", European Journal of Operational Research, **218**(3), 636-647, 2012, doi.org/10.1016/j.ejor.2011.12.005.

[17] H.B. Ticha, N. Absi, D. Feillet, A. Quilliot, "Empirical analysis for the VRPTW with a multigraph representation for the road network", Computers & Operations Research, **88**, 103-116, 2017, doi.org/10.1016/j.cor.2017.06.024.

[18] W. Xu, X. Wang, Q. Guo, X. Song, R. Zhao, G. Zhao, D. He, "Gathering Strength, Gathering Storms: Knowledge Transfer via Selection for VRPTW", Mathematics, **10**(16), 28-88, 2022, doi.org/10.3390/math10162888.

[19] Y.F. El Bahi, L. Ezzine, Z. Aman., H. El Moussami, "Distribution management problem: case of vehicle routing problem with capacity constraints "CVRP" in the Moroccan petroleum sector", In 8th International Conference on Control, Decision and Information Technologies (CoDIT) (1), 1149-1154, IEEE, 2022, doi.org/10.1109/CoDIT55151.2022.9803955.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 8, No. 4, 73-80 (2023)

www.astesj.com

Special Issue on Innovation in Computing, Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

MRI Semantic Segmentation based on Optimize V-net with 2D Attention

Zobeda Hatif Naji Al-azzwi [*,1], Alexey N. Nazarov [2]

[1] *Institution Moscow of Physics and Technology, Department of School of Radio Engineering and computer Technology, 101000, Russian Federation*

[2] *Federal Research Center Computer Science and Control Russian Academy of Sciences, Moscow, Russia, 101000, Russian Federation*

ARTICLE INFO

Article history:
Received:19 April, 2023
Accepted:26 June, 2023
Online: 25 July, 2023

Keywords:
Semantic segmentation
Deep learning,
Attention-V-net

ABSTRACT

Over the past ten years, deep learning models have considerably advanced research in artificial intelligence, particularly in the segmentation of medical images. One of the key benefits of medical picture segmentation is that it allows for a more accurate analysis of anatomical data by separating only pertinent areas. Numerous studies revealed that these models could make accurate predictions and provide results that were on par with those of doctors. In this study, we investigate different methods of deep learning with medical image segmentation, like the V-net and U-net models. Improve the V-net model by adding attention in 2D with a decoder to get high performance through the training model. Using tumors of severe forms, size, and location, we downloaded the BRAST 2018 data set from Kaggle and manually segmented structural T1, T1ce, T2, and Flair MRI images. To enhance segmentation performance, we also investigated several benchmarking and preprocessing procedures. It's significant to note that our model was applied on Colab-Google for 35 epochs with a batch size of 8. In conclusion, we offer a memory-effective and effective tumor segmentation approach to aid in the precise diagnosis of oncological brain diseases. We have tested residual connections, decoder attention, and deep supervision loss in a comprehensive ablation study. Also, we looked for the U-Net encoder and decoder depth, convolutional channel count, and post-processing approach.

1. Introduction

In the past several years, convolutional neural networks (CNNs) have been used to solve issues in the disciplines of computer vision and medical image analysis. By now, deep learning-based image segmentation has gained a solid reputation as a reliable technique for picture segmentation. As the initial and crucial step in the pipeline for diagnosis and treatment, it has been routinely employed to divide homogeneous areas [1]. Today, many people experience illness, including cancer. Brain tumors can be either benign or malignant. A brain tumor is an enlargement of brain cells or cells close to the brain. Brain tissue can give rise to brain tumors. The membranes that encircle the surface of the brain, the pituitary gland, and the pineal gland are neighboring structures. An unchecked cell proliferation that threatens the central nervous system's survival is known as a brain tumor. Primary and secondary brain cancers fall into separate categories. Primary brain malignancies begin in brain cells, as opposed to secondary cancers, which spread to the brain from other organs. Primary brain tumors are gliomas. High-grade glioblastoma (HGG) and low-grade glioblastoma (LGG) are two more kinds of gliomas. The results of brain tumor segmentation can be utilized to generate quantitative measurements for glioblastoma diagnosis and treatment planning. Segmenting 3D modalities is a difficult task involving deviations and errors, in contrast to radiologists manually evaluating MRI modalities to gain quantitative information. [2]. In the medical industry, segmenting an image into its component components to extract features is a crucial task. Picture segmentation, which often comprises semantic segmentation and instance segmentation, is a significant challenge in the study of computer vision (CV).

Multimodal Brain Tumor Image Segmentation (BRATS), a yearly workshop and competition, has been held in recent years to compare several benchmark techniques created to segment the

* Corresponding Author: Zobeda Hatif Naji Al-azzwi, Iraq. Baqubqa, & zhn0hussen@gmail.com

brain tumor [3]. In clinical contexts, common standard imaging modalities include ultrasound, magnetic resonance imaging (MRI), and computed tomography (CT). Because the multiple image contrasts of these MRI protocols can be used to extract crucial supplemental information, numerous prior studies have demonstrated that multimodal MRI protocols can be used to identify brain cancers for treatment planning. T2-weighted fluid-attenuated inversion recovery (FLAIR), T1-weighted (T1), T1-weighted contrast-enhanced (T1c), and T2-weighted (T2) are among the multimodal MRI techniques. Variations in the organs' size, location, and shape make this problem more challenging. Convolutional Neural Networks (CNNs) can, however, be utilized to develop automatic segmentation techniques using MRI data [4], [5]. Deep learning algorithms produce more robust features than conventional discriminative models based on pre-defined features because they learn a hierarchy of progressively more complex task-specific features directly from the data. Deep CNNs have produced outstanding results for segmenting brain tumors [6]-[14]. In this work, we aim to segment Tumor brain MRI volumes. Due to the many appearances the tumor brain can take on in several scans because of deformations and variations in the intensity distribution, this is a difficult assignment. Moreover, due to field inhomogeneity, MRI volumes are frequently impacted by artifacts and distortions. One of the primary problems with brain lesion identification is that, because lesions only affect a small portion of the brain, unsophisticated training algorithms are more likely to choose the pointless choice of null detection. and therefore can extract the features of the three regions with greater accuracy. As a result, it has produced impressive results and made outstanding contributions to the clinical diagnosis and care of patients with brain tumors [15]. There are a lot of studies involving deep learning techniques and explaining them in details [16]-[18]. All instances of a class are found using instance segmentation, with the added ability to distinguish between distinct instances of any segment class[19]. Segmentation method entered in many medical and biological specialties [20],[21].Tumor brain segmentation interesting task to many author [22], [23]. Deep learning important technique for many applications [24].

2. Related work

The researcher In this study [5], They employed ROI masks to restrict the networks' training to relevant voxels and a cascade of two CNNs, inspired by the V-Net design, to segment brain tumors. This architecture facilitates dense training on problems with highly skewed class distributions, such as brain tumor segmentation, by limiting training to the vicinity of the tumor area. The results are based on data from the BraTS2017 Training and Validation sets. In this paper [13], For the segmentation of 3D images, they suggested a volumetric, fully convolutional neural network-based approach. After being trained from beginning to end on MRI volumes of the prostate, the network learns to predict segmentation for the full volume at once. This work introduces an objective function that optimizes during training and is based on the Dice coefficient. They deal with situations where the number of foreground and background voxels differs significantly. Histogram matching and random non-linear adjustments were utilized to augment the data and compensate for the minimal number of annotated volumes that were available for training. In this paper [14] suggested the AGSE-VNet framework for automatically segmenting MRI data from brain tumors. Using the channel relationship, the Squeeze and Excite (SE) module and Attention Guide Filter (AG) module are used to automatically enhance the useful information in the channel and suppress the useless information, and the attention mechanism is used to guide the edge detection and eliminate the influence of irrelevant information. Utilizing the online verification tool for the BraTS2020 challenge, results have been analyzed. Verification focuses on the fact that the Dice scores for the whole tumor (WT), tumor core (TC), and enhanced tumor (ET) are, respectively, 0.68, 0.85, and 0.70. Conclusion Despite the fact that MRI images have varying intensities, AGSE-VNet is unaffected by the tumor's size and therefore can extract the features of the three regions with greater accuracy. As a result, it has produced impressive results and made outstanding contributions to the clinical diagnosis and care of patients with brain tumors. In this [15] the researcher used 2D V-net K for tumor brain segmentation and enhancing predication they used data set BRATS2020.

3. Materials and Methods

3.1. Segmentation

The term "image" refers to a matrix of pixels that are ordered in rows and columns according to coordinates. There are numerous methodologies for constructing an image, including enhancement, detection, compression, segmentation, and more. As an alternative, segmentation is the act of labeling pixels.. Image segmentation is a technique used in the field of image processing that divides a digital image into several subgroups known as "image segments," which functions to simplify prospective processing or interpretation of the image by reducing the complexity of the original image. There is a distinct name for each pixel or component of a picture that belongs to the same category. Semantic segmentation, instance, and panoptic segmentation, are three prominent type of segmentation as shown in figure1.

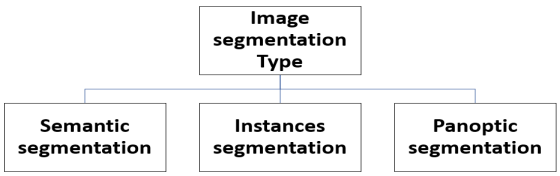


Figure 1: Show the approach of Image segmentation.

One of the most challenging problems in computer vision, particularly in medical image analysis, is semantic segmentation, the act of giving a class label to each pixel. Semantic segmentation helps with the automatic localization and detection of sick structures.[16], Semantic segmentation aims to classify different areas of a picture into categories. The segment could be viewed as a categorization issue at low resolution. Segmentation is used in computer-aided diagnosis to extract anatomical sections and has been a significant research area for information extraction[17]. When utilizing deep learning for semantic segmentation, the learning process is supervised and the training images containing actual classes that have been separated, also known as ground truth data, are connected to the labels. Semantic segmentation is used in numerous industries, including the medical sector.

When numerous instances are combined into one class, the problem description is frequently ambiguous when utilizing this approach. For example, the "people" class may be used to classify the entire crowd in a bustling scene. The semantic segmentation does not provide enough detail for this complex image.

Instance segmentation is a specific type of picture segmentation that focuses on locating instances of objects and defining their bounds. Instance segmentation has drawn interest in several computer vision disciplines, including autonomous vehicle control, drone navigation, and sports analysis. Numerous effective models have recently been created, and they can be divided into two groups: speed-focused and accuracy-focused. For this task to be used in real-time applications, accuracy and inference time are crucial.[19]. Thus, semantic segmentation differs from instance segmentation in that it treats many objects belonging to the same category as a single entity. As opposed to that, instance segmentation detects specific objects within the image[20].

Panoptic segmentation the most current segmentation task, known as panoptic segmentation, can be described as a combination of instance and semantic segmentation, where each instance of an object in the image is separated and its identification is foreseen. Popular activities like self-driving automobiles, where a significant amount of information about the immediate surroundings must be gathered with the aid of a stream of photos, find widespread application for panoptic segmentation algorithms. In figure2 [18] show the panoptic, semantic and instance segmentation.

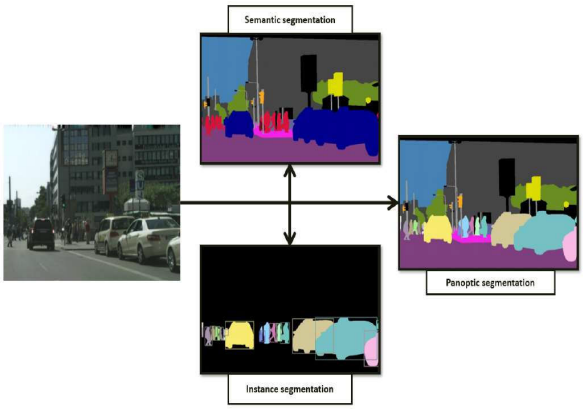


Figure 2: Demonstrate how instance segmentation, semantic segmentation, and panoptic segmentation vary from one another.

The segmentation technique belongs to the class of pixel cases. The first approach makes use of discontinuity to segment an image into lines, edges, and isolated points by identifying them based on rapid changes in local properties. The regions' boundaries are then deduced. By utilizing the homogeneity of spatially dense information, such as texture, intensity, and color, the second method's segmentation results are generated. The third strategy combines boundary-based segmentation and segmentation based on regions. [3]. Most current tumor segmentation techniques can be grouped into one of two big families based on how they segment tumors. The main drawback with image segmentation is over segmentation and high sensitivity to noise. Discriminative techniques focus on select local picture features that seem significant for the tumor segmentation job, directly learning the link between image intensities and segmentation labels without any domain knowledge. The most important picture segmentation techniques include threshold-based, edge-based, region-based, clustering-based, and "deep learning" based on artificial neural networks (3). Model training is used in machine learning-based image segmentation approaches to enhance the system's capacity to recognize key traits. For photographic segmentation tasks, deep neural network technology is highly effective.

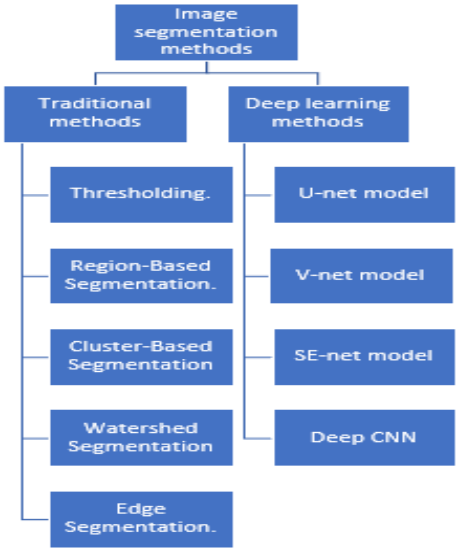


Figure 3: Show the traditional segmentation methods and deep learning methods

3.2. Deep learning methods with segmentation approach

Artificial neural networks, which are the sole foundation of the machine learning subfield of deep learning, will replicate the human brain in a manner like how neural networks do for deep learning. Because Deep Learning models can dynamically learn features from the input, they are ideal for applications like photo recognition, speech recognition, and natural language processing. The three most widely used deep learning designs are feedforward neural networks, convolutional neural networks (CNNs), and recurrent neural networks (RNN). A feedforward neural network (FNN), which has a linear information flow, is the most fundamental type of ANN. FNNs have been widely employed for tasks such as image classification, image segmentation, speech recognition, and natural language processing. Convolutional neural networks (CNNs) that were trained on manually labeled patient datasets are the foundation of most contemporary automatic segmentation methods. For image segmentation, a variety of neural network implementations and architectures are suitable. They frequently share the same essential components. An encoder is a group of layers that uses more specific, in-depth filters to extract visual characteristics. By using its earlier experience with a task that is comparable to this one (like image recognition), the encoder may be able to carry out segmentation tasks. A decoder is a collection of layers that gradually transforms the encoder's output into an interpretation according to the pixel resolution of the input image. To *enhance* model accuracy, skip

connections. Multiple long-range neural network connections that let the model recognize *significant information* at many scales are used. There are many models used to segment images; the most famous and important are U-net models[21] segment for biomedical image , V-net models [13] segment for volume medical image in follow section more description for each one.

a. V-net model:

The [13] proposed On both the left and right sides of the V-Net, there are many layers. The network's right and left halves alternately decompress and compress the signal until it is back to its original size. The network's left side is divided into several stages, each of which can operate in a different configuration. Each phase has between one and three convolutional layers. Each level involves the learning of a residual function. Each step's input is applied to its respective convolutional layers, subjected to non-linear processing, and added to the output of the final convolutional layer to enable the learning of a residual function. Comparison of this design to non-residual learning networks like U-Net ensures convergence.

The volumetric kernels used in each stage's convolutions have a 5x5x5 voxel size. (A voxel in 3D space corresponds to a value on a standard grid. Like how "voxelization" is used in point clouds, the word "voxel" is widely employed in 3D spaces.) The resolution of the compression path is decreased by stride 2 convolution using 2x2x2 voxel wide kernels. As a result, the objective of layer pooling is still accomplished while halving the size of the generated feature maps. The number of feature channels doubles at each stop along the V-compression Net's journey. Convolutional operations have a smaller memory footprint during training because they do not need switches to change the output of pooling layers back to their inputs for back-propagation. sampling backwards. They leave a lower memory footprint when training. It is possible to expand the receptive field by downsampling. A nonlinear activation function is employed with PReLU.

The network extracts feature from the lower resolution feature maps, enhances the spatial support for those features, and collects the required data to construct a two-channel volumetric segmentation. In each step, the inputs are deconvolved to increase their size, and then one to three convolutional layers are added, each of which uses half as many 5x5x5 kernels as the layer before it. Residual function is taught, just like in the left part of the network.

The output size of the final convolutional layer, which has a 1x1x1-kernel size, is the same as the input volume. The two feature mappings are generated. With the use of soft-max voxel-wise segmentation, these two output feature maps provide probabilistic foreground and background segmentations. lateral connections Similar to U-Net, compression results in the loss of location data (left). As a result, horizontal connections are used to transfer the features that were acquired from the left part of the CNN's first stages to the right part. By doing so, better precision of the final contour forecast and the correct part position data have been achieved. The model's convergence is accelerated by these linkages. Figure 4 shows the structure of V-net.

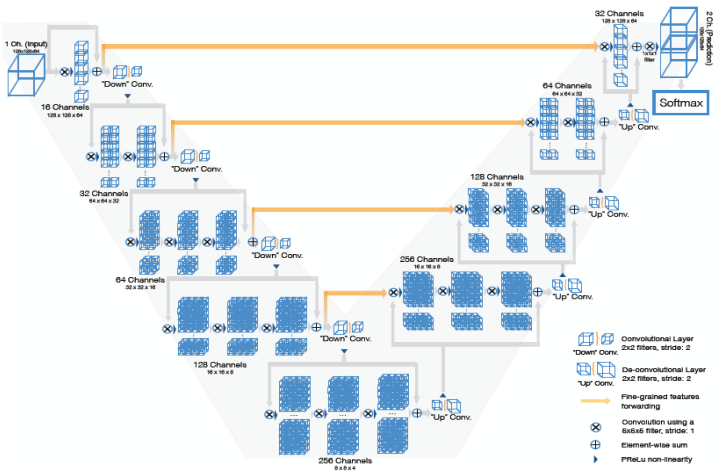


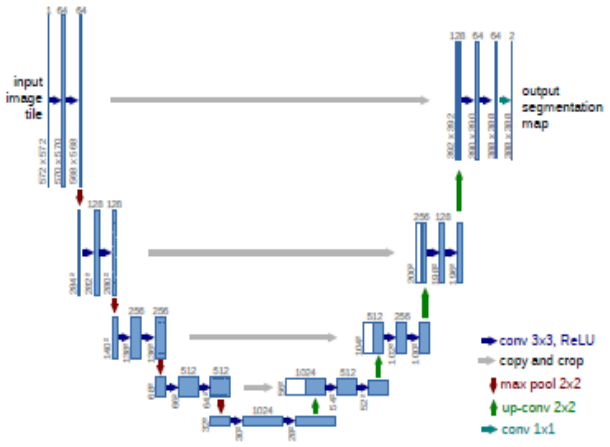
Figure 4: Show the structure of V-net [13]

b. U-net model:

The U-net model is a significant model that includes biomedical imagery. [21], take U shape as shown in figure 4. [21] it's have three part encoder, decoder and bottleneck. For classification-related issues, the tool that turns an image into a vector that can be used for classification is useful. However, in order to segment a picture, the feature map must also be turned into a vector, which must then be used to recreate the original image. This is a hard operation because it is far more difficult to convert a vector into a picture than the contrary, this is the main issue with U net-design. The three components of the U-net model are contraction, bottleneck, and expansion. In the U-Net model, the contraction is employed to expand a vector into a segmented image. This would preserve the structural integrity of the image and significantly lessen distortion.

The architecture can successfully learn the complex structures since there are twice as many kernels or feature maps after each block. A mediator between the contraction layer and the expansion layer is the bottom layer. After two 3X3 CNN layers, a 2X2 up convolution layer is applied.

The number of blocks for both expansion and contraction is the same. Yet, the extension component is where this architecture's core may be located. Several expansion building elements are included in addition to the contraction layer. A 2X2 up sampling layer is added to each block after the first two 3X3 CNN layers to handle the input. To maintain symmetry, the convolutional layer uses half as many feature mappings between each block. But each time, the input is also supplemented by feature maps from the associated contraction layer. As a result, the traits identified during the image's contraction would be considered during its reconstruction. The subsequent 3X3 CNN layer receives the output mapping, with the number of feature mappings matching the necessary number of segments. This model, U-net, has been used with the 2018 Brast data set in this paper to evaluate the optimization of V-net, as shown in Figure 6.



The figure 5: Show the U-net model

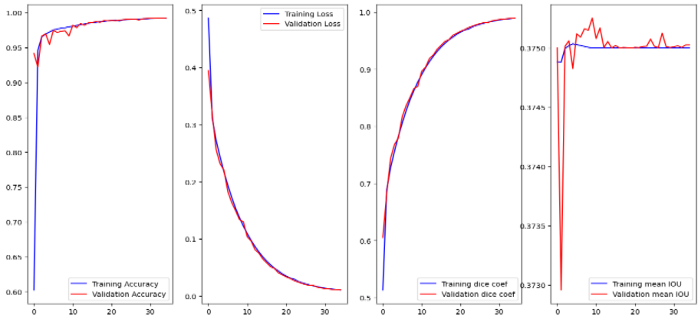


Figure 6: Show the training of U-net model with 2018Brast.

3.3. Attention 2D-layer.

Deep learning approaches such as attention models or attention mechanisms are used to add additional focus to a particular component. Focusing on something special and noting its unique value is what deep learning is all about. The attention mechanism was first made available by[24]. To address the bottleneck problem brought on by a fixed length encoding vector, where the decoder's access to the input's data would be severely limited. Because the representation's dimensionality would be limited to match that of shorter or simpler sequences, this is seen to be especially troublesome for long and/or complex sequences.

The attention layer's goal is to increase the machine translation encoder-decoder model's effectiveness. By integrating all the encoded input vectors in a weighted fashion, with the most pertinent vectors receiving the highest weights, the attention mechanism was developed to enable the decoder to use the most pertinent sections of the input sequence in a flexible manner.

As indicated in the findings section, the attention layer and V-net model were integrated in this study to increase training effectiveness. As seen in figure 7, the attention layer used in this study is composed of a convolution layer followed by Relu activation, convolution, followed by a sigmoid function.

Boost helpful functions and disable less helpful ones for the present work. The image is decompressed after each decoder receives the down sampling characteristics for the corresponding step. In the process of downsampling, the attention block is used to remove background noise and unnecessary details, while the image filter (edge information) directs the structural and feature data for the images. Remember that the model uses the skip connection idea to prevent gradient vanishing.

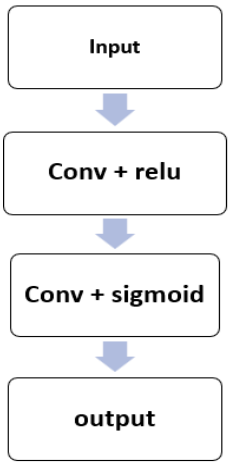


Figure 7: Show the structure of attention layer.

4. Proposed Method

Encoder and decoder components are symmetrical in the network's structure. The encoder is made up of four blocks, each of which contains several convolution layers, batch normalization, and PRelue. Between each block, there is a Max pooling layer with (2,2), which allows for down sampling and equivalent up sampling. All convolution layers have size 5x5with stride [1,1]. A full skip connection is joined with concatenation to keep the feature map and resolution of the image. Using skip connections following sampling at the decoding end, the lower-level features are fused with the first four scales of the feature map of the V-net.

Four blocks make up the decoder side, including an attention layer with a kernel size of (1,1) and a stride of 1. The layer's first three operations are convolution, relu activation, and sigmoid. Encoding equipment Batch normalization facilitates weight initialization and speeds up network training by enabling significantly higher learning rates. While building deeper networks, initializing weights might be challenging. The addition of a 2D attention layer can enhance this model. Many researcher worked on V- net model and optimized it [5,14]. The optimize in this research was by adding attention layer with encoder. There are full skip connections between encoder and decoder.

This model V-net makes use of the Brast 2018 data set to produce feature maps with modest values but concentrated, high-dimensional semantic interpretations after continuous convolution. To provide a segmentation result, the decoder is then repeatedly convolved and up sampled to the original size.

By batch normalizing, the sensitivity to the original beginning weights is reduced. To utilize a suitable stride at the conclusion of each block to reduce its resolution. The network's left side compresses the signal while the right side decompresses it until the signal's original size is reached.[13].

4.1. Evaluation criteria

To Evaluate this work, we need take in our consideration many criteria. We gauge the training of our model using the loss,

accuracy, mean IOU, dice coefficient, precision, sensitivity, and specificity.

a. Loss function: A loss function compares the target and anticipated output values to determine how well the neural network mimics the training data. During training, we make an effort to minimize this difference between the expected and desired outputs. determined by.

$$loss = \frac{1}{n}\sum_{i=1}^{n}(y - y')^2 \quad (1)$$

b. Accuracy: Accuracy is a measure of the model's performance across all classes. It is useful when all classes are given equal weight. It is calculated as the ratio of events that were successfully anticipated to all events that were predicted. One of the measures that is most frequently used when performing categorization is accuracy.. We explain each term in table 1.

$$Acc = \frac{T_p+T_N}{T_p+T_N+F_P+F_N} \quad (2)$$

Tabel.1: Describe the criteria for validity.

T_P	A true positive result is one that the model correctly identified as falling within the positive category.
T_N	When the model correctly predicts the negative class, the result is said to be True Negative.
F_P	False Negative: is a finding when the model forecasts the positive class wrongly.
F_N	False Negative: is a outcome genuine positive instances are labeled as negative

c. Mean IOU: mean intersection over union is used when calculating mean average precision.

$$mean_{IOU} = \frac{T_P}{T_P+F_P+F_N}, \quad (3)$$

d. Dice coefficient: A predicted segmentation's pixel-by-pixel agreement with the associated ground truth can be compared using the Dice coefficient. In image segmentation, the Dice score coefficient is a commonly used metric of overlap. The benefit of dice loss is that it does a good job of handling the class mismatch between the foreground and background pixel counts.

$$Dic = \frac{2|A\cap B|}{|A|+|B|} \quad (4)$$

e. Precision: One indicator of the model's performance is precision, or the quality of a successful prediction the model makes. By dividing the total number of positive predictions by the percentage of true positives, precision is determined (i.e., the number of true positives plus the number of false positives).

$$p = \frac{T_P}{T_P+F_P} \quad (5)$$

f. Sensitivity: One indicator of the model's performance is precision, or the quality of a successful prediction the model makes. By dividing the total number of positive predictions by the percentage of true positives, precision is determined.

$$sen = \frac{T_P}{T_P+F_N} \quad (6)$$

g. Specificity: The fraction of true negatives that the model correctly predicts is known as specificity. Specificity can be defined as the ability of the algorithm or model to predict a genuine negative of each possible category. It is sometimes referred to as the genuine negative rate in literature.

$$Sp = \frac{T_N}{T_N+F_P} \quad (7)$$

4.2. Descriptive for Data Set

The three sources for the BraTS 2018 challenge are indicated, respectively, by the designations "2013," "CBICA," and "TCIA." The training data set contains 20 cases of high-grade glioma (HGG) from the 2013 group, 88 cases from CBIC, and 102 cases from TCIA, as well as 10 cases from 2013 and 65 cases from TCIA participants who had low-grade gliomas (LGG), which are less active and infiltrative. [22] .One of dataset that is frequently used in the world of healthcare is BRATS 2018 which used in this paper include tow folder HHG and LLG each one has 2oo case patients we used 55 cases in this paper.

Per case, BRATS provides four multimodal MRI modalities. The four MRI modalities that the BRATS competition organizers provide, along with associated manual segmentation, include 3D brain MRIs with ground truth brain tumor segmentations annotated by physicians (T1, T1c, T2, and FLAIR): T1-weighted (T1), T1-weighted post-contrast (T1c), T2-weighted (T2), and fluid-attenuated inversion recovery (FLAIR). [6]. The link of data set; which used in this paper BRAST2018 can find on this https://www.kaggle.com/datasets/sanglequang/brats2018/download?datasetVersionNumber=10.

The volume dimension of every MRI in the BRATS 2018 dataset is 240 x240 x155. The data set which used in this study MRI 3D image (voxel) tumor brain which represented by X, Y, Z, 3D, non-informative voxels are found in medical images in general and brain MRIs. The preprocessing with this data started by converted the voxel to pixel it's mean 3D to 2D image to reduce the consuming time. The HGG have be used in this article have four modalities (T1, T1c, T2, and FLAIR). The data set in this paper divided to training and test, the test split to test and validation. The size of training is 2970 with dimension (192X192X4), and the size of test and validation is 990 with dimension (192X192X4). It is difficult to discern between a tumor and normal tissue despite the fact that tumor margins are typically fuzzy and that there is a considerable deal of variation in shape, location, and expression among patients. Despite recent developments in automated algorithms, segmenting brain tumors in multimodal MRI scans remains a difficult task in medical image processing [5]. The biggest obstacle to segmenting brain tumors is the class imbalance. Even less of the brain's tissue between 5 and 15% represents each tumor site [23].

5. Result and discussion

We improved and suggested V-Net-attention 2D based fully convolutional networks to address the brain tumor segmentation challenge. In essence, semantic segmentation is the task that deals with tumor recognition and segmentation. The data set of BRAST2018 divided to training, validation and test set with 35 epochs and the size of batch 8, because high memory consumption is still challenge in 3D-CNNs. The developed model was

implemented on colab and evaluated on brain magnetic resonance images MRI. The findings demonstrate that this proposed technique outperforms several of the already popular derivative models based on CNNs in terms of segmentation performance and tumor recognition accuracy. The experiment started by pre-processing to data set, input slice of the BRAST dataset which feed to model have dimension 192x192x4 (h x w x d, refereed to Hight, width and depth respectively) in this paper. In comparison to earlier deep learning-based studies on this subject, we train U-net on the same BRAST2018 data set as indicated in the preceding section. The results reveal that training improves V-net by adding an attention layer, which has a decent performance on HGG as show in figure8. Take optimizers Adam with learning rate (lr=1e-5).

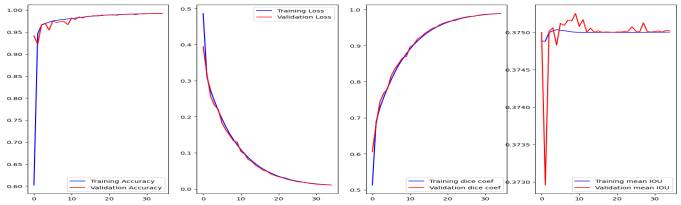


Figure 8: Show the training of V-net attention 2D on the HGG.

The measurement of classification accuracy as well as localization correctness is thought to be required in the analysis of semantic segmentation, which can be complex. The objective is to evaluate how closely the predicted (prediction) and annotated segmentation differ (ground truth). We used different benchmark to evaluate the optimized model V-net Attention 2D where take in consideration loss, Accuracy, Mean_IOU, dice_coef, precision, sensitivity, specificity demonstrated in table2, Instead, it is a helper metric that assesses the level of agreement between the ground truth and prediction.

Table 2: Show the result through training of model V-net attention 2D.

	loss	accuracy	MeanIOU	Dice_coef	precision	sensitivity	specificity
Train result	0.0105	0.99203	0.37501	0.9894	0.9926	0.99156	0.99754
Validation result	0.0108	0.99174	0.37502	0.9891	0.9923	0.99128	0.99744
Test result	0.0107	0.99175	0.37501	0.9892	0.9924	0.99124	0.99747

Here, we improve a strategy for segmenting medical pictures using fully convolutional neural networks that have been end-to-end trained to locate tumors in MRI scans. In contrast to previous recent techniques, Convolutions and BRAST 2018's slice-wise input volume processing are recommended. One of the primary issues with tumor brain is that since these lesions only impact a small portion of the brain, simple training methods are predisposed to the useless option of null detection. Only brain tissue will be used to train the modified V-NET first network model to produce raw data and predict the normal region, as shown in figure 9.

6. Conclusion

In the study developed a deep convolution network-based, fully automatic method for locating and classifying brain cancers in this study. Through tests on a well-known benchmarking dataset (BRATS 2018), which covers both HGG and LGG patients, we have demonstrated that our method can deliver both efficient and robust segmentation relative to manually determined ground truth. Our V-Net 2D attention-based deep convolution networks may also yield outcomes that are comparable for the overall tumor region and even better for the core tumor region when compared to existing state-of-the-art methods. The presented method can be used to generate a patient-specific brain tumor segmentation model without the requirement for manual assistance, which might enable objective lesion assessment for clinical tasks like diagnosis, treatment planning, and patient monitoring.

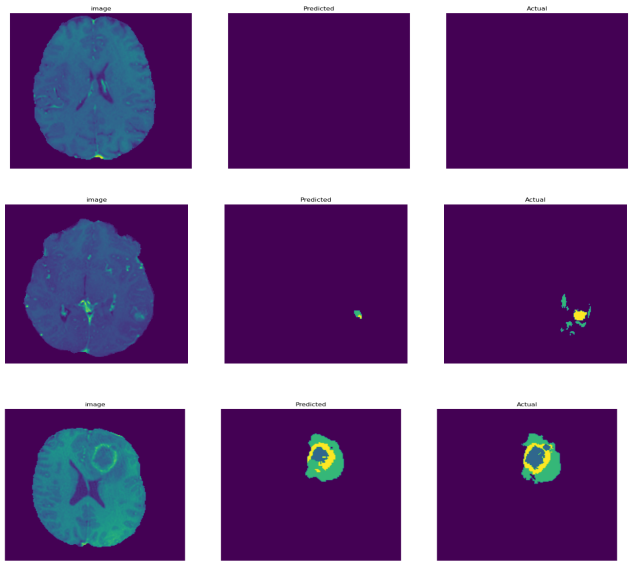


Figure 9: Illustration of the V-net-Attention 2D for normal cases HGG segmentation results in BraTS2018.

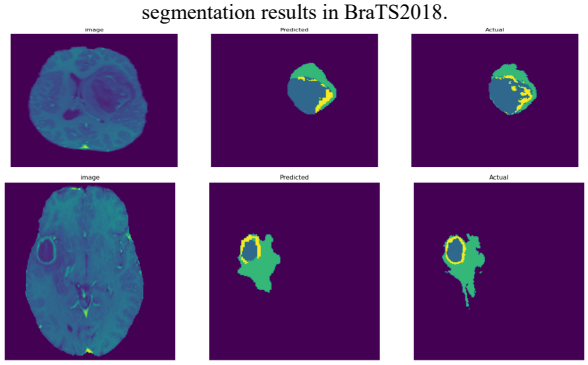


Figure 10: Demonstration of the V-net-Attention 2D segmentation findings for cases of brain tumors and brain tissue from BraTS2018.

Acknowledgment

The accomplishment of this study was supported by a grant from the institution Moscow Institute of Physics and Technology.

References

[1] M.H. Hesamian, W. Jia, X. He, P. Kennedy, "Deep Learning Techniques for Medical Image Segmentation: Achievements and Challenges," Journal of Digital Imaging, **32**(4), 582–596, 2019, doi:10.1007/s10278-019-00227-x.

[2] P. Ahmad, S. Qamar, L. Shen, A. Saeed, "Context Aware 3D UNet for Brain Tumor Segmentation," Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), **12658 LNCS**, 207–218, 2021, doi:10.1007/978-3-030-72084-1_19.

[3] B.H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest, L. Lanczi, E. Gerstner, M.A. Weber, T. Arbel, B.B. Avants, N. Ayache, P. Buendia, D.L. Collins, N. Cordier, J.J. Corso, A. Criminisi, T. Das, H. Delingette, Ç. Demiralp, C.R.

Durst, M. Dojat, S. Doyle, J. Festa, F. Forbes, et al., "The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS)," IEEE Transactions on Medical Imaging, **34**(10), 1993–2024, 2015, doi:10.1109/TMI.2014.2377694.

[4] H.T. Le, H.T. Thu Pham, "Brain tumour segmentation using U-Net based fully convolutional networks and extremely randomized trees," Vietnam Journal of Science, Technology and Engineering, **60**(3), 19–25, 2018, doi:10.31276/vjste.60(3).19.

[5] A. Casamitjana, M. Catà, I. Sánchez, M. Combalia, V. Vilaplana, "Cascaded V-Net using ROI masks for brain tumor segmentation," Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), **10670 LNCS**, 381–391, 2018, doi:10.1007/978-3-319-75238-9_33.

[6] M. Ghaffari, A. Sowmya, R. Oliver, "Automated Brain Tumour Segmentation Using Cascaded 3D Densely-Connected U-Net," Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), **12658 LNCS**, 481–491, 2021, doi:10.1007/978-3-030-72084-1_43.

[7] H. Dong, G. Yang, F. Liu, Y. Mo, Y. Guo, "Automatic brain tumor detection and segmentation using U-net based fully convolutional networks," Communications in Computer and Information Science, **723**, 506–517, 2017, doi:10.1007/978-3-319-60964-5_44.

[8] Q. Jia, H. Shu, "BiTr-Unet: A CNN-Transformer Combined Network for MRI Brain Tumor Segmentation," Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), **12963 LNCS**, 3–14, 2022, doi:10.1007/978-3-031-09002-8_1.

[9] A. Atiyah, K. Ali, "Brain MRI Images Segmentation Based on U-Net Architecture," Iraqi Journal for Electrical and Electronic Engineering, **18**(1), 21–27, 2022, doi:10.37917/ijeee.18.1.3.

[10] V.P. Gladis Pushpa Rathi, S. Palani, "Brain tumor detection and classification using deep learning classifier on MRI images," Research Journal of Applied Sciences, Engineering and Technology, **10**(2), 177–187, 2015.

[11] M.A. Al Nasim, A. Al Munem, M. Islam, M.A.H. Palash, M.M.A. Haque, F.M. Shah, "Brain Tumor Segmentation using Enhanced U-Net Model with Empirical Analysis," 2022.

[12] K. Munir, F. Frezza, A. Rizzi, "Deep Learning Hybrid Techniques for Brain Tumor Segmentation," Sensors, **22**(21), 2022, doi:10.3390/s22218201.

[13] F. Milletari, N. Navab, S.A. Ahmadi, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," Proceedings - 2016 4th International Conference on 3D Vision, 3DV 2016, 565–571, 2016, doi:10.1109/3DV.2016.79.

[14] X. Guan, G. Yang, J. Ye, W. Yang, X. Xu, W. Jiang, X. Lai, "3D AGSE-VNet: an automatic brain tumor MRI data segmentation framework," BMC Medical Imaging, **22**(1), 2022, doi:10.1186/s12880-021-00728-8.

[15] D. Rastogi, P. Johri, V. Tiwari, "Brain Tumor Segmentation and Tumor Prediction Using 2D-VNet Deep Learning Architecture," Proceedings of the 2021 10th International Conference on System Modeling and Advancement in Research Trends, SMART 2021, (August 2022), 723–732, 2021, doi:10.1109/SMART52563.2021.9676317.

[16] Y. Azzi, A. Moussaoui, M.-T. Kechadi, "Semantic Segmentation of Medical Images with Deep Learning: Overview," Medical Technologies Journal, **4**(3), 568–575, 2020, doi:10.26415/2572-004x-vol4iss3p568-575.

[17] J. Moorthy, U.D. Gandhi, "A Survey on Medical Image Segmentation Based on Deep Learning Techniques," Big Data and Cognitive Computing, **6**(4), 2022, doi:10.3390/bdcc6040117.

[18] O. Elharrouss, S. Al-Maadeed, N. Subramanian, N. Ottakath, N. Almaadeed, Y. Himeur, "Panoptic Segmentation: A Review," 2021.

[19] S. Jung, H. Heo, S. Park, S.U. Jung, K. Lee, "Benchmarking Deep Learning Models for Instance Segmentation," Applied Sciences (Switzerland), **12**(17), 1–25, 2022, doi:10.3390/app12178856.

[20] T. Scherr, A. Bartschat, M. Reischl, J. Stegmaier, R. Mikut, "Best Practices in Deep Learning-Based Segmentation of Microscopy Images," Proceedings - 28. Workshop Computational Intelligence, Dortmund, 29. - 30. November 2018. Ed.: F. Hoffmann, 175, 2018.

[21] T.B. Olaf Ronneberger, Philips Fischer, "U-Net: Convolutional Networks for Biomedical Image Segmentation," IEEE Access, **9**, 16591–16603, 2015, doi:10.1109/ACCESS.2021.3053408.

[22] L. Dai, T. Li, H. Shu, L. Zhong, H. Shen, H. Zhu, "Automatic Brain Tumor Segmentation with Domain Adaptation," 1–12, 2018.

[23] L.M. Ballestar, V. Vilaplana, "MRI Brain Tumor Segmentation and Uncertainty Estimation Using 3D-UNet Architectures," Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), **12658 LNCS**(1), 376–390, 2021, doi:10.1007/978-3-030-72084-1_34.

[24] D. Bahdanau, K.H. Cho, Y. Bengio, "Neural machine translation by jointly learning to align and translate," 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings, 1–15, 2015.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 8, No. 4, 91-100 (2023)

www.astesj.com

ASTESJ

ISSN: 2415-6698

FPGA Implementation of 5G NR LDPC Codes

Sahar Fekry Abdel-Momen[*,1], Abdel Halim Abdelnaby Zekry [1], Ashraf Yehia Hassan [2], Wageeda Ibrahim Shaban [3], Mustafa Mohammed Shiple [4]

[1]*Faculty of Engineering, Ain Shams University, Department of Electrical Engineering, Cairo, 11511, Egypt*

[2]*Facultyof Engineering, Benha University, Department of Electrical Engineering, Benha, 13512, Egypt*

[3]*Facultyof Engineering, Benha University, Departmentof Basic Sciences, Benha ,13512, , Egypt*

[4]*Facultyof Engineering, Al Azhar University, Department of Electrical Engineering, NTI, 11371, Giza, Egypt*

ARTICLE INFO

Article history:
Received:09 April, 2023
Accepted:01 July, 2023
Online: 25 July, 2023

Keywords:
5G Fifth generation
NR New Radio
LDPC Low-Density Parity-Check

ABSTRACT

As a result of rising expectations for quality, the employment of advanced technical requirements for future fifth-generation (5G) new radio is required. The error-correction coding method is one of the most important components of a new generation. The 5G NR New Radio Low-Density Parity Check (LDPC) codes, which have been adopted by the 5G standard, are a standout solution in terms of high coding gain, high throughput, and low power dissipation. This paper presents an implementation of 5th generation (5G) New Radio (NR) and 5G NR low-density parity check codes, which are performed with the aid of a proper architecture. LABVIEW will be used in wireless communications to reduce the cost, space, and power. Simultaneously, this increased the speed. The circuit design supports a constraint length of 1360 and a code rate of 0.5. The LDPC encoder and decoder are implemented on an NI MY RIO 1900 ZYNQ FPGA at a 33 MHZ core frequency starter kit. Xilinx Vivado 18.2 series was used for the simulation. The implemented design shows an area overhead reduction of 50% compared with the referenced designs of the Xilinx 7 series device. In MY RIO ZYNQ, the proposed method achieved 21000 LUTs compared with Xilinx 7-series solutions, and it has a much higher throughput (224 vs. 87 and 5 MBit/s), followed by MY RIO ZYNQ, which is better than previous state-of-the-art solutions in terms of area and higher data rates. Moreover, the implemented 5G NR LDPC decoder tested against an additive white Gaussian noise channel (AWGN) and consequently has gained more popularity in many applications.

1. Introduction

It is vital for wireless communication systems to control data corruption by offering an effective solution to errors that occur during the communication process [1]. To satisfy the needs and requirements of users about reducing errors that were created by the channel, for example, noise, fading, etc. Channel coding techniques are implemented at the receiver to correct a significant number of communication defects that influence sent data symbols [2]. The basic purpose of channel coding is to make transmitted symbols more distant so that they are harder to destroy and more dependable when they reach their destination, thereby obtaining a practically error-free signal. Many techniques and strategies are available for the transmission and reception of digital data through noisy channels, but Low-Density Parity-Check (LDPC) is the most reliable.

It offers enhanced coding gain and performance [2]. The Low-Density Parity-Check (LDPC) decoder is an efficient technique for decoding code. However, Low-Density Parity-Check (LDPC) requires an exponential increase in hardware complexity to obtain an improved decoding accuracy [3]. The previously mentioned codes compete to reserve a position inside the control and data channels within the forthcoming 5G NR standard. They all seem to have the same opportunity to win a race. The decision to choose one of these codes must be based on a rigorous evaluation to determine whether it fits the 5G criteria. This requires the physical implementation of these codes. One of the easiest techniques for completing this task is the use of an FPGA device [4]. Adopting this technique shortens the time required for prototype and processing. Because of the stochastic

[*]Corresponding Author: Sahar Fekry Abdel-Momen, Toukh , 13741, Egypt
Email: eng.saharfekry23@gmail.com

nature of the Bit Error Rate (BER) performance test, it may be take a day on a computer, while on an FPGA, it can take hours. This mostly results from the FPGA's ease of usage and parallel processing capabilities.

1.1. Problem definition

According to previous research, various implementation strategies for LDPC codes can be found in the literature. Researchers are working to make LDPC codes less complex, more power efficient, and faster. They also employ (μc) microcontrollers to implement 5G NR LDPC codes on slower platforms. The use of field-programmable gate array (FPGA)-based technology has been suggested to overcome the performance problem of LDPC codes. The fundamental benefit of FPGAs over conventional digital signal processing plates is their capacity to take advantage of parallelism or the reproduction of hardware operations that occur simultaneously in several areas of the chip. The ability to tradeoff between space and speed until late in the design cycle is another benefit of FPGAs [5].

1.2. The suggested way to the study

This study improves the implementation of a soft-decision 5G NR LDPC decoder on a MY RIO ZYNQ FPGA to address issues with complex implementation, area, and speed. The design and implementation of a 5G NR LDPC encoder with a constraint length of 1360 and a rate of 1/2 (a starter kit) have also been discussed. In the starter set, as previously described. The design was coded in VERLILOG and LABVIEW. The simulation will use the Xilinx Vivado 18.2 series

1.3. history of LDPC

The LDPC code is a forward error correction code invented by Robert Gallager in his doctoral dissertation at MIT in the 1960s. Although LDPC code is one of the most practical realizations of Shannon's theory, with high computational complexity for forwarding error correction and highly structured algebraic block codes and convolutional codes, LDPC codes have long been neglected. Approximately 30 years after the invention of the LDPC code, MacKay and Neal proved that the LDPC code can approach the Shannon limit under the condition of combining iterative decoding based on belief propagation [5]. LDPC codes were built using Tanner graph construction. LDPC allows highly efficient data transfer. 5G new radio has been the control of this application, from a commercial field to a military field. The high rate of data transmission is the main advantage of the 5G new radio. But on the other hand, extremely high transmitting data could cause many errors. These errors may damage the data; therefore, we could not obtain correct information [6]. One of the greatest advantages of using LDPC codes is their ability to achieve a near-Shannon limit performance with low decoding complexity. This means that LDPC codes can provide reliable communication even at low signal-to-noise ratios (SNRs) [6]. LDPC decoding is a computationally intensive process that involves iteratively updating the probabilities of transmitted bits based on the received signal. The decoding process can be performed using various algorithms such as belief propagation, message passing, and min-sum.

1.4. Overview 5GNRLDPC Codes

The 5G New Radio (NR) LDPC is a powerful error correction technique used in 5G wireless communication systems. This technology was designed to improve the reliability and efficiency of data transmission in 5G networks, which are expected to be faster and more reliable than their predecessors. Low-Density Parity-Check (LDPC) codes have emerged as a popular choice for error correction in 5G New Radio (NR) coding. LDPC codes are a class of linear error-correcting codes capable of achieving near-Shannon limit performance with low decoding complexity. In this paper, we explore the key features of 5G NR LDPC coding and its benefits for next-generation wireless communication systems [6].

In 5G NR, LDPC decoding was performed using a belief propagation algorithm with layered decoding. Layered decoding involves dividing the code into multiple layers, and decoding each layer separately. This reduces the overall complexity of the decoding process

1.5. Advantages Of LDPC Codes

A key advantage of LDPC codes is their ability to correct a large number of errors with high accuracy [7]. This makes them ideal for use in 5G networks, which require high-speed data transfers with minimal errors. Another advantage of the LDPC code is its flexibility. These codes can be adapted to different types of data transmission, making them suitable for use in a wide range of applications. The technical advantages of 5G new radio can be included in the following points:

- High Data Rates: LDPC codes provide high data rates, which is essential for 5G new radio systems that require high-speed data transmission.
- Low Latency: LDPC codes are double-edged swords; however, in most cases, they have low decoding latency, which is crucial for real-time applications such as autonomous vehicles and remote surgery [8].
- Robustness: LDPC codes are highly robust against channel impairments such as noise and interference, making them ideal for use in 5G new radio systems that operate in challenging environments.
- Low-complexity: LDPC codes investigate the need for 5G implementation [9]. LDPC codes have low encoding and decoding complexity, which reduces the computational requirements of the system and enables efficient implementation of hardware.
- Flexibility: LDPC codes can be designed to meet specific performance requirements, making them highly flexible and adaptable to different use cases in 5G new radio systems [10].
- Standardization: LDPC codes are part of the 5G new radio standard, ensuring interoperability between different vendors' equipment and facilitating the deployment of 5G networks worldwide.

In addition to their technical advantages, LDPC codes are cost-effective. They require less processing power than other error-correction techniques, making them an attractive option for manufacturers looking to reduce costs while maintaining high levels of performance. Field-programmable gate arrays (FPGA)

have become more widespread in recent years, particularly in high-performance embedded applications for digital integrated circuits. Designers' ability to use FPGAs to implement complicated hardware in the field is one of the most important features. Unlike other large-scale integrated circuits that are bound to a specific permanent hardware function, FPGAs can be programmed after they have been built [9].

2. Literature Review

In [10] the author introduced a paper which entitled "5G New Radio Prototype Implementation Based on SDR," where their build an open-source SDR-based platform to realize the transceiver of the physical downlink shared channel (PDSCH) of 5G NR according to Third Generation Partnership Project (3GPP) standard. They provided a prototype for pairing two 5G users using the NOMA technique. In addition, a suitable design for LDPC channel coding was performed. The intermediate stages of segmentation, rate matching, and interleaving were also carried out to realize a standard NR frame. They implemented a LabVIEW code word level for all communication modules of NR starting from CRC, segmentation, and LDPC channel coding and ending with the NOMA and OFMA stages in their experiments. They carried out both simulation and real-time scenarios on the designed 5GNR for the purpose of system performance evaluation and to demonstrate its potential in meeting future 5G mobile network challenges. In [11] the authors published a paper in 2020 with the title "Implementation of Decoder Using LDPC Codes on FPGA". This can improve the speed-decoding technique. 5G technology increases the throughput and latency of LDPC codes and polar codes. In their study, they introduced LDPC codes using the Min Sum (MS) algorithm. the authors built the embedded device successfully and evaluated the performance of the Turbo and LDPC Codes on an AWGN channel with BPSK modulation via Xilinx ISE 14.7 software. Using Xilinx ISE simulation on the target Virtex- 5 FPGA for device XC5Vlx20t-2-ff32, the LDPC encoder and decoder support 375.0 MHz and 390.0 MHz frequency, respectively, to verify the functional simulation and data connection with the intended hardware. In contrast to turbo codes, LDPC codes enable quick switching in FPGA hardware. In addition, LDPC codes deliver superior timing-related parameter outcomes compared to turbo codes. In contrast to turbo codes, which have hardware encoder and decoder latencies of 7.767 ns and 8.64 ns, respectively, the LDPC encoder and decoder have combined path latencies of 6.513 ns and 6.302 ns, respectively. According to the analysis, LDPC codes offer the best solutions for hardware complexity, timing response parameters, and high-frequency support. Given that LDPC codes work very well in robot communication with FPGA hardware, they are the ideal, most effective, and most reliable option for 4G and 5G wireless communication, as well as digital broadcasting. 1n [12] the article, titled Design and Implementation of Optimized LDPC for SDR Applications, was published ,They attempt to enhance the capacity performance and low-power consumption and minimize the decoding complexity. The SDR-based wireless communication technology employs Low-Density Parity-Check codes. With the suggested design, the hardware complexity of telecommunication applications was reduced. Utilizing the Xilinx 12.4 ISE tool and Verilog language simulation, an improved LDPC encoder and decoder were constructed. LDPC encoder and decoder, where the suggested architecture has a latency of just 7.2ns compared to the standard method's 9.1ns. Similar to how the traditional design uses 1.07 mw of electricity and the suggested design uses 0.962 mw, Their investigation demonstrates that the new system uses less latency and power than the standard design. Their improved LDPC encoding and decoding with less latency and power consumption without compromising the performance of the standard design. The Xilinx ISE tool was used to implement and synthesize this technique. Utilizing the Xilinx platform, the proposed system was constructed and tested based on the results .In [13] the Author also introduced a prototype of the LDPC decoder in paper "FPGA Based Design and Prototyping of Efficient 5G QC-LDPC Channel Decoding". In his study, he introduced an effective architecture that consumes 92664 LUTs, 49049 registers, and about 248 BRAMs. The target device was xc7k160t . the author also gives another architecture in his paper "Parallel and Flexible 5G LDPC Decoder Architecture Targeting FPGA". In the code rate of 0.324, the average throughput uniform was 427. This result was noted with less resource consumption.

3. Implementation of 5G-NR LDPC System Using FPGA Chip on the Embedded Computer NI-MyRIO 1900

3.1. The FPGA Card Used

An FPGA is an integrated circuit, the hardware configuration of which can be altered to perform a particular function. The user application can program an FPGA. This operation is performed "sequentially" by the user according to their objectives. The FPGA had fixed logic cells. The user determines the functions of these logic cells and those of the connections [13]. We used an NI MyRIO-1900 with a Xilinx FPGA board for the FPGA project (Figure 1). Control, robotics, and mechatronic systems may incorporate programmable portable NI MyRIO 1900. An analogue input (AI), analogue output (AO), digital input and output (DIO), power output, audio input, and output channel-equipped NI MyRio-1900 device can be remotely or via a USB connection to a computer. The analogue input channels of the system can function between 0 and 5 V and between 2 and 10 V (12 bit, 500 kS/s). Two analogue output channels with a combined resolution of 12 bits at 345 kS/s and an output range of 0–5 V to 4–10 V [14]. Digital inputs and outputs were available for a total of 40 channels. The Xilinx Z-7010-powered device has terminals that generate +3.3 V, +5.0 V, +15.0 V, and -15.0 V of voltage.

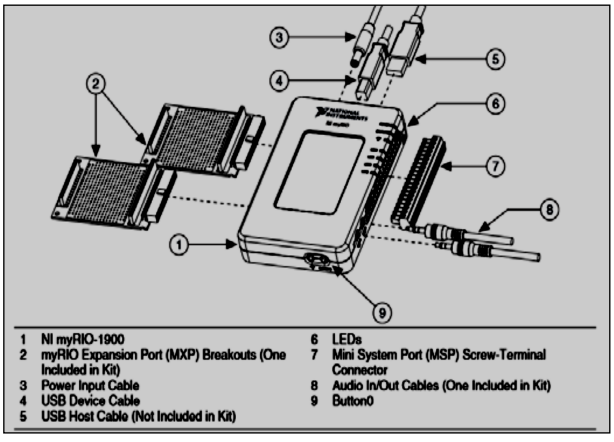


Figure 1: NI MYRIO-1900 and accessories [14].

3.2. Advantages of ZYNQ Board

For cost and Power Reduction, Zynq can combine the features of several components into a single device. When Zynq chips are purchased in bulk, the versatility of the technology enables consumers to use a single chip for several products, thereby lowering costs. Additionally, there is a significant reduction in power usage as a result of the internal processing of data and computations within Zynq rather than data being transmitted from component to component [15].

System Integration with Programmability, Today's cutting-edge businesses can integrate many system components into ever-smaller devices all the way down to a single chip owing to the development of technology that has given rise to the system-on-chip (SoC) concept in the electronics industry. Zynq is a revolution in SoC-level integration [16]. It combines the functionality and power characteristics of an ASIC, the adaptability of an FPGA, and programming simplicity connected to microprocessors into a single device. A 2-chip solution is less dependable and less secure than Zynq because it is a fully integrated system.

Heightened system performance, which distinguishes Zynq from competing products, is the seamless integration of bandwidth between the processing system and programmable logic. This bandwidth is several orders of magnitude larger than what a 2-chip system would be able to provide. As a result, programmers can create accelerators with speeds up to 10 times faster for specific processes [16].

3.3. Why was LabVIEW Chosen?

LabVIEW is a system design tool that employs mathematically based textual and graphical programming techniques to design processes using a graphical user interface. Unlike other common programming methods, LabVIEW includes high-level devices that simulate various communication components and generate various waveforms. In addition, LabVIEW can store data in reports and connect to a variety of measures and sensors. This software has been extensively employed in several industries, including telecommunications, aerospace, and medical research. Because of their adaptability and agility, engineers and scientists need to swiftly develop and test new ideas frequently. One of the main benefits of LabVIEW is its capacity to handle enormous volumes of data, which is crucial for many research applications. As an all-in-one solution for various engineering projects, the software also provides a variety of tools for signal processing, data analysis, and visualization. Overall, LabVIEW is a potent tool that may facilitate and accelerate the design of complicated systems for engineers and scientists [17].

4. Material and Procedure

The transmitter data for the LDPC system consists of an LDPC matrix called the parity check matrix. The data were encoded using the parity-check matrix after generation. We modulated using the BPSK technique, after which it was influenced by the AWGN channel and demodulated before passing it on to the decoder, which decoded it using the min-sum algorithm into the original data.

4.1. LDPC Encoding Stage (Generate the parity-check matrix).

The first step: is to generate a parity-check matrix that defines the relationship between input data and parity bits. This matrix is typically sparse and has a low-density structure.

The second step: The input data is converted into a bit sequence, which is then converted into a bit sequence that can be processed by the LDPC encoder.

The third step: The input data are multiplied with the parity check matrix, and the bit sequence is multiplied with the parity check matrix to generate parity bits.

The fourth step: Adding parity bits to the original data: Generated parity bits are added to the original data to create an encoded message.

The fifth step: Transmitting the encoded message: The encoded message is transmitted over a communication channel

4.2. BPSK Modulation and Channel

To modulate the encoded data, it uses BPSK to convert each bit into another modulated bit. It converts ones to negative 1's. Each zero bit is changed to a positive 1. We take the coded data and then create a loop to generate the modulated data. The AWGN channel model was then used to add Gaussian noise to the modulated data. These modulated data pass through the channel and are corrupted by the AWGN.

4.3. 5G NR LDPC Decoding

The min-sum algorithm is a popular method used in low-density parity-check (LDPC) decoding. It is an iterative algorithm that aims to minimize the sum of the absolute values of the variable node updates during each iteration. The algorithm works by computing the messages sent between variable nodes and check nodes in a bipartite graph representation of the LDPC code.by using LabView "It is one of the most popular engineering software "we can build software for all our system

The implementation of each block is briefly described as follows:

Step 1: The message is generated randomly by using MATLAB, then reading the message from the text file and storing it into a character array to encode all message characters

Table 1: Structure of base graph

Set index (i)	Lifting sizes (Z_c)
0	2, 4, 8, 16, 32, 64, 128, 256
1	3, 6, 12, 24, 48, 96, 192, 384
2	5, 10, 20, 40, 80, 160, 320
3	7, 14, 28, 56, 112, 224
4	9, 18, 36, 72, 144, 288
5	11, 22, 44, 88, 176, 352
6	13, 26, 52, 104, 208
7	15, 30, 60, 120, 240

Step 2: Fill the base graph by selecting one of the existing Base Matrices. Base Graph (Base Matrix.There are 102 codes derived from only two base graph matrices by lifting each with every one of the 51 lifting sizes. There are two types of base graphs for the

3GPP standard. The first, BG 1, had 46 rows and 68 columns. The second is BG 2, which has 42 rows and 52 columns. The choice between them was based on the coding rate and size of the information bits. To get the size of information bits, you should know the value of the lifting size (Zc). For each base graph, there are various values for lifting sizes ranging from 2 to 384 [18].The maximum number of information bits can be calculated using the following formulas, depending on the number of base graphs.

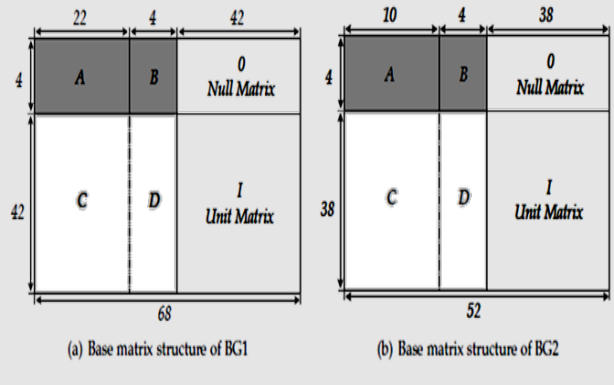


Figure 2: Region division and parameters of base matrix [19].

Step 3: converting Base graph into parity check matrix

Parity Check Matrix: The Parity check matrix was built from the base graphs. Each element in the base graph is replaced by a Zc × Zc matrix. This replacement depends on the value of each element and applies the following rules:

- If the value of the element is (-1), then the Zc × Zc matrix is a matrix of zeros.
- If the value of the element is (0), the Zc × Zc matrix is an identity matrix.
- For each other values, the Zc × Zc matrix will be an identity shifted matrix with the value of the number

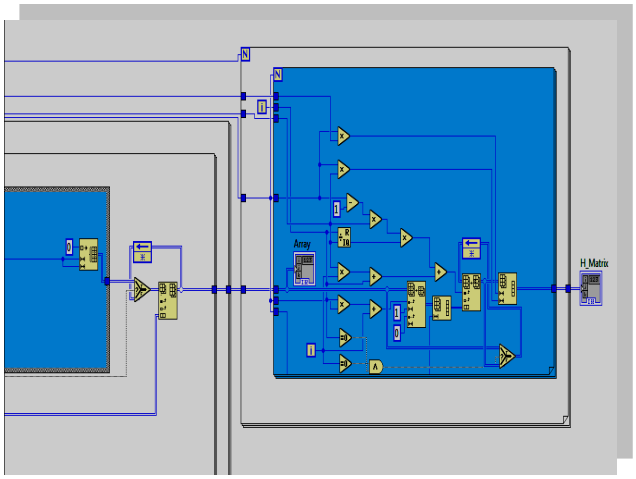


Figure 3: LABVIEW Implementation of H Matrix from block diagram

- Load base graph using read delimited Spreadsheet.vi block.
- Loop each element in the base graph and check the value of this element.
- Develop a corresponding Zc × Zc matrix for each element.
- The developed matrix is rotated to correct the parity-check matrix.

Step 4: Adding parity bits to the messages bits to encoding its.

The main goal of LDPC encoding is to add parity bits to the messages bits. The addition of parity bits to each code block increases the reliability of sent data to develop the code word. The code word is the encodeddata used to transfer it from the transmitter to the receiver.Assume we have an M vector that represents the massage bits donated by M_1, M_2, and M3… M_k. And the parity bits in the P vector which donated by P_1, P_2, and P3… P_n. The code word C is created by concatenating these two vectors which is following that: $C = [MP]$. To confirm that the encoding stage has been done successfully, we should apply the next equations: $C^T = \begin{bmatrix} M^T \\ P^T \end{bmatrix}$ And $H \times C^T = 0$Then developing the parity check matrix following its rules. After that, you could get the parity bits by using this formula: $H \times C^T = 0$ [20]. Finally, the code word is generated easily by adding the parity bits to the information bits [20].

Step 5: Recover the original data Min-Sum Algorithm (MSA). A block diagram is shown below represents the implemented system.

- Initialize specified array called R_{cv} this array contain both variable nodes and check nodes.
- Any non-zero elements in R_{cv} is equal to its relative non-zero elements included in the received data (transmitted data after channel).
- Start to loop every row in R_{cv} with the maximum iteration value.
- Calculate the log of every row and its sign by multiply all signs for each row.

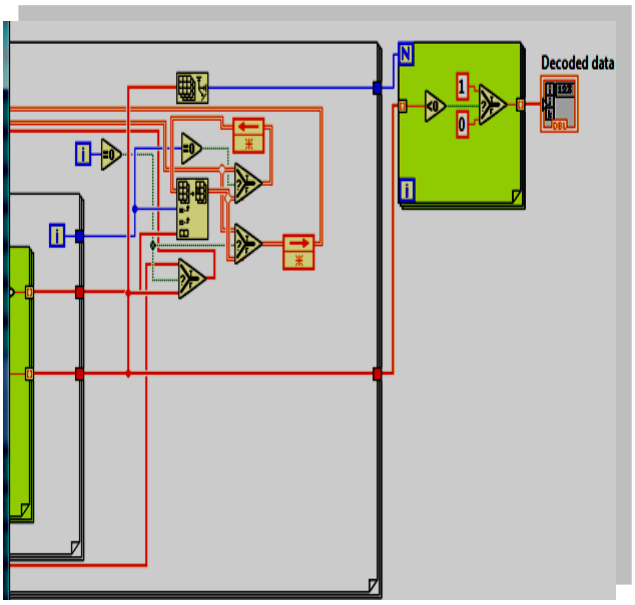


Figure 4: LABVIEW Implementation of 5G-LDPC Decoder

Step 6: Performance testing stage

This section presents the iterative decoding performance of our LDPC codes with different iterations. For the LDPC code, we chose the LDPC code with a code length of 1360 bits and a code rate of 0.5. The min-sum decoding algorithm is used for the LDPC

codes, and the maximum number of iterations is set to 30 because it is sufficient for the convergence of the decoding algorithm. The simulation was performed under an AWGN channel with a BPSK constellation. It should be noted that, to verify the performance of the product codes, we notice that we achieve increased coding gain of our system.

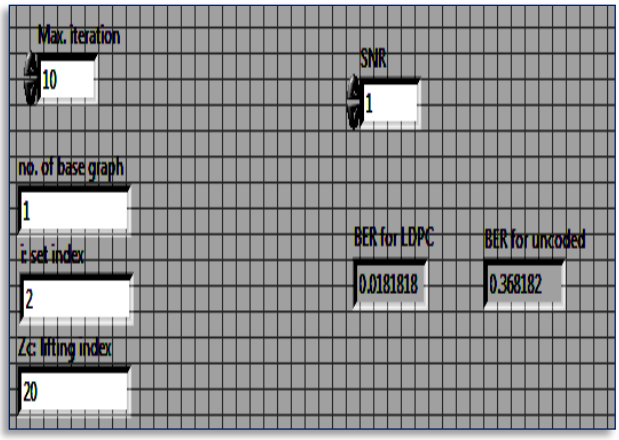


Figure 5: Output of the BPSK in LABVIEW with and without LDPC

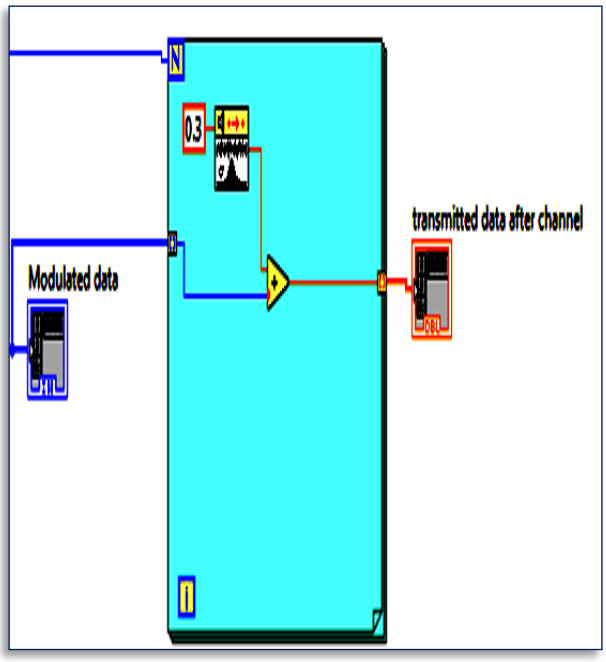


Figure 6: LABVIEW Implementation of transmitted data after channel

We repeat this operation about five times with a different value of SNR. The result was recorded for the LDPC coding data and the uncoded data. The results show the relationship between bit error rate (BER) and signal-to-noise ratio (SNR).For LDPC-coded and uncoded data, when SNR increases, It indicates that the BER is decreasing on a fast track. However, the BER rate for uncoded data is still high, although the SNR has increased. The Min-sum decoder was finished by the LabVIEW FPGA, and a successful compilation was necessary to burn the program onto a Zynq plate card.

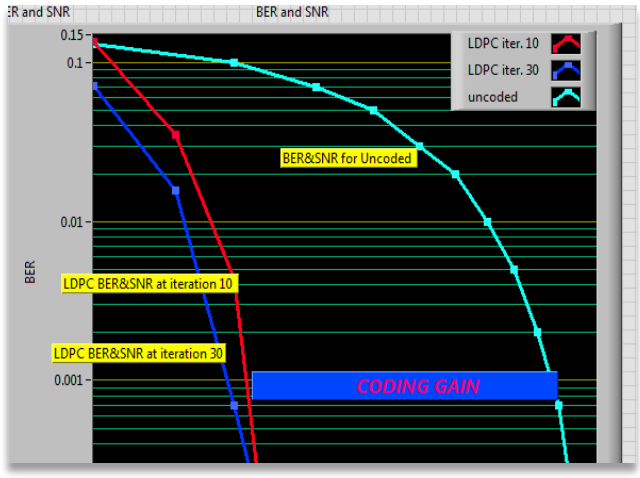


Figure 7: LABVIEW Implementation of BER vs. Eb/No with LDPC

5. The performance of the proposed method by MATLAB

Simulation outcomes are predicted using BER values compared to signal-to-noise ratio values. Our system is written in MATLAB to get the estimated relationship between BER and SNR. Figure 8 shows the relation between BER and SNR in MATLAB. Performance of 5G LDPC code with BG1, BG2, uncoded for code rate of 1=2, code word length of 1360 bits for the min-sum algorithm for the proposed system. The simulation results shown that, we achieves about 6.0dB coding gain. Estimating the Bit Error Rate and Sigma **sys**tem. The simulation results show that we achieve about 6.0 dB of coding gain. Estimating the Bit Error Rate and signal-to-noise ratio involves a number of variables. Where the anticipated value of the seed is no more than 4. As a result, the following table, Table II, shows a comparison of the performance of the output with and without LDPC code**.** The Noise Ratio involves a number of variables. Where the anticipated value of the seed is no more than 4. As a result, the following table,

Table 2: Shows a comparison of the performance of the output with and without LDPC code

parameters	NO LDPC	With LDPC [20]	Our proposed
n	100	200	1360
m	50	100	440
Seed in	-1	-2	-2
EB/NO	5.2	6.18	2.38
Max. iteration	30	30	30

Table 2 show that LDPC codes perform more effectively when comparing resource usage and Eb/No with a variety of LDPC encoder and decoder parameters. Our proposed LDPC achieves small SNR value.

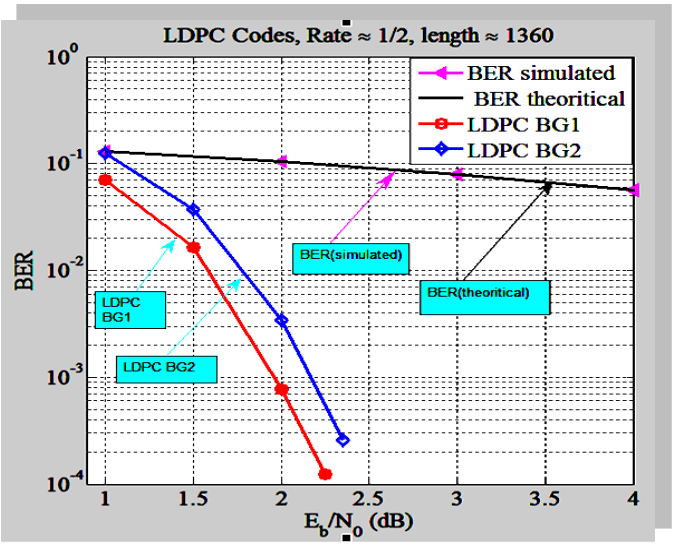


Figure 8: Show that relation between BER & SNR by MATLAB

5.1. FPGA Scheme

- On the PC, install the drivers for the NI yRIO-1900 device and the FPGA module.
- The "create project" button is pressed after the LabVIEW program has opened. The "MyRIO custom FPGA project" option is then chosen on the newly opened interface.
- The project name and MyRIO connection method are chosen.
- .Content is added to the project's main folder at the designated location.
- Next, choose "New VI" from the context menu when you right-click on "FPGA Target" in the next Figure interface. This will add a new VI to the "FPGA Target" section [21]. The block diagram and VI's front panel will also be shown. The gadget is set up in accordance with the chipset.
- The LabVIEW project will be saved with the registered name once it has been developed in the VI. The project is eventually completed. As a result, the
- FPGA in the NI MyRIO-1900 device is set up in accordance with the chipset.

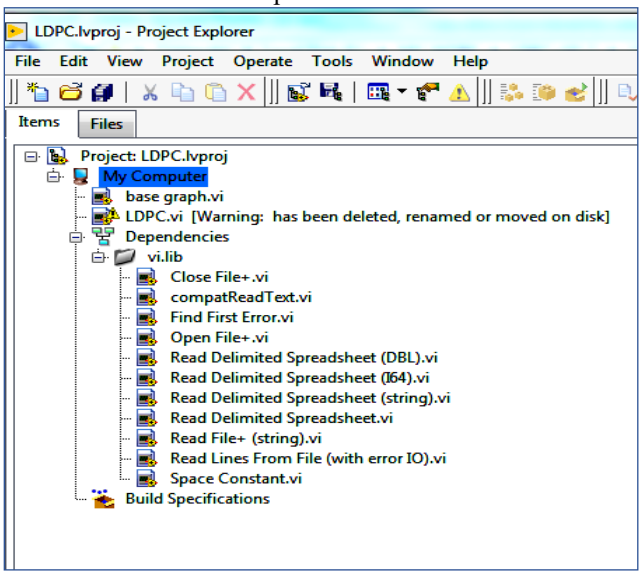


Figure 9: FPGA scheme creation interface

Figure 9 shows what the program looks like on my device where This research, it is thoroughly taught how to develop an FPGA project using the LabVIEW program and load it into the FPGA chip on any target device. The NI MyRIO-1900 was chosen as the study's target device

6. Second method For LDPC Implementation by Using HDL Design

6.1. 5G-NR Encoder Synthesis And Simulation Result

The 5G-NR LDPC system is individually described in Verilog. The different modules in the LDPC system design are designed by the Vivado 18.2 Schematic tool

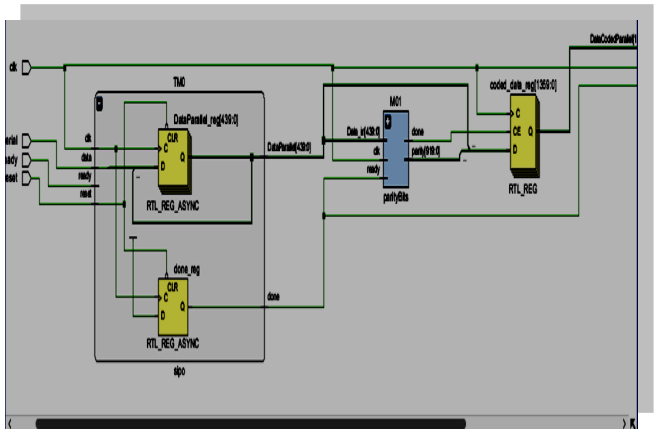


Figure 10: RTL of the Encoder

Figure 10 shows that The RTL for LDPC encoder for constraint length 1360 and bit rate 1/2 has been developed and synthesis is done shown that the resulting shapes from synthesis on LDPC encoder.

- The following table show that the Encoder hardware resources utilization.

Table 3: Encoder Hardware resources utilization.

Resource	utilized	Available	Utilization
LUT	390	70560	0.55
FF	596	141120	0.42
IO	3	82	3.66
BUFG	1	196	0.51

Table 3 displays specifics regarding the utilization of the hardware resources. It provides data on the number of slices, flip-flops, LUTs, IOBs, global buffers, and BUFG

The following simulation shows how input is provided to the encoder once the data read assertion has been made and how the encoder is asserted following the completion of the encoding

operation. The LDPC encoder yields meaningful results, as demonstrated in Figure 11

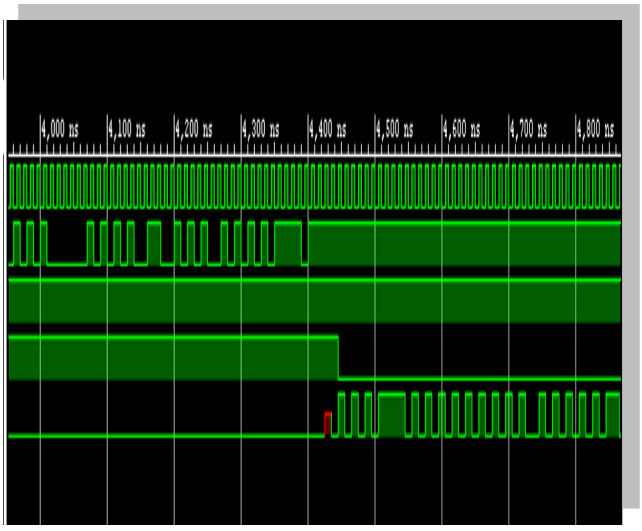


Figure 11: Simulation results of the Encoder output obtained from Vivado simulation.

We notice that the output stat appears after 4.400 ns. The channel sends the coded bits to the decoder after encoding. There are no errors in this simulation. It is possible to observe the asserted control signals for each as shown in Figure 11.

6.2. 5G-NR Decoder Synthesis And Simulation Result

The following Shape shows the process of decoding. This figure shows that the input is entered in parallel. The same thing appears in the output, which also appears in parallel. It shows that 14 bits appear in one microsecond.

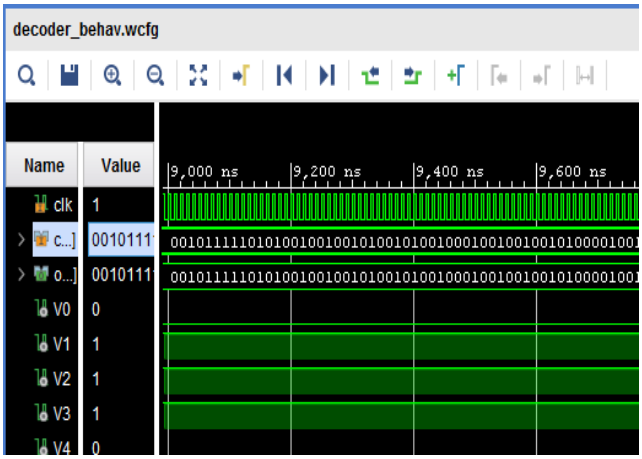


Figure 12: Simulation results of the decoder output obtained from Vivado 18.2

A shown in Figure12 that the simulation output from the decoder, if we compare the data input of the LDPC encoder shown in Figure 11 with the output of the LDPC decoder decided in Figure 12, it is shown that the NRLDPC decoder perfectly estimates the original data. Considering the effect of noise, if the encoded data is corrupted, then the decoder should also be able to retrieve the original message sequence. The decoder can detect and correct the errors with excellence.

7. Device utilization reports

It was shown that the resulting shape from the synthesis of the architecture was placed and routed in a ZYNQ FPGA with the highest speed grade (-2) using Vivado 18.2 from Xilinx. The following shown is the part from circuit of the LDPC decode.

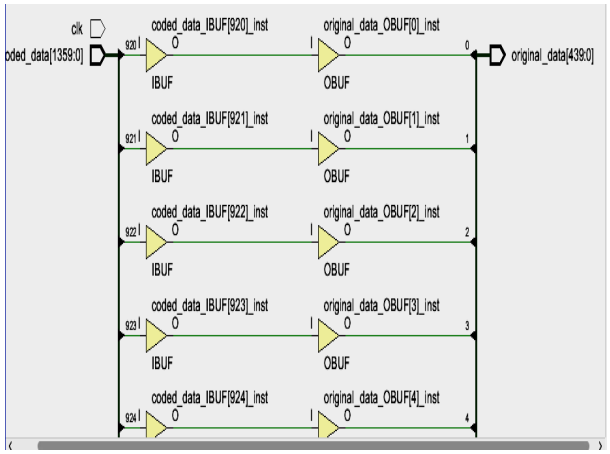


Figure 13: RTL of LDPC decoder unite

Figure 13 shows that The RTL LDPC Decoder for constraint length 1360 and bit rate 1/2 has been developed and synthesized.

All architectures were synthesized, placed, and routed in a kit FPGA with the highest speed grade (-2) using Vivado 2018.2, The ZYNQ (xczu3eg-sbva484-1-e) was chosen to build the system using Vivado 2018.2 program, as shown in Table IV below.

Table 4: Hardware Decoder Utilization for Our Proposed Design

Resource	Utilized	Available	Utilization
LUTs	159	70560	0.28
LUTRAM	58	28800	0.20
FF	2275	14112	1.61
IO	4	82	4.88
BUFG	1	196	0.51

7.1. Timing and power summary

- Speed grade: -2
- Minimum period: 30.3035 ns
- Maximum frequency: 33 MHz
- Minimum input arrival time before the clock: 1.245 ns.
- Maximum output required time after the clock: 0.809 ns

8. Comparative Analysis between Various Fpgas Devices

It shows that from our proposed implementation, we have an area overhead reduction of more than 50% compared to referenced designs, and we can achieve more than 21000 LUTs compared with other designs LUTs, where other designs consume

more FFs and BRAMs. Furthermore, it has much higher throughput (224 vs. 87 and 5 MB/s). So ZYNQ's (xczu3eg-sbva484-1-e) FPGA is better than previous state-of-the-art solutions in terms of area. Moreover, the implemented 5G NR LDPC decoder was tested against an additive white Gaussian noise channel (AWGN) and consequently has gained more and more popularity in many applications. The frequency of the 5G NR LDPC decoder obtained is 33 MHZ with a minimum period of 30.3035 ns. The main limitation of the implemented decoder appears to be small lifting factor values. For them, it is necessary to freeze the decoder for several clock cycles during one iteration to provide enough time for Check message memories to update. This reduces the decoder's throughput and power. The Vivado tool reports a total on-chip power of 5.897W. After Making a comparison between our results with another LDPC architectures in[22], [23], and [24]. In [22], used the Xilinx Kintex-VII architecture. In [23] used an FPGA, a Xilinx Virtex 7. In [24] Xilinx Kintex Ultrascale FPGA.

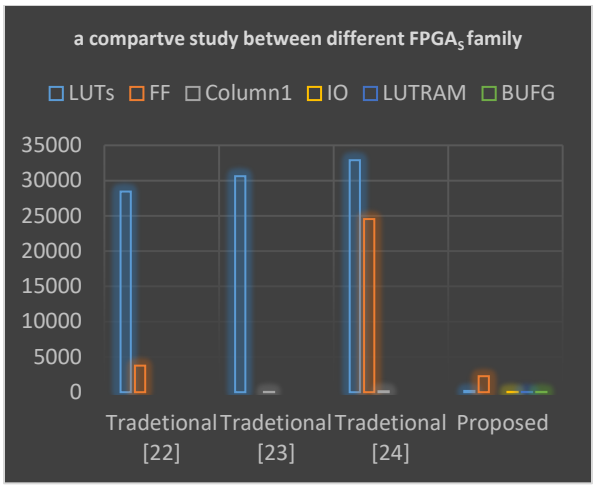


Figure 14: Show that a comparative study between Implementation results using different FPGAs family kits

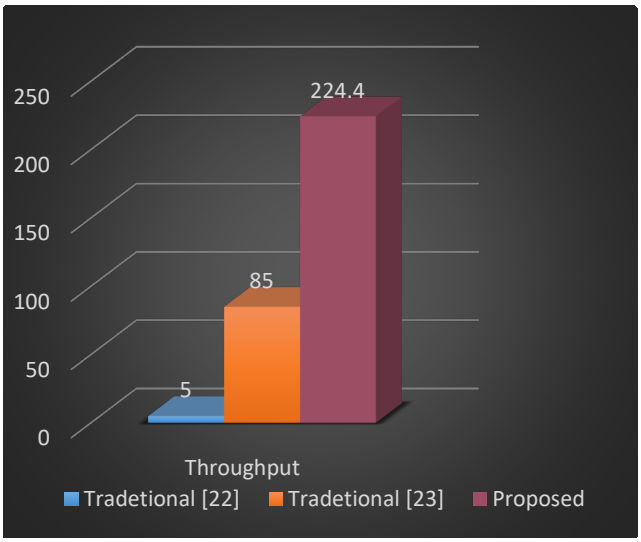


Figure 15: Performance comparison of Throughput consumption

Our proposed implementation is done using ZYNQ Xilinx Ultrascale (xczu3eg-sbva484-1-e). However [23] has a better throughput value than [22], architecture resources have been very high according to throughput value. Our proposed implementation has 38.77% improvement rate in throughput than [23]. On other hand, [24] consumes 137.5 BRAM which is very high. But In our implementation, we replaced the BRAMs in [24] with 58 LUTRAMs. Our implementation progress is due to the proposed architecture. This method saves the resources that are used for converting the base graph to the parity check matrix. The architecture of ZYNQ Ultrascale allows the parallelism of the design in order to increase throughput. So, we notice that if we look at the table as well, we will find that our proposal is to save power, and this is the best thing in modern systems, which is to save power. Our implementation provides saving resources utilizations compared to other designs.

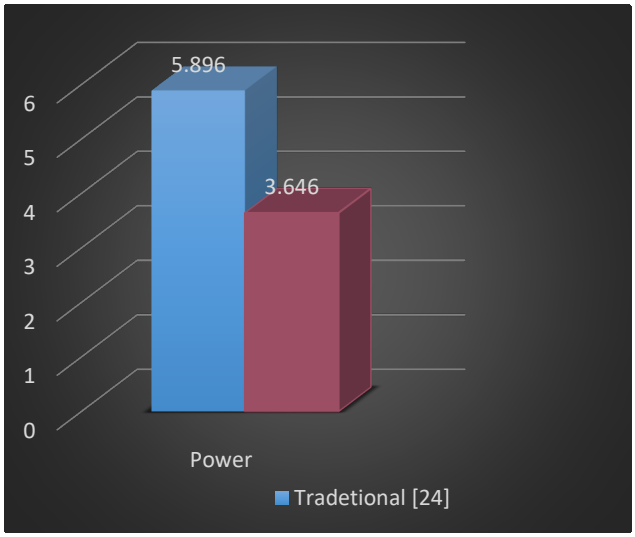


Figure 16: Performance comparison of Power consumption

In Figure15, it shows the difference between our proposed design and [24]. The design we introduce saves about 2.251 mW. This is a regular result when using less resources.

9. Conclusion

5G-NR LDPC decoding is the best technique for decoding codes. In this work, an area overhead reduction has been achieved compared with respect to previous work at the same code rate. 5G-NR LDPC decoders with constraint length 1360 and code rate 1/2 have been successfully implemented on the Virtex-6 FPGA NI MY RIO Zynq-1900 device of the NI family and realized using Xilinx Vivado 18.2. The MIN-SUM algorithm is employed in wireless communication to decode the 5G-NR LDPC codes; those codes are used in every strong digital communication system. We have implemented an area-optimized VLSI architecture for the 5G NR LDPC Encoder and 5G NR LDPC Decoder. The 5G NR LDPC decoder allows safe data transmission via error correction, and the original message can be recovered accurately. This proposed 5G NR LDPC algorithm presents a reduction in power and cost and, at the same time, an increase in processing. Speed. 5G-NR LDPC code has better performance than other LDPC codes, especially at low Eb/N0 values. This means we save on power. Furthermore, it is much better if we compare the un-coded performance. This is besides enhancing throughput. Among the advantages of LDPC is that, by increasing the number of iterations,

the performance is enhancedmore and more**.** In this work, area and power overhead reductions have been achieved compared with reference works at the same code rate. Field Programmable Gate Array (FPGA) technology is considered a highly configurable option for implementing many progressive signals. In future work, we will try to improve the latency of the LDPC decoder.

Conflict of Interest

The authors declare no conflict of interest

References

[1] G. Suchitra, "Performance Analysis of Ldpc Code in Digital Communication System Using LabVIEW," Hindusthan Journal of Information Communication and, **1**(1), 11–18, 2020.

[2] J. Xu, Z. Zheng, K. Tian, "Low-Density Parity-Check Codes: Research Status and Development Direction," Journal of Information Security, **13**(04), 257–271, 2022, doi:10.4236/jis.2022.134014.

[3] K. Huo, Z. Hu, D. Liu, Design and Implementation of Shared Memory for Turbo and LDPC Code Interleaver, Wireless Communications and Mobile Computing, **2022**, 2022, doi:10.1155/2022/5782199.

[4] A. Devrari, A. Kumar, A. Kumar, "FPGA Performance Analysis of LDPC and Turbo Codes for Communication System," 2022.

[5] M. Pregara Zachary Saigh Advisors, I. Soo Ahn Yufeng Lu, "Low Density Parity Check Code Implementation Senior Project Final Report Students," 2013.

[6] A.A. Ovchinnikov, A.A. Fominykh, "Decoding of LDPC codes for 5G standard using source distribution," 2020 Wave Electronics and Its Application in Information and Telecommunication Systems, WECONF 2020, 0–4, 2020, doi:10.1109/WECONF48837.2020.9131454.

[7] K.W. Jensen, Polar Codes and LDPC Codes in 5G New Radio.

[8] M. Indoonundon, T. Pawan Fowdur, "Overview of the challenges and solutions for 5G channel coding schemes," Journal of Information and Telecommunication, **5**(4), 460–483, 2021, doi:10.1080/24751839.2021.1954752.

[9] V.L. Petrović, D.M. El Mezeni, A. Radošević, "Flexible 5g new radio ldpc encoder optimized for high hardware usage efficiency," Electronics (Switzerland), **10**(9), 2021, doi:10.3390/electronics10091106.

[10] L.Y. Hosni, A.Y. Farid, A.A. Elsaadany, M.A. Safwat, "5G New Radio Prototype Implementation Based on SDR," Communications and Network, **12**(01), 1–27, 2020, doi:10.4236/cn.2020.121001.

[11] G.M.G. Madhuri, A.V.K. Sravanthi, A.R. sri, G.S. Mrunalini, G.S. Kumar, "Implementation of Decoder Using Ldpc Codes on Fpga," International Journal of Engineering Applied Sciences and Technology, **04**(12), 354–357, 2020, doi:10.33564/ijeast.2020.v04i12.062.

[12] D.S. Lenin, H. Shekhar, "Design and Implementation of Optimized LDPC for SDR Applications," … Journal of Engineering, Management, & Applied, 1–10, 2021, doi:10.14456/ITJEMAST.2021.15.

[13] J. Nadal, A. Baghdadi, "FPGA based design and prototyping of efficient 5G QC-LDPC channel decoding," Proceedings of the International Workshop on Rapid System Prototyping, **2020-Septe**, 2020, doi:10.1109/RSP51120.2020.9244853.

[14] U. Guide, "NI myRIO-1900," 1900.

[15] M. Ali Altuncu, T. Guven, Y. Becerikli, S. Sahin, "Real-Time System Implementation for Image Processing with Hardware/Software Co-design on the Xilinx Zynq Platform," International Journal of Information and Electronics Engineering, **5**(6), 473–477, 2015, doi:10.7763/ijiee.2015.v5.582.

[16] A.A. Rafiq, M. Yusuf, Pujono, "Implementation of Digital Image Processing Using NI myRIO and Arduino Mega 2560 as Controller on Rover Bogie Robot," Proceedings - 2018 International Conference on Applied Science and Technology, ICAST 2018, (March 2020), 210–215, 2018, doi:10.1109/iCAST1.2018.8751506.

[17] S. Mhaske, H. Kee, T. Ly, A. Aziz, P. Spasojevic, "FPGA-Based Channel Coding Architectures for 5G Wireless Using High-Level Synthesis," International Journal of Reconfigurable Computing, **2017**, 2017, doi:10.1155/2017/3689308.

[18] L. Wang, "Implementation of Low-Density Parity-Check codes for 5G NR shared channels," Degree Project Computer Science and Engineering, 2021.

[19] Y. Tian, Y. Bai, D. Liu, "Low-latency qc-ldpc encoder design for 5g nr," Sensors, **21**(18), 2021, doi:10.3390/s21186266.

[20] P. Fang, "Application Analysis of Low Density Parity Check Code for Wireless Network," IOP Conference Series: Earth and Environmental Science, **634**(1), 2021, doi:10.1088/1755-1315/634/1/012043.

[21] E. Y., H. Çıtak, M. Çoramık, "A FPGA Project with MYRIO," International Advanced Technologies Symposium, (December), 19–22, 2017.

[22] A. Verma, R. Shrestha, "A New Partially-Parallel VLSI-Architecture of Quasi-Cyclic LDPC Decoder for 5G New-Radio," in 2020 33rd International Conference on VLSI Design and 2020 19th International Conference on Embedded Systems (VLSID), IEEE: 1–6, 2020, doi:10.1109/VLSID49098.2020.00018.

[23] J.C. Liu, H.C. Wang, C.A. Shen, J.W. Lee, "Low-Complexity LDPC Decoder for 5G URLLC," Asia Pacific Conference on Postgraduate Research in Microelectronics and Electronics, **2018-Octob**, 43–46, 2018, doi:10.1109/PRIMEASIA.2018.8597812.

[24] A. Katyushnyj, A. Krylov, A. Rashich, C. Zhang, K. Peng, "FPGA implementation of LDPC decoder for 5G NR with parallel layered architecture and adaptive normalization," Proceedings of the 2020 IEEE International Conference on Electrical Engineering and Photonics, EExPolytech 2020, 34–37, 2020, doi:10.1109/EExPolytech50912.2020.9243997.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 8, No. 4, 101-110 (2023)

www.astesj.com

ASTESJ

ISSN: 2415-6698

Application of Lean Practices in Food Supply Chain: The Case of Morocco

Fadwa Bouhannana*, Akram El Korchi

Systems engineering laboratory and decision support, National School of Applied Sciences, Ibn Zohr University, Agadir, Morocco

ARTICLE INFO

Article history:
Received:10 April, 2023
Accepted:11 July, 2023
Online: 26 August, 2023

Keywords:
Lean manufacturing
Practices
Food Supply chain
Performance
Measure

ABSTRACT

Recent studies show the benefits of lean manufacturing implementation in agri-food industries to improve operational and environmental performance. However, only a restricted number of studies have addressed the implementation of lean practices in food companies located in developing countries. This study aims to assess the current implementation status of lean practices in Moroccan agri-food companies, particularly small and medium-sized enterprises, and to examine their impact on operational and environmental performance. Responses from 45 agri-food companies were collected through a questionnaire. The results show that the degree of implementation of lean practices in the Moroccan food industry is generally average. On the other hand, some lean practices are implemented more frequently compared to others., e.g. customer involvement, employee involvement, supplier involvement and total productive maintenance are the most implemented. On the other hand, pull and setup are not used very much. The findings also demonstrate the positive impact of lean practices on both operational and environmental performance.

1. Introduction

The term "lean" was initially coined in 1988 by [1] to characterize the Toyota production system. Subsequently, in 1990, [2] defined lean manufacturing as a system that achieves equivalent outputs while utilizing fewer inputs, resulting in increased value for the end-user. This particular definition highlights the importance of waste identification and elimination [3]. Empirical studies have consistently demonstrated that lean manufacturing has positive impacts on productivity, quality, and both customer and employee satisfaction [4], [5].

Numerous studies in the literature have demonstrated the advantages of lean implementation, which can take the form of quantitative improvements, like production cycle time, inventory, defects, and waste, or qualitative benefits like customer satisfaction, good communication, and job satisfaction [6]. These findings are consistent with previous research by [7] and [8], which established a positive association between lean implementation and operational performance. Additional studies, conducted by [9] and [10], have also provided evidence of the advantages of lean implementation on operational performance. The literature suggests that lean implementation not only leads to improved operational performance [9], [10], but also helps companies achieve their environmental objectives [11].

Numerous studies have explored the suitability of implementing lean manufacturing in the food industry. [12], [13], while others have chosen to ignore the agri-food sector due to the challenges related to the seasonal variability and storages issues [14].

Various studies have been conducted on lean implementation in agri-food companies of developed countries, whereas, few research has focused on the context of developing countries. This study aims to bridge this gap by examining the extent of lean practice implementation in agri-food companies in a developing country, such as Morocco. It also examines the effects of lean practices on the operational and environmental performance of these companies. The primary objective is to assess the current status of lean practices in Moroccan agri-food companies, particularly small and medium-sized enterprises, and evaluate their impact on operational and environmental performance. Responses from 45 agri-food companies were collected through a questionnaire. This study addresses the effect of control variables such as firm size [15]–[17], location and the existence of the lean project. Finally, according to the author's knowledge, this study is the first attempt assessing the impact of lean practices on operational and environmental performance in Moroccan

* Corresponding Author: Fadwa Bouhannana, fadwabouhannana@gmail.com

agribusiness and analyze the present status of implementation of lean manufacturing practices among food Moroccan agri-food companies (small and medium-sized). Except for [18], who explored an empirical study in Moroccan agri-food firms to determine the principles of lean adopted by these firms. What makes this study limited is the fact that they chose a sample of only nine companies.

The study endeavors to provide answers to the following two research questions:

1. To what extent are lean practices implemented in agri-food companies in Morocco?
2. What is the influence of lean practices implementation on operational and environmental performance?

The paper follows the following structure: Section 2 provides a comprehensive literature review. Section 3 outlines the conceptual framework utilized in the study. The research methodology is detailed in Section 4. The study's findings are summarized in Section 4. The implications of these results are discussed in Section 5, and the paper concludes with final conclusions in the last section.

2. Theoretical background

2.1. Lean manufacturing

The success initially attained by Toyota, followed by various organizations worldwide, has prompted numerous companies to embark on lean projects [19]. Presently, lean has gained popularity not only in developed countries but also in select developing nations [20], [21]. However, each organization implements lean practices based on its unique circumstances, making it impossible to follow a single, standardized recipe for guaranteed success [22].

Lean manufacturing is a business philosophy that focuses on optimizing customer value while minimizing waste across various areas of operation. This includes addressing issues such as overproduction, waiting time, defects, overprocessing, excess inventory, unnecessary motion, and underutilized talent. In the agri-food sector, Lean principles have been applied to various aspects of food production, from farm to table. A study conducted by [23] that found that Lean principles can be applied effectively in the agri-food sector to improve processes such as supply chain management, production planning, and process flow. In [24], the author summarized into different principles: defining value, defining the value chain, flow and standardization.

Studies have consistently demonstrated that the adoption of Lean principles yields favorable outcomes in terms of both operational and environmental performance. In [23], the author found that Lean principles can improve supply chain management, production planning, and process flow, leading to increased efficiency and reduced waste in the agri-food sector. Another study by [13] found that Lean implementation can help agri-food businesses reduce costs and improve quality, leading to increased competitiveness.

Numerous studies have provided robust evidence indicating that Lean practices have a significant and positive impact on multiple dimensions of operational performance [24], [25].

Several studies have identified and adopted five key operational performance indicators as outlined [12], [15]. These indicators encompass reduced inventory [26], improved quality [26], [27], increased productivity [3], [28], shortened cycle time [28], [29], and enhanced delivery time [30], [31]. In addition, the adoption of Lean principles can indeed contribute to improved environmental performance, given its emphasis on waste reduction and minimizing negative environmental impacts. A study conducted by [32] investigated the influence of the five core Lean methods on four environmental performance measures. (material utilization, energy consumption, non-product outputs, and pollutants released). They found that TPM and JAT had a significant impact on environmental performance. However, continuous improvement only had an effect on material utilization. In [33], the author found that Lean implementation in the meat processing industry reduced waste and improved production efficiency, leading to reduced energy use and decreased greenhouse gas emissions. These findings suggest that Lean implementation can have a positive impact on both operational and environmental performance in a variety of industries.

2.2. Implementing lean practices in the agri-food companies

The agri-food industry holds a vital position within the Moroccan economy [18], [34]. The agri-food industry, recognized as a significant pillar of the national economy [35], is distinguished by its diverse agricultural activities, expanding export markets, and potential for sustainable development. Despite the important position of this sector, most companies face several challenges: operational, financial, commercial, and technological [18]. The agri-food industry necessitates a multifaceted production process, dealing with a diverse range of products characterized by varying levels of perishability. This industry also faces challenges associated with production time variations and the need to meet diverse customer demands.

Current literature emphasizes the significance of implementing lean manufacturing practices in the food industry as a means to improve operational efficiencies [36]–[39]. However, the implementation of lean practices in the agri-food sector presents unique challenges, primarily due to seasonal variations, large-scale production, processing complexities, and issues related to storage. It is important to recognize that lean principles and tools, initially developed for mass production of non-perishable goods, may not have universal applicability across all industries, including the food industry [40].

Numerous studies have explored the adaptability of lean manufacturing principles in the context of the food industry [12], [13]. However, some studies have overlooked the agri-food sector due to its unique challenges associated with seasonal variability and storage issues [41]. Nevertheless, research conducted by [10] and [42] has demonstrated the substantial benefits that lean manufacturing can bring to the food industry. Again, lean can help

agri-food companies reduce non-value-added time, reduce waste, and increase the percentage of operational value-added activities [10], [42].

Over the years, there has been a noticeable surge in research focusing on lean practices within the agri-food industry. This increasing trend can be attributed to the growing recognition of the importance of waste reduction and the expanding population within the agri-food supply chain [43], [44]. However, limited research has provided tangible evidence of the actual impact resulting from the implementation of lean practices [45] with few exceptions such as output projections derived from modeling-based [46]. Most studies in this domain have predominantly relied on case studies, action research, or perception-based impact evaluations, which possess limited generalizability in assessing the effects [38]. According to [47], a significant proportion of studies conducted in the food industry are based on case studies, with only a limited number relying on empirical surveys [48]. Consequently, the results pertaining to the application of lean manufacturing in the food sector remain somewhat contentious. In [49], the author demonstrated how the VSM technique can be applied to Finnish food plants. In a study conducted by [50], a case study approach was employed to examine medium-sized cookie companies. The results indicated that the application of lean techniques such as 5S, kaizen, changeover, and TPM (Total Productive Maintenance) led to improvements in equipment availability, reduction in material waste, and enhancements in overall product quality.

The objective of this research is to assess the level of implementation of lean practices in agri-food companies within developing countries, with a specific focus on Moroccan companies. The study uses a conceptual model of lean manufacturing by [3] that covers both internal and external practices. The ten key components of lean manufacturing identified by Shah and Ward cover supplier engagement, customer involvement, and internal company issues. Operational performance is measured using [48], while environmental performance is assessed using various sources, including [51]–[56].

3. Development of the conceptual model

Scholarly journals have recently published many articles on the impacts of lean practices on operational and environmental performance. The literature review identifies commonly associated practices between lean practices and performance metrics. The lean conceptual model, presented in Table 1, is based on this review and divided into three levels: concepts, strategies, and practices. The model incorporates sections from different literature models and studies.

4. Methodology

This study was conducted in Morocco, within agri-food companies. The sectors of activity chosen in this study are canned goods: fish, capers, jams, olives, and tomatoes. A questionnaire was designed to measure the implementation of lean practices and their impact on operational and environmental performance. Prior to its administration, the questionnaire underwent testing and validation by three managers and two academics to ensure its reliability and validity. The questionnaire has five sections. The first section is reserved for general information about the company. The second section concerns the implementation of lean practice. The third section is used to extract the respondents' perceptions regarding implementing these practices on performance. A manager in the Ministry of Industry was contacted to obtain the database of Moroccan companies.

A total of 230 companies operating in the agri-food sector received the questionnaire for participation in the study. A total of 45 respondents answered the questionnaire.

The Likert scale that was adopted in this questionnaire had 5 points. The companies were requested to indicate the extent of lean practice implementation within their respective organizations. (1= Not implemented, 5= Strongly implemented).

5. Results

5.1. Descriptive analysis

In this study, 45 Moroccan companies participated, with 72% confirming the implementation of a lean project in their respective companies, while 28% reported not having a lean project. The majority of the participating companies were small and medium-sized enterprises (70%), with the fish industry representing 33% and the fruits and vegetables industry representing 27% of the sectors. The companies' distribution across regions was relatively equal, with 31% from the north, 34% from the center, and 35% from the south regions (Figure 1). Regarding the work experience of the respondents, notable percentages include 22% with 10 years of experience, 16% with 6 years of experience, and 13% with 16 years of experience. Figure 1 demonstrates that Logistics with 29%, Quality with 24%, and Production with 16% were the three departments that significantly participated in the study.

5.2. Implementation level of lean practices

In response to the initial inquiry, a descriptive analysis was performed to evaluate the extent of adoption of all Lean practices (as presented in Table 2).

The results show that the implementation of Lean practices in Moroccan agri-food firms is relatively moderate, with an average score of 3.15. Nevertheless, the questionnaire data reveals a considerable variation in the implementation level of each practice. Specifically, practices associated with preventive maintenance, employees, customers, suppliers, and flow are highly implemented in Moroccan agri-food companies, while practices belonging to the categories of Pull, Setup, and SPC are less widely adopted.

Three non-parametric tests were administered to examine if there were any differences in the responses and to assess the scores of each variable outlined in Table 3. The initial test utilized in this study was the Friedman test, which aimed to assess the perceived differences in the levels of implementation of lean practices. The findings indicate that there is a noteworthy distinction in the level of implementation of certain practices : Pull, supplier involvement, setup and TPM, as depicted in Table 2, with p-values less than 0.05.

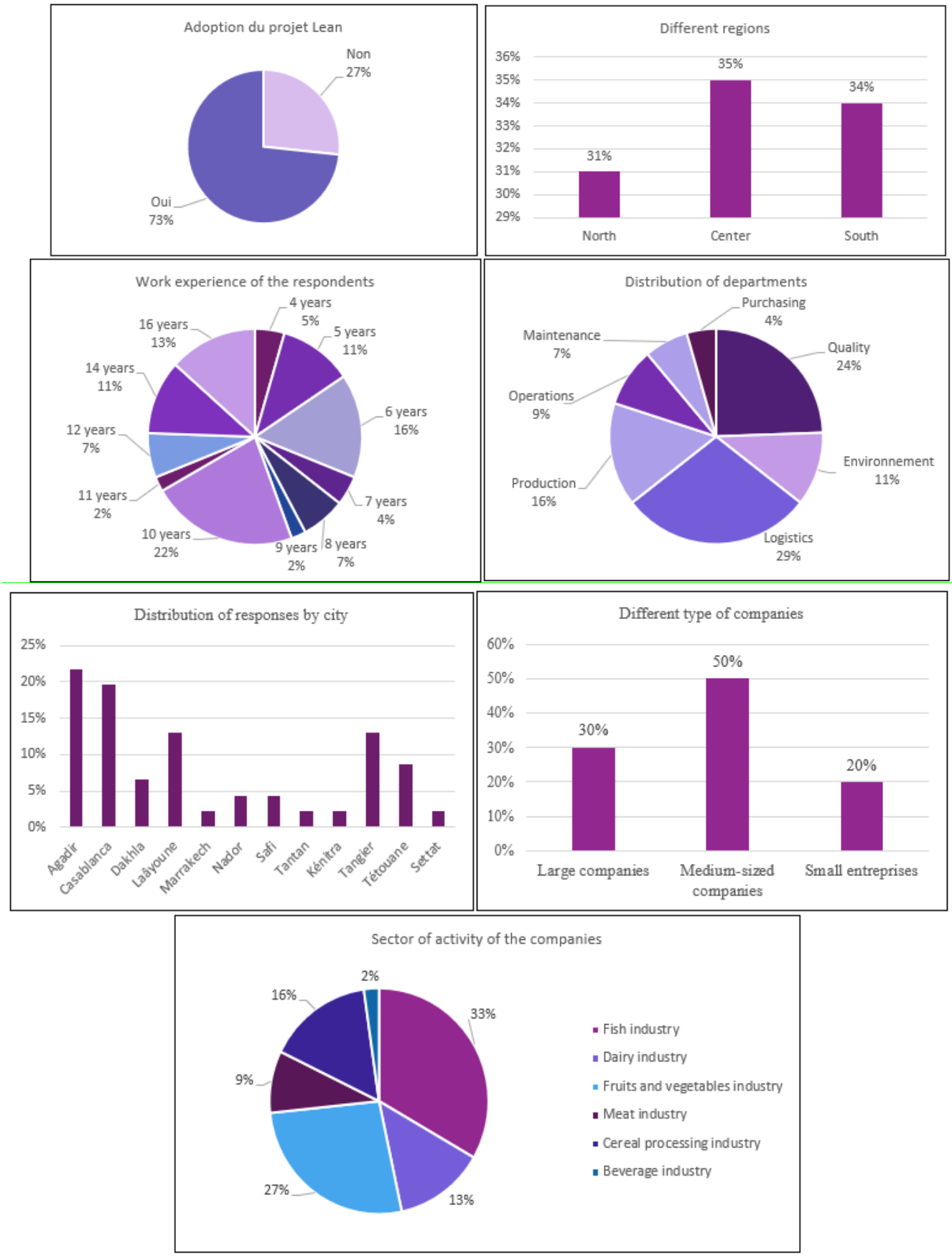


Figure 1: Descriptive analysis

Table 1: Conceptual model for lean implementation

Concepts	Strategies	Pratiques	References
Lean	Waste reduction	5S	[57], [58]
		Reduction in the size of product batches	
		Reduction of excess inventory	
		Reducing scrap rate	
		Layout of the workshop in order to optimize the process sequences	
	Balancing of production lines	Balancing of production lines	[57], [58]
		Stabilization of the rate of production machines	
		Elimination of bottlenecks	
		Synchronization of the production line rhythm according to customer demand	
	Pull	Pull production system	[48], [58], [59]
		Use of the Kanban	
		Production at one station is governed by the demand emanating from the subsequent station.	
	Flow	The products are categorized into families based on their similar processing requirements	[48], [58], [59]
		The factory layout is determined by the product families.	
	Setup	Reduced changeover time	[58], [59]
		The production cycle time is closely monitored to ensure a swift response to customer requests	
		Use of techniques such as SMED (Single Minute Exchange of Die(s)) to reduce changeover time	
	SPC	SPC (Statistical Process Control)	[57], [58]
		Ishikawa diagram	
		Use of visual tools (Control charts...)	
		Statistical techniques to reduce variance in processes	
		Process capability analysis prior to product launch	
	Total Preventive Maintenance	Preventive maintenance plan	[48], [58], [59]
		Time is set aside each day to plan activities related to equipment maintenance	
		Equipment maintenance checklists are posted regularly in our workshops	
	Employee involvement	Workshop employees are trained to perform several functions	[48], [58], [59]
		Workshop employees play a crucial role in problem-solving teams.	
		Shop floor employees participate in process and product improvement efforts	
	Supplier involvement	We maintain regular and close contact with our suppliers.	[58], [59]
		Our primary suppliers follow a just-in-time delivery approach, providing goods to the factory as needed.	
		We are proactively working to minimize the number of suppliers in each category.	
	Customer Involvement	Our customers provide us with feedback regarding the quality and delivery performance of our products.	[58], [59]
		We deliver to our major customers just-in-time	
		Our customers frequently share market demand with our sales department	

		Our customers are directly engaged in shaping our product offerings.	
Performance	Operational performance	Reduction of the production cycle time	[48]
		Reduction in delivery time	
		Increasing productivity	
		Reduction of customer complaints	
		Production cost reduction	
		Product Quality Improvement	
		Scrap Rate Reduction	
	Environmental performance	Reducing of consumption of hazardous/toxic materials	[52]
		Minimization of air emissions.	[51]–[53], [55]
		Minimization of water consumption	[51]–[53], [55]
		Minimization of solid waste	[52]
		Reduction of environmental accidents	[52], [55]
		Reducing CO2 emissions from transportation	[55]
		Decrease of cost for energy consumption	[52]
		Decrease food wastes	[54]

Table 2: Implementation of lean practices in Moroccan agri-food companies

Abbreviation	Section	Practices	Mean
S1	Waste reduction	5S	2.96
		Reduction in the size of product batches	3.07
		Reduction of excess inventory	3.14
		Reducing scrap rate	3.25
		Layout of the workshop in order to optimize the process sequences	3.11
S2	Balancing of production lines	Balancing of production lines	3.46
		Stabilization of the rate of production machines	3.29
		Elimination of bottlenecks	3.11
		Synchronization of the production line rhythm according to customer demand	3.43
S3	Pull	Pull production system	2.96
		Use of the Kanban	2
		Production at one station is governed by the demand emanating from the subsequent station.	2.89
S4	Flow	The products are categorized into families based on their similar processing requirements	3.54
		The factory layout is determined by the product families.	3.07
S5	Setup	Reduced changeover time	2.68
		The production cycle time is closely monitored to ensure a swift response to customer requests.	3.32
		Use of techniques such as SMED (Single Minute Exchange of Die(s)) to reduce changeover time	2.21
S6	SPC	SPC (Statistical Process Control)	2.54
		Ishikawa diagram	2.93
		Use of visual tools (Control charts...)	2.68
		Statistical techniques to reduce variance in processes	2.68
		Process capability analysis prior to product launch	3.21
S7		Preventive maintenance plan	3.57

	Total Preventive Maintenance	Time is set aside each day to plan activities related to equipment maintenance	3.32
		Equipment maintenance checklists are posted regularly in our workshops	3.14
S8	Employee involvement	Workshop employees are trained to perform several functions	3.54
		Workshop employees play a crucial role in problem-solving teams	3.46
		Shop floor employees participate in process and product improvement efforts	3.25
S9	Supplier involvement	We are frequently in close contact with our suppliers	4
		Our primary suppliers follow a just-in-time delivery approach, providing goods to the factory as needed.	3.39
		We are proactively working to minimize the number of suppliers in each category	2.68
S10	Customer involvement	Our customers provide us with feedback regarding the quality and delivery performance of our products.	3.71
		We deliver to our major customers just-in-time	3.75
		Our customers frequently share market demand with our sales department	3.57
		Our customers are directly engaged in shaping our product offerings.	3.5

Wilcoxon test is a second test which aimed to compare the levels of utilization of diverse lean practices. The results suggest that there is a considerable contrast among the lean practices linked to Pull, statistical process control, Flow, Setup, and supplier involvement. However, there is no substantial variation observed in the other practices, with p-values greater than 0.05.

As part of this study, an inferential analysis was carried out to investigate the relationship between the control variables (presence of a lean project and company size) and the execution of lean practices in agri-food enterprises. The Kruskal-Wallis test indicate that a significant contrast exists between the companies that have a lean project and those that do not, in terms of the practices related to Pull, employee involvement, and TPM with p-values less than 0.05. However, the analysis did not find any significant difference in the degree of implementation of lean practices between small and mid-size companies. The p-values obtained were greater than 0.05, indicating that the observed differences were not statistically significant.

5.3. Impact of lean practices on performance

To address the second research question, survey participants were asked to rate the impact of implementing lean practices on both operational performance (e.g., reduction in delivery time, increased productivity, reduction in production cycle time) and environmental performance (e.g., reduced water consumption, food waste reduction, solid waste reduction) using a Likert scale ranging from 1 to 5. In this scale, 1 indicated no effect, 2 represented a minor effect, 3 denoted a neutral response, 4 indicated a moderate effect, and 5 signified a major effect. The results of this analysis can be found in Table 4 and reveal that the implementation of lean practices significantly affects several aspects of operational performance, such as reducing production costs, enhancing product quality, and increasing productivity. Additionally, the results show that lean practices substantially impact environmental performance, particularly in reducing water consumption, energy consumption cost, and solid waste.

Table 3: The results of Friedman test

Section	Chi-Square	df	Asymp.Sig.
S1	2.330	4	0.703
S2	8.348	3	0.023
S3	23.952	2	<0.01
S4	10.010	1	<0.01
S5	15.662	2	<0.01
S6	7.164	4	0.086
S7	5.128	2	0.37
S8	2.402	2	0.311
S9	22.611	2	<0.01
S10	6.211	3	0.063

Table 4: The performance of lean practices

Section	Code	Mean
Operational performance	Reduction of the production cycle time	3.82
	Reduction in delivery time	3.82
	Increasing productivity	3.96
	Reduction of customer complaints	3.79
	Production cost reduction	4
	Product Quality Improvement	4
	Scrap Rate Reduction	3.89
Environmental performance	Reduction of hazardous/toxic material consumption	3.5
	Minimization of atmospheric emissions (CO2 and other gases)	3.36
	Minimization of water consumption	3.82
	Solid waste reduction	3.68

Reduction of environmental accidents	3.61
Reducing CO2 emissions from transportation	3.36
Lowering the cost of energy consumption	3.79
Food Waste Reduction	3.5

6. Discussion

The significance of this study's findings lies in their evaluation and discussion specifically within the context of Moroccan agri-food companies. The results indicate that the majority of these companies have already adopted some form of lean practices. The most commonly implemented practices in this sector include those related to suppliers, customers, employees, and preventive maintenance. This aligns with the findings of [12], who studied the utilization of lean practices in European food SMEs, except that they observed a lower implementation of employee-related practices. However, this study also found a difference compared to [48], in Moroccan companies, the flow-related practices are more widely adopted, whereas European SMEs implement these practices less frequently. On the other hand, both this study and the results of [48] show that the pull practices have low implementation.

It appears that some lean practices associated with Pull are challenging to implement in food companies. The results suggest that this is largely due to the limited use of Kanban. The reason for this weakness could be attributed to the difficulty in predicting demand within the agri-food sector. The limited utilization of Kanban emerged as a prominent factor contributing to this difficulty, which can be linked to the inherent unpredictability of demand in the agri-food sector [56]. For instance, weather conditions greatly influence the demand for certain products, with a sunny afternoon witnessing a significant surge compared to a rainy afternoon [48].

Furthermore, the results indicate a low adoption of techniques aimed at reducing setup time. This finding can be explained by the intricate nature of production processes, involving diverse materials and compositions. It appears paradoxical that despite the inherent characteristics of the food sector, such as perishability and short production cycles [48], some lean practices face challenges in implementation within this industry.

This study demonstrates the substantial impact of lean practices on both environmental and operational performance, with notable effects such as reduced production costs, enhanced product quality, decreased water consumption, and lower energy consumption costs. One hypothesis could focus on the potential impact of lean practices on cost reduction and efficiency in Moroccan agri-food companies. It could be posited that by implementing lean principles such as waste reduction, just-in-time production, and continuous improvement, these companies can achieve significant cost savings and enhance overall operational performance. Other hypothesis could be proposed that by adopting lean methodologies, agri-food companies can enhance product quality, safety, and compliance with regulatory standards.

A key challenge lies in reconciling lean practices with the specific requirements of the agri-food sector, which deals with perishable goods, complex supply chains, and stringent food safety regulations. The active involvement of governmental agencies and industry associations in promoting lean practices can provide the necessary resources, knowledge-sharing platforms, and incentives to facilitate its adoption. However, certain barriers may impede the effective implementation of lean practices in Moroccan agri-food companies. These can include resistance to change, insufficient awareness or comprehension of lean principles, and inadequate investment in technology and infrastructure.

7. Conclusion

This study offers an overview of the growth of the agri-food industry, with a concentrate on identifying the most commonly implemented lean practices in Moroccan companies and how they can improve operational and environmental performance in the sector. However, it should be noted that the extent of this study is restricted to the lean practices of agri-food companies. Despite this limitation, the results provide valuable insights for company managers by giving them an understanding of the level of lean practice adoption in the Moroccan agribusiness sector and their effects on operational and environmental performance. This paper offers valuable practical implications for managers by presenting an overview of the present status of implementing lean manufacturing in food processing SMEs. While one limitation of this study is its focus on a limited sample of food companies in Morocco, it nonetheless serves as a foundation for future empirical research in this domain. Given the significance of the food sector in the forthcoming years, further exploration and investigation in this area are essential to enhance our understanding and application of lean practices in the industry. Future research could expand the study to cover agribusinesses in other developing countries, in order to generalize the findings. The results of this study have the potential to inform the creation of a tool to assess lean practices in the Moroccan agri-food companies.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

The authors would like to thank the anonymous reviewers and editor for their helpful suggestions that greatly improved the original manuscript.

References

[1] F.J. Krafick, “Triumph of the Lean Production System,” Sloan Management Review, **30**(1), 41–52, 1988.
[2] J.P. Womack, D.T. Jones, D. Roos, “The machine that changed the world,” Business Horizons, **35**(3), 81–82, 1992, doi:10.1016/0007-6813(92)90074-J.

[3] R. Shah, P.T. Ward, "Lean manufacturing: context, practice bundles, and performance," Journal of Operations Management, **21**(2), 129–149, 2003.

[4] F. Bouhannana, "Integrating Lean , Green and Agile in Supply Chain Management : A Systematic Literature Review Laboratory of Systems Engineering and Decision Making," 12–14, 2019.

[5] R. Mann, D. Kehoe, "An Evaluation of the Effects of Quality Improvement Activities on Business Performance," International Journal of Quality & Reliability Management, **11**(4), 29–44, 1994, doi:10.1108/02656719410057935.

[6] J. Bhamu and K. Singh Sangwan, "Lean manufacturing: literature review and research issues," *Int. J. Oper. Prod. Manag.*, **34**(7), 876–940, 2014, doi:10.3390/f9040186.

[7] F. Bouhannana and A. El Korchi, "A systematic literature review to measure lean, green and agile in manufacturing organisations," *Int. J. Oper. Res.*, **1**(1), 1, 2022, doi: 10.1504/IJOR.2022.10050966.

[8] V. Yadav, R. Jain, M.L. Mittal, A. Panwar, A. Lyons, "The impact of lean practices on the operational performance of SMEs in India," Industrial Management and Data Systems, **119**(2), 317–330, 2019, doi:10.1108/IMDS-02-2018-0088.

[9] R. Borges Lopes, F. Freitas, I. Sousa, "Application of lean manufacturing tools in the food and beverage industries," Journal of Technology Management & Innovation, **10**(3), 120–130, 2015.

[10] Boston Consulting Group, "Lean food-and-beverage manufacturing: Lower costs, better products, improved sustainability," 2015, doi:10.1109/ISBIM.2008.40.

[11] P. Centobelli, R. Cerchione, and R. Singh, "The impact of leanness and innovativeness on environmental and financial performance: Insights from Indian SMEs," *Int. J. Prod. Econ.*, 212, 111–124, 2019, doi:10.1016/j.jclepro.2013.10.015.

[12] M. Dora, M. Kumar, D. Van Goubergen, A. Molnar, X. Gellynck, "Food quality management system: Reviewing assessment strategies and a feasibility study for European food small and medium-sized enterprises," Food Control, **31**(2), 607–616, 2013.

[13] E. Psomas, J. Antony, N. Bouranta, "Assessing lean adoption in food SMEs: evidence from Greece," International Journal of Quality & Reliability Management, **35**(1), 64–81, 2018.

[14] A. Marques-Mendes, M.J. Santos, "Strategic CSR: an integrative model for analysis," Social Responsibility Journal, **12**(2), 363–381, 2016.

[15] T. Bonavia, J.A. Marin, "An empirical study of lean production in the ceramic tile industry in Spain," International Journal of Operations & Production Management, **26**(5), 505–531, 2006.

[16] T. Brush, A. Karnani, "Impact of plant size and focus on productivity: An empirical study," Management Science, **42**(7), 1065–1081, 1996.

[17] R. Shah, P.T. Ward, "Defining and developing measures of lean production," Journal of Operations Management, **25**(4), 785–805, 2007.

[18] A. Farissi, M. El Oumami, Z. Beidouri, "Assessing Lean Adoption in Food Companies : The Case of Morocco," **12**(May 2020), 5–14, 2021, doi:10.14716/ijtech.v12i1.3837.

[19] T. Bortolotti, S. Boscari, P. Danese, "Successful lean implementation: Organizational culture and soft lean practices," International Journal of Production Economics, **160**, 182–201, 2015.

[20] F. Bouhannana and A. Elkorchi, "Trade-offs among lean, green and agile concepts in supply chain management: Literature review," in *2020 IEEE 13th International Colloquium of Logistics and Supply Chain Management (LOGISTIQUA)*, IEEE, 2020, 1–5, doi: *10.1109/LOGISTIQUA49782.2020.9353930*

[21] M. Ghosh, "Lean manufacturing performance in Indian manufacturing plants," Journal of Manufacturing Technology Management, **24**(1), 113–122, 2013.

[22] S. Bhasin, "Prominent obstacles to lean," International Journal of Productivity and Performance Management, **61**(4), 403–425, 2012.

[23] A. Muñoz-villamizar, J. Santos, P. Grau, E. Viles, J. Santos, D. Sebastian, P. Grau, "Trends and gaps for integrating lean and green management in the agri-food sector," 2019, doi:10.1108/BFJ-06-2018-0359.

[24] P. Carlborg, D. Kindström, C. Kowalkowski, "A lean approach for service productivity improvements: synergy or oxymoron?," Managing Service Quality: An International Journal, **23**(4), 291–304, 2013.

[25] G. Onofrei, J. Prester, B. Fynes, P. Humphreys, F. Wiengarten, "The relationship between investments in lean practices and operational performance: Exploring the moderating effects of operational intellectual capital," International Journal of Operations & Production Management, **39**(3), 406–428, 2019.

[26] R.R. Fullerton, W.F. Wempe, "Lean manufacturing, non-financial performance measures, and financial performance," International Journal of Operations & Production Management, 2009.

[27] K.O. Cua, K.E. McKone-Sweet, R.G. Schroeder, "Improving performance through an integrated manufacturing program," Quality Management Journal, **13**(3), 45–60, 2006.

[28] S. Taj, L. Berro, "Application of constrained management and lean manufacturing in developing best practices for productivity improvement in an auto-assembly plant," International Journal of Productivity and Performance Management, **55**(3/4), 332–345, 2006.

[29] C. Droge, J. Jayaram, S.K. Vickery, "The effects of internal versus external integration practices on time-based performance and overall firm performance," Journal of Operations Management, **22**(6), 557–573, 2004.

[30] B.B. Flynn, E.J. Flynn, "An exploratory study of the nature of cumulative capabilities," Journal of Operations Management, **22**(5), 439–457, 2004.

[31] P. Nieminen, J. Takala, "Achieving better on-time-delivery performance with the help of internal dependencies in the production," International Journal of Management and Enterprise Development, **3**(1–2), 181–190, 2006.

[32] J.A. Garza-Reyes, V. Kumar, S. Chaikittisilp, K.H. Tan, "The effect of lean methods and tools on the environmental performance of manufacturing organisations," International Journal of Production Economics, **200**, 170–180, 2018.

[33] K. Zokaei and D. Simons, "Performance improvements through implementation of lean practices: a study of the UK red meat industry," *Int. Food Agribus. Manag. Rev.*, **9**(1030-2016–82410), 30–53, 2006, doi:10.1136/bmj.322.7278.98.

[34] M. Asmae, E.-N. Abdelali, and H. Brahim, "Critical success factors of Lean implementation in Moroccan agro food supply chain," in *2022 14th International Colloquium of Logistics and Supply Chain Management (LOGISTIQUA)*, May 2022, 1–6. doi: 10.1109/LOGISTIQUA55056.2022.9938121.

[35] H. ROCHDI and M. EL ATASSI, "Impact de l'intégration relationnelle inter-organisationnelle sur la performance logistique des chaines logistique d'exportation: Cas des CLE des primeurs de la région Sous-Massa," *Rev. Int. Cherch.*, **2**(2), 2021, doi:10.1007/s00170-012-3988-4.

[36] N. Bateman, "Sustainability: the elusive element of process improvement," *Int. J. Oper. Prod. Manag.*, 2005, doi:10.1016/j.ijpe.2018.09.031.

[37] M. Dora, M. Kumar, X. Gellynck, "Determinants and barriers to lean implementation in food-processing SMEs–a multiple case analysis," Production Planning & Control, **27**(1), 1–23, 2016.

[38] P. F. Larasati and N. B. Mulyono, "Exploring The Current State Of Lean Practices In The Agri-Food Supply Chain: A Systematic Literature Review," *Int. J. Appl. Bus. Res.*, 168–186, Jul. 2022, doi: 10.35313/ijabr.v4i2.269.

[39] D. Powell, S. Lundeby, L. Chabada, and H. Dreyer, "Lean Six Sigma and environmental sustainability: the case of a Norwegian dairy producer," *Int. J. Lean Six Sigma*, **8**(1), 53–64, 2017.

[40] B. Heymans, "Lean manufacturing and the food industry," *J. Technol. Manag. Innov.*, **10**(3), 120–30, 2015, doi:10.1007/978-3-642-40081-0.

[41] V. Sanchez Rodrigues, M. Kumar, "Synergies and misalignments in lean and green practices: a logistics industry perspective," Production Planning & Control, **30**(5–6), 369–384, 2019.

[42] B. Heymans, "Lean manufacturing and the food industry," *J. Technol. Manag. Innov.*, **10**(3), 120–30, 2015, doi:10.1007/s40092-019-0313-y.

[43] H. De Steur, J. Wesana, M. K. Dora, D. Pearce, and X. Gellynck, "Applying Value Stream Mapping to reduce food losses and wastes in supply chains: A systematic review," *Waste Manag.*, 58, 359–368, 2016, doi:10.1016/j.jclepro.2017.09.171.

[44] D. Pearce, M. Dora, J. Wesana, and X. Gellynck, "Toward sustainable primary production through the application of lean management in South African fruit horticulture," *J. Clean. Prod.*, 313, 127815, 2021.

[45] M. Melin and H. Barth, "Value stream mapping for sustainable change at a Swedish dairy farm," *Int. J. Environ. Waste Manag.*, **25**(1), 130–140, 2020, doi:10.1016/j.jclepro.2018.01.121.

[46] N. E. Caicedo Solano, G. A. García Llinás, and J. R. Montoya-Torres, "Towards the integration of lean principles and optimization for agricultural production systems: a conceptual review proposition," *J. Sci. Food Agric.*, **100**(2), 453–464, 2020, doi:10.1108/JHOM-09-2016-0165.

[47] N. V. K. Jasti, A. Sharma, and R. Kodali, "Lean to green supply chain management: a case study," *J. Environ. Res. Dev.*, **6**(3) 890–899, 2012, doi:10.1016/j.ijpe.2021.108060.

[48] M. Dora, D. van Goubergen, M. Kumar, A. Molnar, X. Gellynck, "Application of lean practices in small and medium-sized food enterprises," British Food Journal, **116**(1), 125–141, 2014, doi:10.1108/BFJ-05-2012-0107.

[49] U. Lehtinen and M. Torkko, "The lean concept in the food industry: A case study of contract a manufacturer," *J. Food Distrib. Res.*, **36**(856-2016–56436), 57–67, 2005, doi:10.1016/j.procir.2016.05.067.

[50] N. Upadhye, S. G. Deshmukh, and S. Garg, "Lean manufacturing in biscuit manufacturing plant: a case," *Int. J. Adv. Oper. Manag.*, **2**(1–2), 108–139, 2010.
[51] W.S. Al-Ghwayeen, A.B. Abdallah, "Green supply chain management and export performance: The mediating role of environmental performance," Journal of Manufacturing Technology Management, **29**(7), 1233–1252, 2018, doi:10.1108/JMTM-03-2018-0079.
[52] A. Al-Sheyadi, L. Muyldermans, K. Kauppi, "The complementarity of green supply chain management practices and the impact on environmental performance," Journal of Environmental Management, **242**, 186–198, 2019.
[53] M. Dieste, R. Panizzolo, "The effect of lean practices on environmental performance: An empirical study," Lean Engineering for Global Development, 225–258, 2019.
[54] M. Feng, W. Yu, X. Wang, C.Y. Wong, M. Xu, Z. Xiao, "Green supply chain management and financial performance: The mediating roles of operational and environmental performance," Business Strategy and the Environment, (April 2017), 1–14, 2018, doi:10.1002/bse.2033.
[55] S. Laari, J. Töyli, L. Ojala, "The effect of a competitive strategy and green supply chain management on the financial and environmental performance of logistics service providers," Business Strategy and the Environment, **27**(7), 872–883, 2018, doi:10.1002/bse.2038.
[56] M. Trujillo-gallego, W. Sarache, M.A. Sellitto, "Environmental performance in manufacturing companies : a benchmarking study performance," 2020, doi:10.1108/BIJ-05-2020-0225.
[57] A. Belhadi, F.E. Touriki, S. El, "Development of a Lean Assessment Tool for Small and Medium Sized-Enterprises," 361–369, doi:10.1007/978-3-319-58409-6.
[58] R. Shah and P. T. Ward, "Defining and developing measures of lean production," *J. Oper. Manag.*, **25**(4), 785–805, Jun. 2007, doi: 10.1016/j.jom.2007.01.019.
[59] B. Abdelilah, A. El Korchi, M. Amine Balambo, "Agility as a combination of lean and supply chain integration: how to achieve a better performance," International Journal of Logistics Research and Applications, 1–29, 2021.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 8, No. 4, 111-121 (2023)

www.astesj.com

Special Issue on Innovation in Computing, Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Colorized iVAT Images for Labeled Data

Elizabeth Dixon Hathaway[1,*], Richard Joseph Hathaway[2]

[1]*Department of Health & Human Performance, University of Tennessee at Chattanooga, Chattanooga, 37403, United States*

[2]*Department of Mathematical Sciences, Georgia Southern University, Statesboro, 30460, United States*

ARTICLE INFO

Article history:
Received:19 May, 2023
Accepted:19 July, 2023
Online: 26 August, 2023

Keywords:
Data visualization
Cluster heat map
iVAT image
Clustered tendency
Labeled data

ABSTRACT

A 2-dimensional numerical data set $X = \{x_1,\ldots,x_n\}$ with associated category labels $\{l_1,\ldots,l_n\}$ can be accurately represented in a 2-dimensional scatterplot where color is used to represent each datum's label. The colorized scatterplot indicates the presence or absence of spatial clusters in X and any special distribution of labels among those clusters. The same approach can be used for 3-dimensional data albeit with some additional difficulty, but it cannot be used for data sets of dimensions 4 or greater. For higher dimensional data, the improved Visual Assessment of cluster Tendency (iVAT) image can be used to indicate the presence or absence of cluster structure. In this paper we propose several new types of colorized iVAT images, which like the 2-dimensional colorized scatterplot, can be used to represent both spatial cluster structure and the distribution of labels among clusters.

1. Introduction

Let $O = \{o_1,\ldots,o_n\}$ be a set of n objects (e.g., people, trees). Suppose we have two types of information about the objects in O: (1) a set of corresponding *object data* $X = \{x_1,\ldots,x_n\} \subset \Re^s$, where for each k, datum x_k gives values for s different *features* (e.g., height, weight) of object o_k; and (2) a set of *labels* $L = \{l_1,\ldots,l_n\}$ indicating to which of several known *categories* each object belongs. The purpose of this paper is to introduce simple visual displays that allow us to compare the *cluster structure* in X with the *category structure* in L. Here we use the terms *cluster* or *spatial cluster* to refer to a group of close, similarly valued data points in X. Since we are dealing with visual displays, our eyes can be used to assess cluster structure. An interesting discussion of definitions of clusters from both the human and computer points of view is given in [1]. The potential uses of the new displays include: (1) checking to see if a cluster-based classification scheme (e.g., nearest centroid classification) might work well for labeling future objects; (2) providing a visual approach to feature selection for X by comparing displays using different subsets of features; (3) providing a simple visual approach for identifying clusters of interest, and (4) simply checking to see if the cluster structure of X aligns with the category structure of L.

In the case of s = 2 features we do not need a new approach as we can represent all the available information of object data X and labels L in a single, colorized scatterplot. We illustrate this below in Fig. 1 with a scatterplot of the hypothetical school data example from [2] involving 26 students (the objects) with their corresponding scaled SAT and high school GPA scores (the object data X) and their freshman math course outcomes of Pass or Fail (the labels or categories L). Note that the scatterplot reveals the clusters of the object data X along with the distribution of the labels among those clusters. Stated another way, we see that the colorized scatterplot allows us to check the *alignment* of the object data clusters with the label categories. Ignoring color, we can reasonably identify two spatial clusters: a lower left cluster (lower achievers) and an upper right cluster (higher achievers). Now also considering color, we see that the Fail category mostly aligns with the lower achievers while the Pass category mostly aligns with the upper achievers. We will refer to a datum x_k as a *minority point* if it is labeled differently from most of its nearest neighbors in X, and with this term we note that there are two minority points among lower achievers and one among higher achievers. The scatterplot makes it very clear which cluster should be targeted for additional learning support, which is the group of students who achieved less in high school.

A colorized scatterplot is all that is needed for labeled 2-dimensional data, but how do we display similar visual information for labeled data with s > 2? It is certainly possible to look at many colorized scatterplots of 2-dimensional slices of X but this can give misleading or at least incomplete information

* Corresponding Author: Elizabeth Dixon Hathaway, elizabeth-hathaway@utc.edu

about the actual situation when all s features are simultaneously considered. In 2 dimensions the simple (uncolored) scatterplot serves as the canvas to which we add color in order to display both object data and label information. What is the corresponding canvas for higher dimensional object data? The answer given here is the improved Visual Assessment of (cluster) Tendency (iVAT) image [3,4] or a similar image derived from it. In a single grayscale image iVAT is able to represent much of the cluster information for object data of any dimension, and so we choose it as a reasonable scatterplot substitute when s > 2. Color is added according to a variety of schemes given later. Presenting and testing these new Colorized iVAT (CiVAT) and related images is the purpose of this paper.

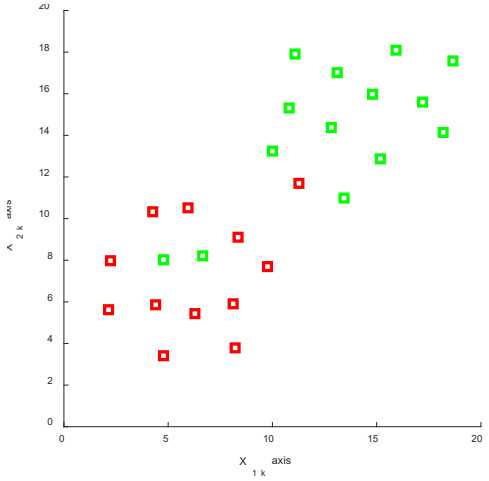


Figure 1. Green indicates Pass and Red indicates Fail.

We close this section by describing the organization of the remainder of this paper. Section 2 gives the necessary background regarding notation, terminology and iVAT images. The next section introduces several CiVAT and related approaches. Section 4 gives 10 examples using the new image-based techniques, seven of which use 2-dimensional object data. The point of using 2-dimensional examples is to provide training on the proper interpretation of the iVAT-based images by directly comparing the images with available colorized scatterplots; this direct comparison cannot be done when the object data dimension s ≥ 4. The final section gives some additional discussion and questions for future research.

2. Background

Our notation for objects O = {o_1,o_2,…,o_n}, object data X = {x_1,x_2,…,x_n} and category labels L = {l_1,l_2,…,l_n}were given in the last section. Additionally, the number of clusters is denoted by c and the number of label categories is denoted by d. In any particular example it is not necessarily true that the clusters of X align with the categories of L or even that c = d. An alternative to object data X that is very important in the following is called relational data, which is represented using an n × n matrix R, whose ijth entry R_{ij} = the relationship between objects o_i and o_j. With relational data, rather than describing objects directly (using the features of object data X), an object is described by saying how similar or dissimilar it is to each of the other objects. Relational data is natural in some cases such as when you try to describe the music of a particular musical band by saying how much it is like or unlike the music of other known bands.

In this paper the type of relationship described by R will always be dissimilarity between pairs of objects in O. Sometimes there may only be relational data R available and no object data X, but whenever there is object data, it is very easy to get corresponding relational data. In all of the following we will convert X to R using squared Euclidean distances:

$$R_{jk} = \text{dissimilarity}(o_j,o_k) = (x_{1j}-x_{1k})^2 + (x_{2j}-x_{2k})^2 + \ldots + (x_{sj}-x_{sk})^2, \text{ for } 1 \le j,k \le n \quad (1)$$

We mention that the much more difficult inverse problem of going from R to some object data set X is addressed using multidimensional scaling [5]. Note that the relational data matrices based on (1) will always satisfy for $1 \le j,k \le n$:

$$R_{jj} = 0 \quad (2a)$$

$$R_{jk} \ge 0 \quad (2b)$$

$$R_{jk} = R_{kj} \quad (2c)$$

Relational dissimilarity matrices are important to us because we can visually represent them using a single grayscale digital image, which leads us to a discussion of the original Visual Assessment of cluster Tendency (VAT) image from [6]. To represent a dissimilarity matrix R as a digital image, we interpret each matrix element R_{jk} as the gray level of the (j,k) pixel in the image. Sometimes we also refer to the actual pixel in row j column k of the image by R_{jk}. To make later colorization schemes simpler to describe we will assume that the dissimilarity matrix R has been scaled by dividing each element by the largest element of the original (unscaled) R, which means that the largest element in the new scaled R is 1 and the smallest is 0. In the representation of R as an image, the value 1 generates a white pixel and 0 generates a black one. Values in (0,1) generate various shades of gray. All images generated for this paper were done using the image display command **imagesc** in Matlab.

Does an image display of R in (1) give interpretable, visual information about the cluster structure in the object data X used to generate R? It turns out that the ordering of the data in X, equivalently the ordering of the rows and columns of R, is crucial to the answer to this question. The VAT procedure reorders the rows and columns of R to correspond to a reordering of the object data in X so that nearby data points are numbered consecutively (as nearly as possible). Precisely, we relabel the data according to the following scheme. The first reordered data point $x_{(1)}$ is taken to be a datum on the outer fringe of the data set X. Then the second (reordered) data point $x_{(2)}$ is chosen to be the remaining one closest (i.e., least dissimilar) to $x_{(1)}$. Then $x_{(3)}$ is chosen as the remaining point closest to any previously selected point (either $x_{(1)}$ or $x_{(2)}$). This is continued at each step by adding the next unchosen point that is closest to the set of previously chosen ones. In practice the reordering of the data of X is done virtually through the permutation of rows and columns of R, eventually producing the VAT reordered R*, as stated later in the VAT algorithm. The VAT procedure for reordering is related to Prim's Algorithm [7] for generating the minimal spanning tree of a graph, and this is discussed in [1].

We demonstrate the importance of ordering by revisiting the school data shown in Fig. 1 using figures taken from [2]. Fig. 2 gives a scatterplot of the data in its original, random ordering, along with the image for its corresponding relational data R, calculated using (1). In the scatterplot of Fig. 2(a), the first ordered point is represented by a red square and the rest of the ordering is represented using line segments connecting consecutively ordered data.

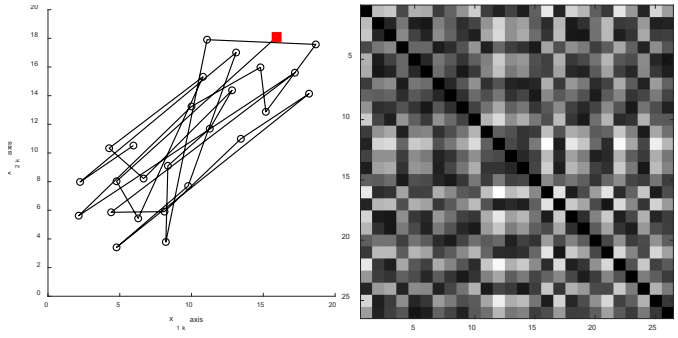


Figure 2: (a) Original X ordering starting with red square. (b) Image of R (uses the original X ordering).

Does the image in Fig. 2(b) usefully describe the cluster structure of the data? No. We cannot determine the number of clusters, the numbers of data in the clusters, the separation of the clusters, or even if there are any clusters. Now we will consider what happens if we reorder the data X (equivalently, the rows and columns of R) according to the VAT procedure. Fig. 3 gives the VAT reordered scatterplot for the data along with a display of the image corresponding to the VAT reordered R*. Notice in Fig. 3(a) that the first 13 data points are relatively close to each other as are the last 13 points. This closeness should generate darker pixels in the top left and bottom right 13 × 13 diagonal blocks of the image. Additionally, the relatively larger inter-cluster distances should produce lighter pixels in the other areas of the image. Fig. 3(b) confirms our expectations and gives representations of the two clusters as the two dark diagonal blocks. Appreciate that VAT produces this type of information directly from R, and there is no dependence on the actual dimensionality of the object data X.

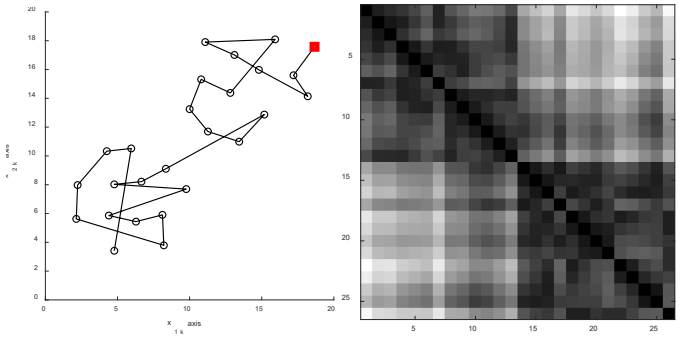


Figure 3: (a) VAT reordering starting with red square. (b) VAT image (uses the VAT reordered R*).

After seeing a few examples, it becomes easy to interpret VAT images in many cases. Blocks on the diagonal correspond to clusters. The number of clusters corresponds to the number of diagonal blocks. The numbers of objects in clusters corresponds to the sizes (i.e., the number of rows or columns of pixels) of the diagonal blocks. The degree of separation between clusters corresponds to the degree of grayscale contrast between diagonal blocks and off-diagonal block areas. When the clusters are compact and well separated, the VAT image shows a stark contrast of nearly black diagonal blocks with nearly white areas elsewhere. The VAT image in Fig. 2 is somewhat muddy in appearance because the clusters are not compact and well separated. The VAT algorithm follows.

Algorithm 1. VAT Reordering [6]

Input: R—n × n dissimilarity matrix

Step 1 Set J = {1, 2, …, n}, I = ∅; P(:) = (0, 0, …, 0) .

Step 2 Select $(i,j) \in \underset{p\in J,\ q\in J}{\arg\max} \{R_{pq}\}$.

Set P(1) = i; I = {i}; and replace J ← J – {i} .

Step 3 For r = 2, …, n:

Select $(i,j) \in \underset{p\in I,\ q\in J}{\arg\min} \{R_{pq}\}$.

Set P(r)=j; replace I←I∪{j} and J←J–{j}

next r.

Step 4 Obtain the VAT reordered dissimilarity matrix R^* using the ordering array P as $R^*_{ij} = R_{P(i)P(j)}$ for $1 \le i, j \le n$.

It is possible to use the VAT image as the basis for colorized versions but it turns out that there is an enhanced version, known as improved VAT (iVAT), which was first proposed in [3] and is based on replacing the original dissimilarities with the set of derived minimax distances. It is useful to understand how the minimax data are defined. Think of a complete graph with vertices $\{v_1,\ldots v_n\}$ and interpret R as giving edge lengths (i.e., distances) between each pair of vertices; that is R_{jk} is the direct distance, i.e., edge length, from v_j to v_k. What is the minimax distance from v_j to v_k? Let p be any path from v_j to v_k and let e_p be the maximum edge length (from R) used in path p. Then the minimax distance between vertices v_j to v_k is defined to be the minimum value of e_p over all possible paths p from v_j to v_k, and this value is taken to be the new minimax dissimilarity (or distance) between x_j and x_k.

There are predictable effects resulting from the switch to minimax data. First, all points within a cluster become nearer and more equidistant to each other, and this often provides more uniformity of tone in the diagonal blocks. Second, many or all points within one cluster share a common minimax distance to points in another cluster, which also gives off-diagonal block areas a more uniform appearance. Usually there is more observable contrast between diagonal blocks and off-diagonal areas using minimax data. Lastly, it is important to note that while both Euclidean and minimax distances work great for separated hyperspherical clusters, it is true that minimax distances can better handle cases of chainlike and stringy clusters by keeping intra-cluster pairwise distances more uniformly small. Fig. 4 gives the iVAT image corresponding to the school data.

Notice the increased uniformity of both diagonal and off-diagonal blocks. Note also the high degree of contrast. The **imagesc** command in Matlab does a virtual rescaling of the matrix range of values to [0,1] so that every VAT or iVAT image produced by it will include black pixels on the diagonal and at least one pure white pixel off the diagonal. Comparing Fig. 4 to Fig. 3(b) we see the usual situation; iVAT images provide cleaner and clearer descriptions of cluster structure than VAT images. For this reason, we choose to use iVAT as the basis for our higher dimensional substitute for scatterplots; it will represent the cluster

structure and added color will represent the label category information.

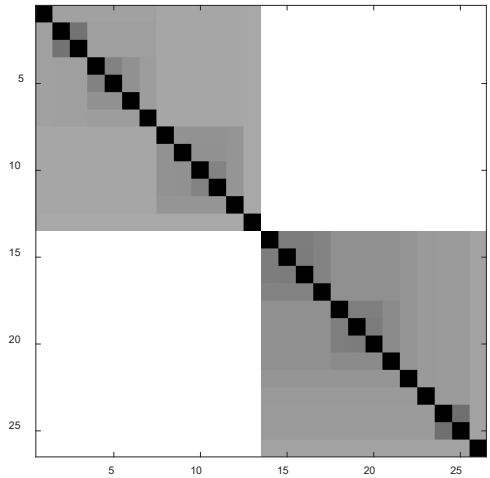


Figure 4: iVAT image (uses the VAT reordered minimax distances).

The originally proposed iVAT from [3] is done in a relatively inefficient manner. It is shown in [4] that greater efficiency (i.e., an order of magnitude of complexity) can be gained by first applying VAT to the original relational data R to get the VAT reordered R*, and then finding the minimax form of R* which we denote throughout the following by R'*. The special properties of a VAT reordered R* allow certain opportunities to gain efficiency in calculating the minimax distances. To be clear, the efficient technique in [4] produces an equivalent result to that obtained by finding the minimax version of R and then applying VAT. The efficient approach for minimax distance calculation is given next in Algorithm 2.

Algorithm 2. Efficient iVAT Calculation of Minimax Dissimilarities [4]

Input: R^*—VAT reordered n × n dissimilarity matrix
Step 1 Initialize minimax dissimilarity matrix $R'^* = [0]^{n\times n}$
Step 2 For r = 2, …, n

$$j = \underset{k=1,\ldots,r-1}{\arg\min} \{R^*_{rk}\}$$
$$R'^*_{rc} = R^*_{rc}\ , c = j$$
$$R'^*_{rc} = \max\ \{R^*_{rj}\ , R'^*_{jc}\ \}, c = 1, \ldots, r\text{-}1, c \neq j$$

next r
Step 3 R'^* is symmetric, thus $R'^*_{rc} = R'^*_{cr}$

To calculate an iVAT image for an original set of randomly ordered relational data R, we first apply Algorithm 1 to transform R to the VAT reordered R*, and then apply Algorithm 2 to R* to get the final iVAT matrix R'*. The display of R'* gives a nice image as in Fig. 4. A thorough and recent survey of VAT-based visualization schemes is given in [8], and discussion relating VAT to clustering methodology can be found in [1]. The VAT and iVAT images are part of a much larger set of visualization methodology known as cluster heat maps. A useful partial survey of heat maps through 2009 is found in [9]. In the next section we put color to the canvas.

3. Ordering and Colorization Schemes

In this section we introduce four schemes for adding color to the iVAT image R'*: two that use exactly the iVAT matrix R'* and two that use a label-reordered version of it. Note that the two using a label-reordered version still require application of iVAT in order to efficiently get the minimax distances which give us the clearest pictures.

We start by using an example to describe exactly what is meant by the Label Reordering (LR) of iVAT. In our example suppose that R'* is produced by iVAT and that the column and row orderings of R'* correspond to the object ordering

$$o_3, o_8, o_2, o_5, \mathbf{o_4}, o_1, o_7, o_6. \quad (3)$$

Also suppose that the corresponding label array for this iVAT ordering of objects is L* = [2, 1, 1, 2, **1**, 2, 1, 1]. We can readily see in this example that there are 8 objects and 2 categories and can easily determine the category of each object. For example, since o_4 is the fifth object in the iVAT ordering of (3) and L*(5) = 1, we know that o_4 is in category 1. What exactly is the LR for this example? It is the ordering obtained by first pulling out all the category 1 objects from (3), going from left to right, and then concatenating to it the list of all category 2 objects, going from left to right. Doing this for the example gives the LR order of objects

$$o_8, o_2, o_4, o_7, o_6, o_3, o_5, o_1. \quad (4)$$

In doing the LR reordering the objects are never actually referenced. Instead, L* is used to find the permutation that would take (3) to (4) and then the permutation is applied directly to L* and the rows and columns of R'*. Note that the LR permuted form of L* is simply [1, 1, 1, 1, 1, 2, 2, 2]. The process is continued in the obvious way if there are more than 2 categories.

So the two ordering schemes used later are iVAT and LR. Before stating the two colorization schemes we give a little background for manipulating digital images in Matlab. Suppose that R is a 2-dimensional matrix representing a monochrome image that we wish to colorize. We first need to scale R and then convert it to red, green, blue (RGB) format. This can be done in Matlab using the pair of statements **R = R ./ max(max(R))** and **R = cat(3,R,R,R)**. Execution of these two statements converts R to a 3-dimensional structure where $R_{i,j,1}$, $R_{i,j,2}$ and $R_{i,j,3}$ respectively give numerical values for the red, green and blue components of pixel (i,j). It is helpful to scale R immediately before conversion to RGB so that changes to the color components done by the colorization schemes will have consistent, predictable impact. In our implementations we use the following associations in Table 1 between category (or label) number, color and RGB values:

Table 1: Category-Color Associations

Category Number	Name of Associated Color	[R,G,B] Values
1	red	[1,0,0]
2	green	[0,1,0]
3	blue	[0,0,1]
4	cyan	[1,1,0]
5	magenta	[1,0,1]
6	yellow	[0,1,1]
7 or more	black	[0,0,0]

This color scheme is implemented and referenced in the algorithms using the label color matrix C given by

$$C = \begin{bmatrix} 1 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 & 0 \end{bmatrix}^T \quad (5)$$

Sometimes we will refer to R_{ij} as the matrix or structure entry and sometimes as the corresponding pixel. Also, the notation "**:**" should be understand (as in Matlab) to indicate free range of that parameter. In other words, C(2,:) refers to row 2 of C. As an example of using C, the color for category 3 has [R,G,B] values C(3, :) = [0, 0, 1] which is blue. More colors can easily be included by adding rows to C. The choice of C in (5) suffices for us as we do not have examples with more than 7 different label categories.

We now describe the two colorization schemes. The first type of colorized image is said to be Diagonally Colorized (DC), which when paired with the two ordering schemes gives the two images DCiVAT and DCLR. Here each diagonal pixel, is assigned the label color of its corresponding object. For an example, let R** denote the result of scaling and then converting to RGB format either the iVAT matrix R'* or the LR version of it. If the third row (and column) of the matrix R** corresponds to an object that is in category 2, then the (3,3) pixel is made (pure) green by assigning the RGB values of C(3,:) = [0, 1, 0] to $R^{**}_{3,3}$. This is a simple approach that gives clear information if the number of data is very small, but a single diagonal pixel may be very difficult or impossible to see if the number of data is large. For large data sets, we include the option of coloring *B* additional bands of elements to the immediate right and underneath the diagonal pixels; in the previous example that would mean to not just colorize $R^{**}_{3,3}$ but to also change RGB color values of $R^{**}_{3,4}$, $R^{**}_{3,5}$, …, $R^{**}_{3,\min(n,3+B)}$ and $R^{**}_{4,3}$, $R^{**}_{5,3}$, …, $R^{**}_{\min(n,3+B),3}$ to all match $R^{**}_{3,3}$. This thickened diagonal band will be visible for images of any size, as long as the number of additional bands B is approximately equal to n/50 or larger. Examples in the next section will give the reader a feel for how many additional bands are needed. Algorithm 3 which gives the first pair of our colorized images, DCiVAT and DCLR, is given next.

Algorithm 3. Diagonal Colorization: DCiVAT and DCLR

Input: R—n × n dissimilarity matrix
L—1 × n array where L(j) = category label of o_j
T—string giving order type to use: iVAT or LR
B—scalar in {0,1,2,…,~n/25} denoting the number of additional color bands
C—c × 3 matrix giving label color RGB values

Step 1 Apply Algorithm 1 (VAT) to R to get R* and 1 × n permutation array P

Step 2 Obtain the reordered label array L* using the permutation array P by L*(j) = L*(P(j)) for $1 \le j \le n$.

Step 3 Apply Algorithm 2 (iVAT) to R* to get R'*

Step 4 If T = iVAT, then do not alter R'*
If T = LR, then overwrite R'* by its LR version and L* by its LR version

Step 5 Scale R'* to have all elements in [0,1] and convert the scaled matrix R'* to RGB color format structure R**

Step 6 For r = 1, …, n
Change RGB color values of diagonal pixel R^{**}_{rr} to C(L*(r), :)
next r

Step 7 If B > 0, Change RGB color values of B pixels to right and below to match the new color of the diagonal pixel.
For r = 1,…,n
Change RGB color values of $R^{**}_{r,r+1}$, $R^{**}_{r,r+2}$, …, $R^{**}_{r,\min(n,r+B)}$ and $R^{**}_{r+1,r}$, $R^{**}_{r+2,r}$, …, $R^{**}_{\min(n,r+B),r}$ to all match C(L*(r), :)
next r

The second type of colorized image is said to be Block Colorized (BC), which when paired with the two ordering schemes gives BCiVAT and BCLR. In the block-colorization approach we color a pixel R^{**}_{rt} whenever it is intracategory, i.e., the row and column objects corresponding to that pixel share the same category. This procedure will colorize many pixels compared to the diagonal-colorization approach and if care is not taken then cluster information depicted by various shading patterns could be destroyed by overly intense off-diagonal colorization. Rather than make complete changes to the effected pixels, we redefine them as the average of the old RGB values and the new category RGB ones; i.e., we change the RGB color values of intra-category pixels R^{**}_{rt} to equal the average of (old) R^{**}_{rt} and C(L*(r), :). This has an analogous effect to using stain and not paint on a piece of wood; we get the advantage of color but can still see the grain underneath the color. Algorithm 4 is given next which gives the last two of our colorized images, BCiVAT and BCLR.

Algorithm 4. Block Colorization: BCiVAT and BCLR

Input: R—n × n dissimilarity matrix
L—1 × n array where L(j) = category label of o_j
T—string giving order type to use: iVAT or LR
C—c × 3 matrix giving label color RGB values

Step 1 Apply Algorithm 1 (VAT) to R to get R* and 1 × n permutation array P

Step 2 Obtain the reordered label array L* using the permutation array P by L*(j) = L(P(j)) for $1 \le j \le n$.

Step 3 Apply Algorithm 2 (iVAT) to R* to get R'*

Step 4 If T = iVAT, then do not alter R'*
If T = LR, then overwrite R'* by its LR version and L* by its LR version

Step 5 Scale R'* to have all elements in [0,1] and convert the scaled matrix R'* to RGB color format structure R**

Step 6 Change RGB color values of intra-category pixels
For all (r,t) ∈ {1,…,n} × {1,…,n} with L*(r) = L*(t)
Change the RGB color values of R^{**}_{rt} to equal the average of (old) R^{**}_{rt} and C(L*(r), :)

In the next section we report on various experiments involving the four approaches. Which type of colorization is better, DC or BC? Which type of ordering is better, iVAT or LR? Does any combination of colorization and ordering produce a useful higher dimensional analogue to the 2-dimensional colorized scatterplot?

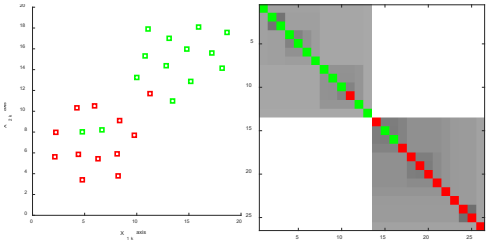
4. Experiments

In the first part of this section, we give all four new images for each in a set of seven 2-dimensional examples. The 2-dimensional examples allow us to compare the images directly to colorized scatterplots, and this will help us achieve two goals: (1) develop some skill in properly interpreting the new colorized images, and

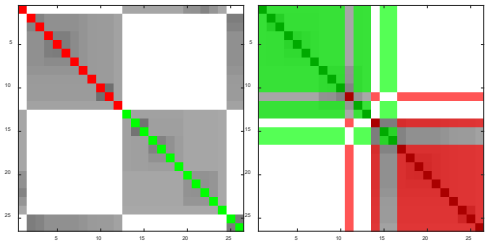
(2) determine if some of the new colorized approaches work better than others. Note that there is always color consistency between scatterplots and colorized images; i.e., scatterplot points of a certain color correspond to image pixels of the same color.

Example 1. Hypothetical School Data from Section 1.

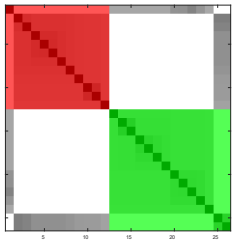
Fig. 5 repeats the colorized scatterplot for the School Data along with its four corresponding new images. Note that DCiVAT in Fig. 5(b) most clearly represents the situation of the single minority point in the green cluster and the two minority points in the red cluster. BCiVAT in 5(d) gives the same information but the colorized strips outside of the main diagonal blocks are a little distracting, at least initially. The DCLR and BCLR images in Fig. 5(c) and 5(e), respectively, disrupt the main diagonal block structure of the iVAT image by reordering according to category, but these images still convey the information of the scatterplot. Note how the off-diagonal-block patches of gray are essential for interpretation; for example, the vertical off-diagonal gray bars in Fig. 5(e) indicate that the last two green points are near (or in) the red cluster. No extra diagonal bands are needed (i.e., B = 0) for DCiVAT and DCLR because of the small sample size n = 26. With each example, notice which type of image most transparently and naturally conveys the essential cluster and category information in the scatterplot.



(a) Scatterplot (b) DCiVAT



(c) DCLR (d) BCiVAT

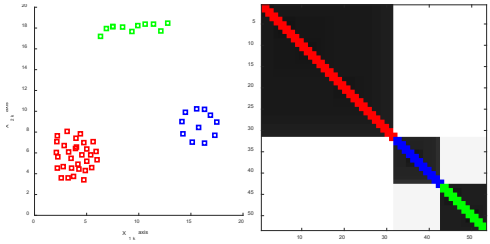


(e) BCLR

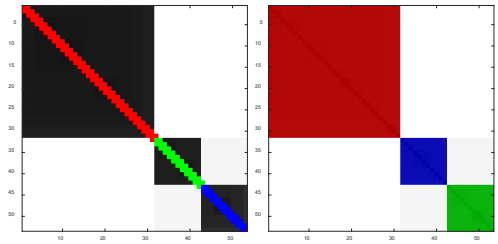
Figure 5: Colorized scatterplot and four images; sample size n = 26; extra diagonal bands B = 0 for DCiVAT and DCLR.

Example 2. Data with Perfect Alignment between Spatial Clusters and Label Categories.

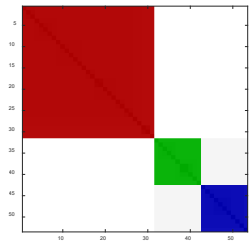
Fig. 6 gives the scatterplot and corresponding images. This example covers the ideal case where the clusters are well separated and perfectly aligned with the label categories. In this simple case each approach perfectly captures the essential information in the scatterplot. Each image shows the presence of 3 well-separated clusters, each filled with points of a single label category. Diagonal blocks, each of single color with light off-block areas, imply alignment between clusters and categories is total. Relative block sizes indicate the red cluster has about 3 times as many points as the blue and green clusters. The iVAT-produced minimax distances allow the stringy clusters to be handled just as naturally as the cloud clusters. The diagonals of DCiVAT and DCLR are slightly thickened for the sample size n = 53 using B = 1 extra bands of pixel coloration.



(a) Scatterplot (b) DCiVAT



(c) DCLR (d) BCiVAT



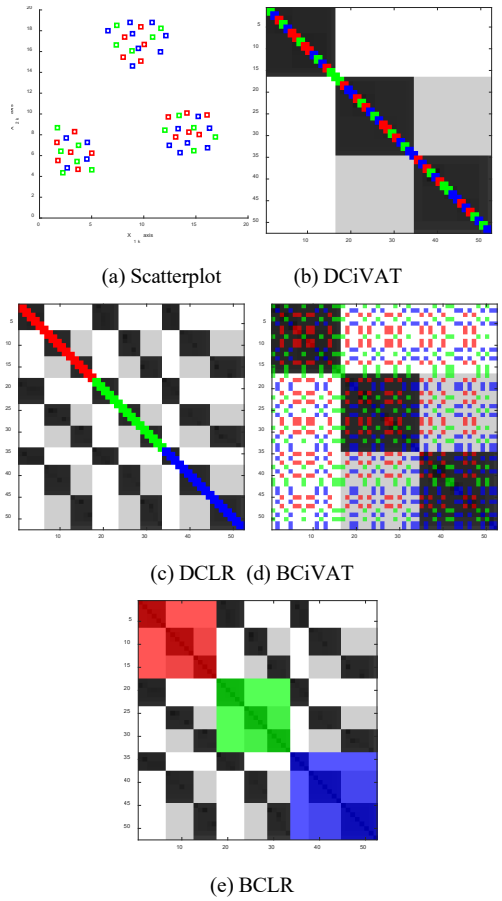
(e) BCLR

Figure 6: Colorized scatterplot and four images; sample size n = 53; extra diagonal bands B = 1 for DCiVAT and DCLR.

Example 3. Data with No Alignment between Spatial Clusters and Label Categories.

Fig. 7 gives the scatterplot and corresponding images. The scatterplot in Fig. 7(a) shows 3 well-separated spatial clusters having virtually no alignment between clusters and label categories. Once again, the DCiVAT image gives all the information in a clear fashion. It shows there are three spatial clusters and that each cluster has a mixture of points from each of the three label categories. The BCiVAT image is similarly informative but the colorful off-diagonal pixels are somewhat distracting. Note that the reordering of DCLR and BCLR has seriously disguised the spatial cluster information, and to reconstruct it you have to notice that the red objects appear in 3

small clusters, as do each of the green and blue objects. This could all be explained by having 3 mixed clusters as in the scatterplot but having to do this type of reconstruction is awkward at best.



(a) Scatterplot (b) DCiVAT

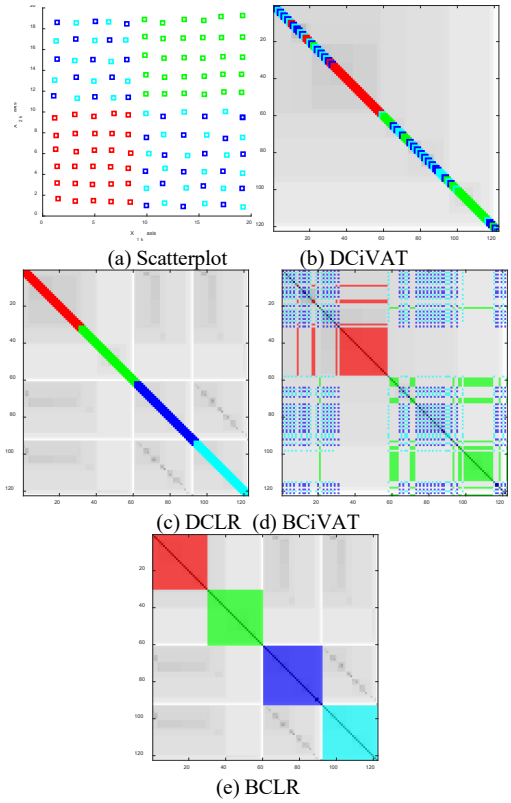
(c) DCLR (d) BCiVAT

(e) BCLR

Figure 7: Colorized scatterplot and four images; sample size n = 52; extra diagonal bands B = 1 for DCiVAT and DCLR.

Example 4. Data with No Spatial Clusters.

Fig. 8 gives the scatterplot and corresponding images. In this example we have a data set with virtually no cluster structure. In each of the images, a uniform gray tint and an absence of dark diagonal blocks indicate a lack of spatial clusters. The scatterplot contains additional information, though. There is some segregation of red and green points, while the blue and cyan ones are mixed together. Do any of the images in Fig. 8 represent this? The DCLR and BCLR do not give the color organization information in an easily interpreted form. No doubt the ghost diagonals in the bottom right-hand quarter of the DCLR and BCLR images contain some of this information but it is difficult to interpret. On the other hand, the iVAT ordering in DCiVAT and BCiVAT shows extensive intermingling of blue and cyan, along with some segregated groups of red and green. It is not a perfect representation of the scatterplot but it is consistent with what is shown there and it is instantly interpretable. Lastly, note the larger sample size n = 122 gives a need for B = 3 extra bands to widen the colorized diagonals in DCiVAT and DCLR.



(a) Scatterplot (b) DCiVAT

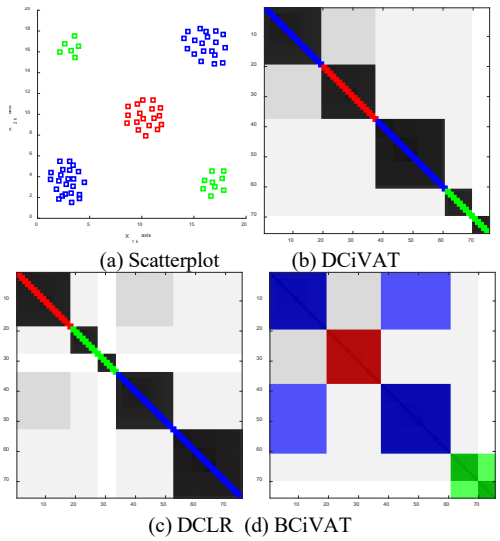
(c) DCLR (d) BCiVAT

(e) BCLR

Figure 8: Colorized scatterplot and four images; sample size n = 122; extra diagonal bands B = 3 for DCiVAT and DCLR.

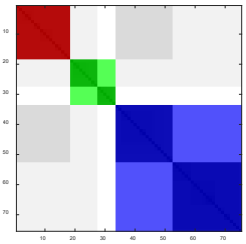
Example 5. Data with More Spatial Clusters than Label Categories.

Fig. 9 gives the scatterplot and corresponding images. This example is quite easy for all four approaches. The scatterplot shows 5 well-separated clusters: 1 red cluster, 2 blue clusters and 2 green clusters. This information is indicated most clearly by DCiVAT and DCLR. The BCiVAT and BCLR images are also consistent with the scatterplot and easily understood, but slightly more awkward because of the lighter blue and green off-diagonal blocks. Notice that the blue diagonal blocks in DCiVAT are not adjacent while the green ones are, and this shows that adjacency of blocks does not necessarily imply nearness of corresponding spatial clusters.



(a) Scatterplot (b) DCiVAT

(c) DCLR (d) BCiVAT

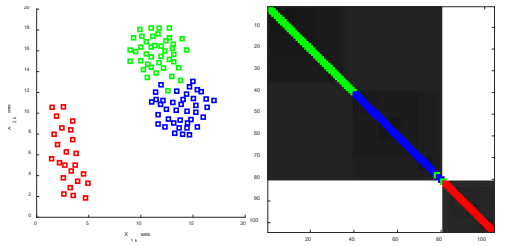


(e) BCLR

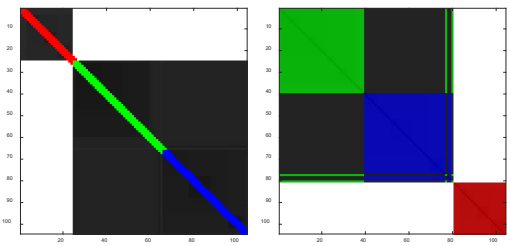
Figure 9: Colorized scatterplot and four images; sample size n = 75; extra diagonal bands B = 1 for DCiVAT and DCLR.

Example 6. Data with Touching Spatial Clusters #1.

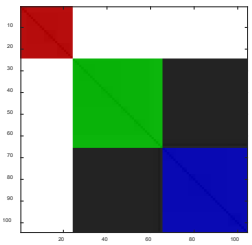
Fig. 10 gives the scatterplot and corresponding images. Several of the previous examples have dealt with extreme cases but we include one here with a more likely property: the existence of spatial clusters that are not well separated. The scatterplot in Fig. 10(a) shows a data set with a red, green and (mostly) blue cluster. The green and blue clusters are touching and could arguably be considered a single cluster. Notice what the images indicate for this example. At first glance, DCLR and BCLR seem to do a good job, indicating the presence of 3 clusters, each cluster populated by points of a single label category. The results for DCiVAT and BCiVAT are essentially the same but with the additional information that a couple of the green points may be invasive to the blue cluster. In each image the large dark block containing the blue and green areas indicate that those two clusters are very near each other and could arguably be considered a single cluster.



(a) Scatterplot (b) DCiVAT



(c) DCLR (d) BCiVAT

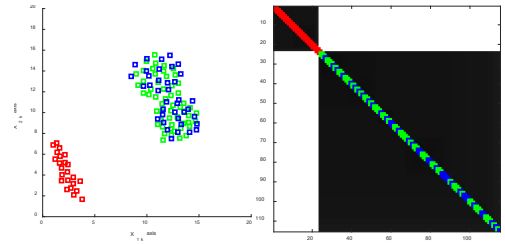


(e) BCLR

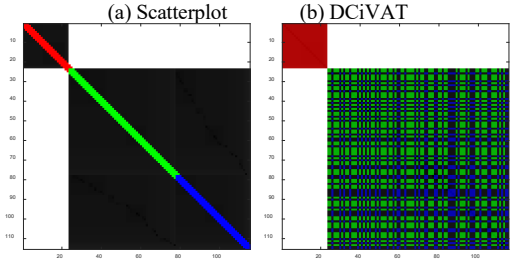
Figure 10: Colorized scatterplot and four images; sample size n = 104; extra diagonal bands B = 2 for DCiVAT and DCLR.

Example 7. Data with Touching Spatial Clusters #2.

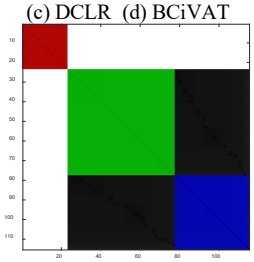
Fig. 11 gives the scatterplot and corresponding images. First note the spatial similarity of this and the previous example that is evident in a comparison of the Fig. 10(a) and Fig. 11(a) scatterplots. The real difference is all about the labeling of the points in the touching clusters. Notice how the DCiVAT and BCiVAT images in Fig. 11 perfectly represent the situation; there is a distant cluster of red points and a single cluster of points with commingled green and blue labels. On the other hand, DCLR and BCLR represent this and the data from Example 6 in an essentially identical manner, and we view this as a serious fault of the label reordering schemes.



(a) Scatterplot (b) DCiVAT



(c) DCLR (d) BCiVAT



(e) BCLR

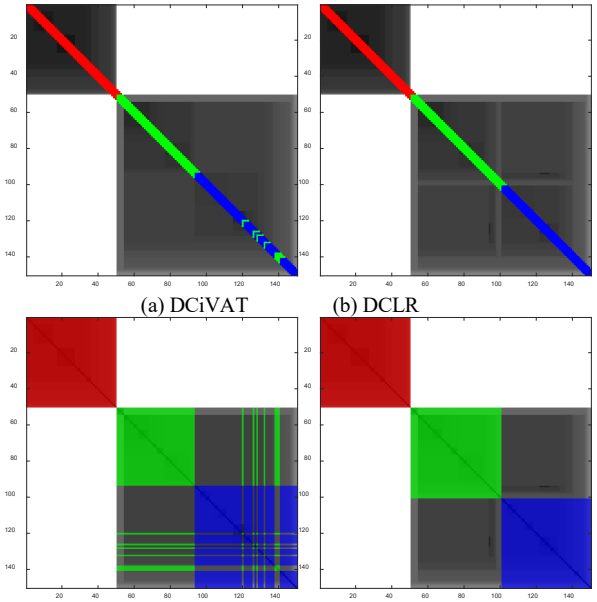
Figure 11: Colorized scatterplot and four images; sample size n = 115; extra diagonal bands B = 2 for DCiVAT and DCLR.

In the second group of examples we will apply all four approaches to higher-dimensional data sets taken from the UCI Machine Learning Repository at https://archive.ics.uci.edu/ml/index.php. The data sets used for the experiments below can be found by searching there for 'iris', 'divorce' and 'seeds'. For each data set we only give the four new images since no scatterplot is possible.

Example 8. Iris Data [10].

Fig. 12 gives the corresponding images. This data is considered to be real-valued continuous. Fisher's Iris Data [10] has probably been used more often in demonstrating clustering and classification schemes than any other data set. In this example we use the "bezdekIris.data" version of the data set from UCI. It consists of 150 4-variate object data describing the Setosa (red labels), Versicolor (green labels) and Virginica (blue labels) subspecies of iris. The data set contains 50 observations for each

of the 3 subspecies, with each observation giving a measurement of sepal length, sepal width, petal length and petal width. Various colorized scatterplots of 2-dimensional slices of the Iris Data are available in the literature (e.g., page 82 of [1]) and the images in Fig. 12 are consistent with what is known about it. The Setosa (red) cluster is well separated from the other two overlapping clusters. Note that the DCiVAT and BCiVAT images indicate the possible intrusion of the Virginica cluster by Versicolor data.



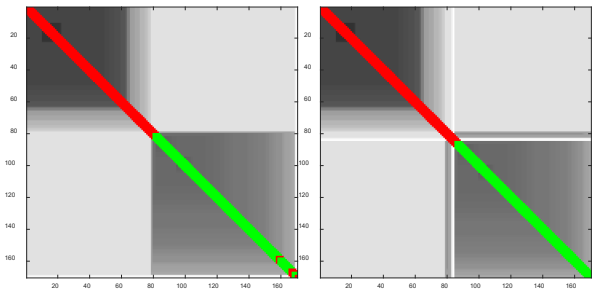
(a) DCiVAT (b) DCLR

(c) BCiVAT (d) BCLR

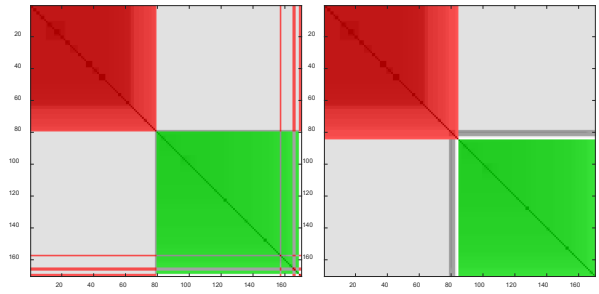
Figure 12: Four images for 4-variate Iris Data; sample size n = 150; extra diagonal bands B = 3 for DCiVAT and DCLR

Example 9. Divorce Data [11].

Fig. 13 gives the corresponding images. This data is considered to be integer valued and is from [11]. It consists of 170 54-variate data from 84 divorced (red label) and 86 married (green label) couples. Each person interviewed gave answers of the form 0 (never), 1 (seldom), 2 (average), 3 (frequently), 4 (always) to each of 54 questions such as "I feel right in our discussions" and "I wouldn't hesitate to tell her about my wife's inadequacy". All images indicate fairly separated clusters that are well aligned with the categories. There may be a bit of a trail of red transition points leading from the core of the red cluster over to the green cluster. The DCiVAT and BCiVAT images also indicate possible red intrusion into the green cluster. Appreciate that we have done something analogous to a colorized scatterplot even though we have 54-dimensional data!



(a) DCiVAT (b) DCLR

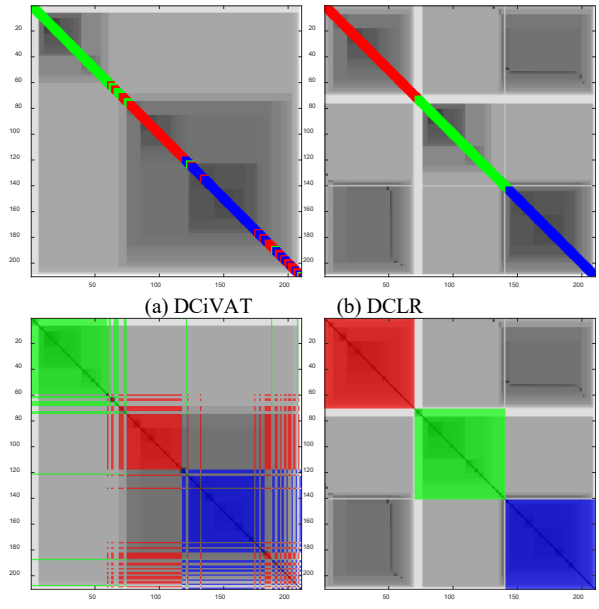


(c) BCiVAT (d) BCLR

Figure 13: Four images for Divorce Data; sample size n = 170; extra diagonal bands B = 4 for DCiVAT and DCLR.

Example 10. Seeds Data [12].

Fig. 14 gives the corresponding images for our final example. The real-valued, continuous Seeds Data is from [12] and consists of 210 7-variate observations of 3 varieties of wheat: Kama (red label), Rosa (green label), and Canadian (blue label). The data is divided equally among the 3 categories. Each datum consists of measurements of 7 geometric parameters of wheat kernels: area, perimeter, compactness, length, width, asymmetry coefficient and length of kernel groove. All of these measurements are obtained using a high-quality imaging of the kernel done by a soft X-ray technique. The DCiVAT and BCiVAT images indicate the presence of poorly separated clusters. There appears to be a green labeled cluster and a large red-blue one. The red-blue one has a red core part, a blue core part, and a mixed part. The DCLR and BCLR images may at first seem to indicate 3 cleanly aligned spatial clusters but the dark off-diagonal block is a reminder that there is considerable red and blue intermingling. This example is probably typical of many; it provides an imperfect but useful glimpse of some of the information that we seek.



(a) DCiVAT (b) DCLR

(c) BCiVAT (d) BCLR

Figure 14: Four images for Seeds Data; sample size n = 210; extra diagonal bands B = 5 for DCiVAT and DCLR.

5. Discussion

In this paper we presented four colorized images that simultaneously display cluster and category information for

labeled object data sets of any dimension. These four approaches are obtained by combining a colorizing scheme (Diagonalized or Block) with an ordering scheme (iVAT or LR). Having tested all approaches on the diverse set of examples in Section 4, we can now assess DCiVAT, DCLR, BCiVAT and BCLR. We believe that the superior method among the four is DCiVAT. In every 2-dimensional experiment it produced an accurate reflection of the corresponding scatterplot in a way that is most easily understood. We will say more about DCiVAT after we give specific criticisms of the other three approaches.

The label reordered approaches DCLR and BCLR may only give good results in the simplest possible case, when there are well-separated clusters with (nearly) perfect alignment between clusters and label categories. This situation is demonstrated in Example 2 where all colorization schemes did equally well. We found no situation where LR ordering works better than iVAT ordering and some where it absolutely fails, such as that seen by comparing Examples 6 and 7; in those examples DCLR and BCLR gave nearly identical results for two very different situations. In other examples LR ordering preserves, but scrambles, the cluster information so that visual interpretation of the images is [complicated; Example 3 gives a clear case of such cluster obfuscation. To summarize, LR is never better, sometimes flat wrong and at other times unnecessarily awkward. Additionally, the process of first doing iVAT and then undoing that ordering to get LR adds computational overhead above that of DCiVAT and BCiVAT.

In comparing DCiVAT to BCiVAT we know the ordering is the same and so it is only about the relative merits of the two different colorization schemes. Preference between these is admittedly subjective. There are cases where a BCiVAT image is very clear in its depiction of the data set, as in Example 7, and other cases where there is enough off-diagonal action to be distracting, as in Examples 3 and 5. It is possible that adjustments to BCiVAT that lighten the off-diagonal part of the image might very well improve its readability, but will it then produce something clearer than DCiVAT? We do not think so. Our conclusion is that BCiVAT is good, but DCiVAT is also good and never produces visually awkward results. DCiVAT is also the most straightforward of the four schemes to describe or program. Simply do iVAT and then colorize the diagonal pixel to match the category labels. Adding extra bands of colorization is easily done in the routine by adding one double loop. While our Matlab routines were written to allow the user to specify the number of additional colorization bands B, this choice could be automated in the DCiVAT program by using something like **B = floor(n/25)**. Python users can get the core VAT and iVAT routines written by Ismaïl Lachheb at https://pypi.org/project/pyclustertend/ .

This paper is concluded with a few ideas about possible future research involving DCiVAT or related colorization schemes. The first opportunity may be to colorize other VAT-type approaches. Many VAT type procedures have been devised since the original paper [6], and [8] gives a thorough survey of most of them with a total of 184 references. For example, there are various big data variants of VAT beginning with [13] and continuing through more recent work as in [14]. There are VAT-based procedures for rectangular (i.e., not square) relational data matrices [15,16] and even streaming data [17]. Are there opportunities in this vast set of methodology to add useful colorization, for depicting either label information or something else? Color is effectively used in [18] to draw attention to parts of 3-dimensional scatterplots. Are there more opportunities to usefully colorize? Might a block colorization scheme be more natural than diagonal colorization in some cases, such as for rectangular data matrices?

Another opportunity may be to refine our ability to interpret iVAT and colorized iVAT images. Did the examples in Section 4 teach us everything about DCiVAT interpretation or would a more extensive and systematic study yield more? Is it possible to use a colorized iVAT image to visually display the performance of a classifier, where correctly classified objects can be visually differentiated from incorrectly classified ones? Is it possible to tweak BCiVAT in a way that lowers its off-diagonal level of distraction? We should think more creatively about adding colorization to further increase the usefulness of VAT approaches.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] J.C. Bezdek, Elementary Cluster Analysis: Four Basic Methods that (Usually) Work. Gistrup, Denmark: River Publishers, 2022.

[2] E.D. Hathaway, R. Hathaway, Diagonally colorized iVAT images for labeled data, submitted to the IEEE International Conference on Data Mining, Orlando, FL, USA, (2022), doi: 10.1109/ICDMW58026.2022.00043.

[3] L. Wang, T. Nguyen, J. Bezdek, C. Leckie, K. Ramamohanarao, iVAT and aVAT: enhanced visual analysis for cluster tendency assessment, in: M.J. Zaki, J.X. Yu, B. Ravindran, V. Pudi (Eds.), Advances in Knowledge Discovery and Data Mining. PAKDD 2010. Lecture Notes in Computer Science, 6118. Springer, Berlin-Heidelberg, 16-27, 2010. doi: 10.1007/978-3-642-13657-3_5.

[4] T.C. Havens, J.C. Bezdek, An efficient formulation of the improved visual assessment of cluster tendency (iVAT) algorithm, IEEE TKDE. **23**, 568-584, 2011. doi: 10.1109/TKDE.2011.33.

[5] I. Borg, P.J.F. Groenen, Modern Multidimensional Scaling: Theory and Applications, 2nd Edition. Springer, New York, 200).

[6] J.C. Bezdek, R.J. Hathaway, VAT: A tool for visual assessment of (cluster) tendency, Proc. IJCNN, IEEE Press. (2002) 2225-2230, doi: 10.1109/IJCNN.2002.1007487.

[7] R.C. Prim, Shortest connection networks and some generalizations, Bell System Technical Journal. **36**(1959) 1389-1401, doi: 10.1002/j.1538-7305.1957.tb01515.x.

[8] D. Kumar, J.C. Bezdek, Visual approaches for exploratory data anlysis: A survey of the visual assessment of clustering tendency (VAT) family of algorithms, IEEE SMC Magazine. **6**, 10-48, 2020. doi: 10.1109/MSMC.2019.2961163.

[9] L. Wilkinson, M. Friendly, Michael, The history of the cluster heat map, The American Statistician. **63,** 179-184, 2009. doi: https://doi.org/10.1198/tas.2009.0033.

[10] R.A. Fisher, The use of multiple measurements in taxonomic problems, Annals of Genetics. **7**(1935) 179-188, doi: https://doi.org/10.1111/j.1469-1809.1936.tb02137.x.

[11] M. Yöntem, K. Adem, T. İlhan, S. Kılıçarslan, Divorce prediction using correlation-based feature selection and artificial neural networks, Nevşehir Hacı Bektaş Veli University SBE Dergisi, **9**(1), 259-273, 2019. Retrieved from https://dergipark.org.tr/en/pub/nevsosbilen/issue/46568/549416.

[12] M. Charytanowicz, J. Niewczas, P. Kulczycki, P.A. Kowalski, S. Lukasik, S. Zak, A complete gradient clustering algorithm for features analysis of X-ray images, Information Technologies in Biomedicine, Ewa Pietka, Jacek Kawa (eds.), Springer-Verlag, Berlin-Heidelberg, 15-24, 2011. doi: https://doi.org/10.1007/978-3-642-13105-9_2.

[13] J.M. Huband, J.C. Bezdek, R.J. Hathaway, bigVAT: Visual assessment of cluster tendency for large data sets, Pattern Recognition. **38**, 1875-1886, 2005. doi: https://doi.org/10.1016/j.patcog.2005.03.018.

[14] L.H. Trang, P. Van Ngoan, N. Van Duc, A sample-based algorithm for visual assessment of cluster tendency (VAT) with large datasets, Future Data and Security Engineering (Lecture Notes in Computer Science, 11251) J. Kung, R. Wagner, N. Thoai, M. Takizawa (eds.), Springer-Verlag, Cham,

Switzerland, 145-157, 2018. doi: https://doi.org/10.1007/978-3-030-03192-3_11.
[15] J.C. Bezdek, R.J. Hathaway, J.M. Huband, Visual assessment of clustering tendency for rectangular dissimilarity matrices, IEEE Trans. Fuzzy Syst. 15, 890-903, 2008. doi: 10.1109/TFUZZ.2006.889956.
[16] T.C. Havens, J.C. Bezdek, A new formulation of the coVAT algorithm for visual assessment of clustering tendency in rectangular data, Int. J. of Intelligent Systems. **27**, 590-612, 2012. doi: https://doi.org/10.1002/int.21539.
[17] D. Kumar, J.C. Bezdek, S. Rajasegarar, M. Palaniswami, Adaptive cluster tendency visualization and anomaly detection for streaming data, ACM Trans. On Knowledge Discovery from Data. **11,** 1-40, 2016. doi: https://doi.org/10.1145/2997656.
[18] I.J. Sledge, J.M. Keller, Growing neural gas for temporal clustering, Proc. Int. Conf. Pattern Recognition, 1-4, 2008. doi: 10.1109/ICPR.2008.4761768.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 8, No. 4, 122-130 (2023)

www.astesj.com

ASTESJ

ISSN: 2415-6698

SEVEN ReImagined: A Transmedia Storytelling Evolution Proposal

Joana Braguez*

Escola Superior de Tecnologia e Gestão de Viseu - IPV , CIAC - Centro de Investigação em Artes e Comunicação, Universidade Aberta, CEIS20 -Centro de Estudos Interdisciplinares, Universidade de Coimbra, Viseu, 3500, Portugal

ARTICLE INFO

Article history:
Received:26 June, 2023
Accepted:20 August, 2023
Online: 26 August, 2023

Keywords:
Digital Media Art
Transmedia Storytelling
Social Art
Sin

ABSTRACT

"Seven ReImagined" is an innovative transmedia storytelling project that reshapes the exploration of the seven deadly sins in a modern context. Building upon the original artwork "Seven", this venture incorporates traditional media, digital tools, and the latest immersive technologies to cultivate profound user engagement and interaction. The project's objective is to enhance understanding of timeless moral themes, encourage self-reflection, and foster communal participation. The project centres around a physical and virtual exhibition (via Metaverse), where the artworks dedicated to each sin invite viewers into a reflective and immersive journey. The inclusion of a confessional booth enriches the narrative by allowing viewers to anonymously express and share their interpretation of their own sins. The journey further extends to a dedicated website and a podcast series, serving as a hub for a comprehensive narrative, providing in-depth information about each artwork, and fostering an engaging global dialogue around these universal themes. Engagement through social media platforms allows the project to reach varied audiences, harnessing the participatory culture of these platforms to stimulate reflections and dialogues. By marrying artistic creativity with technological innovation, "Seven ReImagined", creates a multifaceted dialogue on ancient moral wisdom in our contemporary society, providing a profound platform for self-reflection and communal participation. The project's innovation lies in seamlessly blending traditional art with cutting-edge immersive technologies, offering a fresh perspective on ancient moral concepts. Future iterations hold the potential for enhanced sensory experiences, collaborative educational initiatives, data-driven insights, and diverse exhibition formats. In summary, "Seven ReImagined" creatively fuses art and technology, engaging in multifaceted dialogues about ancient moral wisdom. Serving as a platform for introspection and global discourse, the project reaffirms the enduring relevance of fundamental human themes.

1. Introduction

This paper is an extension of work originally presented in ARTeFACTo2022MACAO conference [1]. The nucleus of this research, "Seven", was born from an intricate dissection of Bosch's seminal work circa 1500, "The Seven Deadly Sins". What started as a historical art review soon revealed a startling modern resonance, prompting us to delve into a myriad of inquiries. The author questions primarily revolved around the adaptability of such archaic themes to address contemporary concerns. What techniques, mediums, or materials could be employed to render these themes relatable to the issues of the present day? The research, therefore, extended beyond the canvas and ventured into the world of past and present sin representation in art, gleaning interpretations from a wide array of artists with diverse cultural and historical backgrounds. The exploratory journey led to an intriguing discovery - the Seven Modern Sins. These served as a linchpin, not only shaping the evolution of "Seven" but also providing a crucial bridge between the historical context of sin and the contemporary social issues the author aimed to probe. This extended paper aims to shed light on the metamorphosis of the artefact into a storytelling project It aims to unravel both the reasons behind and the process of this evolution - why this shift is meaningful, and how is envisaged and actualized this transformation. As we delve deeper into the paper, we first elucidate upon the key concepts that form the bedrock of this article: Digital Media Art (DMA), transmedia storytelling, and Social Art. We then embark on an exploratory journey of the art piece itself. From its creation process and the elements that form its backbone to the interactions it incites with the audience, we cover it all. The narrative then transitions to a comprehensive contemplation of the piece, dissecting its impact, exploring its repercussions, and extrapolating its future potential.

In conclusion, the paper provides a holistic view of the journey of "Seven" so far and ponders over future pathways. It does not

* Corresponding Author: Joana Braguez, jbraguez@estgv.ipv.pt

merely conclude, but opens up dialogues and questions, paving the way for future work and potential enhancements.

2. Framework

2.1. Media Digital Art – brief considerations

Since the dawn of the 1960s, the world has seen an extraordinary progression of media forms, transitioning from the initial stages of the digital revolution to our current epoch of social media ubiquity. The nomenclature for these technological art forms varies greatly, encompassing labels like Digital Media Art, Media Art, Artemídia (in Brazil), New Media Art, Digital Art, Electronic Art (as termed in Austria), Computer Art, Multimedia Art and Interactive Art. These terminologies emerged during the sixties but only gained prominence in the nineties. In this paper, the term Digital Media Art (DMA) is preferred, primarily due to its congruity with the Portuguese version (the author's mother tongue) and its expansive nature. DMA, as defined in this context, encapsulates all aesthetic discourse forms that employ digital media and qualitative computational techniques to construct digital or computational artefacts. These artefacts present communicational or informational alternatives [2].

Several movements, such as Dada, Pop Art, Fluxus, Kinetic Art, Minimalism, and more, have served as a bedrock for DMA. For instance, Dadaism, partially a reaction to the industrialization of war weaponry and the mechanical reproduction of text and images, mirrors DMA's response to the information technology revolution and digitization [3]. Dadaism introduced or altered various concepts, such as authorship, appropriation, and even the definition of art. For example, Duchamp's "Fountain" was instrumental in challenging these notions. Appropriation practices employed by Dadaists, ranging from renowned artworks like the Mona Lisa to mundane objects [4], resonate within DMA, perhaps due to the simplification brought about by emerging technologies. Such a concept of appropriation is also typical of Pop Art. Many DMA works, akin to Pop Art, are intertwined with the commercial culture [3]. Artists linked with Kinetic Art and Fluxus, such as Nam June Paik, exhibited a keen interest in examining the interactions between media, society, and technology. Their efforts, often materializing as installations and sculptures, manipulated audio-visual technologies to question the authority of mass media. This critical exploration continues in the work of many DMA artists [5], primarily seen in the spheres of Hacking, Tactical Media, and Cybernetic Art. Conceptual Art, according to the authors in [3] significantly paved the way for DMA, emphasizing ideas over the end product. Consequently, DMA is inherently more conceptual in nature. Beryl Graham and Sarah Cook, as cited in [6], argue that new media art is process-centric rather than product-focused. The most relevant practices are fluid, varying in time, space, and authorship. Such a viewpoint suggests that DMA is not preoccupied with aesthetics, like other artistic practices, but rather with functionality.

There are three characteristics that define DMA: processes of interactivity, connectivity, and computability [6]. Another author adds participation to this list [6]. These processes prevail in DMA, not due to artistic or social intentions, but due to the inherent capabilities of the media. The author posits that DMA's concern is not just process over product, but the process as facilitated by the media. The procedural aspect of DMA, marked by its potential for interaction, connectivity, and participation, adeptly invites various forms of user involvement. These qualities are instrumental in the design of a social artwork, a point that will be elaborated on further in the subsequent sections.

2.2. Transmedia Storytelling

Transmedia Storytelling, in short, is stories told through different media; but the concept has nothing simple and there are other adjacent, such as cross media, multiple platforms, hybrid media and intertextual commodity, transmedia worlds, transmedia interactions, multimodality, intermedia; in all, we can perceive an idea of various media, multiple, that perhaps mix, or merge into each other and mainly we can understand that all are based on narratives expressed through a coordinated combination of languages and media or platforms [7]. According to media scholar Henry Jenkins, these concepts align with his theory of media convergence, which creates a content flow across different media channels. This convergence suggests a blurring or complete dissolution of boundaries between different media [8]. In his seminal work, "Convergence Culture: Where Old and New Media Collide", the author [9] argues that media convergence is not just a technological process, but a cultural phenomenon involving new forms of exchanges between producers and users of media content, changing their relationships [10]. the author [9]maps the interrelationship of media convergence, participatory culture, and collective intelligence, three elements essential to the concept of transmedia storytelling. Media convergence, in his view, relates to the flow of content across multiple media platforms, the cooperation between various media industries, and audiences' migratory behaviour towards the content they desire. This convergence has a potential impact on aesthetics, due to its grassroots expression and transmedia storytelling; on knowledge and education, due to collective intelligence and new media literacy; on politics, due to new forms of public participation; and on the economy, through the web 2.0 business model [10]. Transmedia storytelling can also heighten audience engagement and enjoyment, potentially capturing diverse audiences when execute correctly [8]. The concept entered the public debate around 1999, despite early exploration of transmedia narratives in cinema since the late 1980s [11]. In [12] the author breaks down the concept of transmedia storytelling into ten critical elements: the first, already described above, is related to the fact that transmedia storytelling represents a process where the integral elements of a fiction are dispersed through multiple channels, with the aim of creating a unified and coordinated experience, but each media should give its unique contribution, and the information should not be repeated, that is, one should not take a story and adapt it to the different channels, but use for different for each channel, in this way, there will not be a single source that gathers all the information; in the second point the author refers that transmedia storytelling reflects the economic consolidation of media, this is related to the incentive that a media conglomerate has in spreading its brand or franchise in several media platforms, although this transmedia expansion is an economic imperative, it can enable more expansive and immersive stories; in the third point the author explains that transmedia stories are not based on individual characters or specific plots, but rather, in complex worlds that sustain multiple characters and their stories, this process triggers an encyclopaedic impulse in both readers and writers; in the fourth point the author talks about extensions (never adaptations) referring that this can take on different functions, one of them being to add more realism; in the fifth point he explains the economic potential that transmedia storytelling has by creating multiple entry points for different audiences; in the sixth point it is explained the importance of each part being accessible in an autonomous way and also contributing to the whole, in a unique

way, for example, I do not need to watch the Tomb Rider movie to play the game or vice versa; in the seventh point we can understand that the projects that have better results are those where the same artist shapes the story in all the channels involved or in projects where a strong collaboration between teams is fomented; in the eighth point we understand that the transmedia storytelling is the aesthetic ideal for an era of collective intelligence. The transmedia narratives also function as textual triggers that trigger the production, evaluation and archiving of information. The transmedia storytelling disperses information in such a way that no consumer can know everything, forcing him to talk to others, in this way he becomes a hunter and gatherer, travelling between several narratives to form a coherent whole; in point nine we realize that this information is not simply dispersed, they make available functions and goals that readers assume as an enactment of the story in their daily lives; in the last point we understand that some gaps and even excesses allow readers to continue the stories through speculations, they are unauthorized expansions. The reader becomes a creator. On this last point, the author [7] introduced the concept of the "consumer" as the VUP - viewer/user/player. He explains that the VUP transforms the story by using their natural psychological and cognitive abilities to allow the work to surpass the medium and states that it is in the transmedia game and the decentralized authorship that the story is finally revealed, so the VUP becomes the true producer of the work. To the VUP the author proposes adding a "C" to VUP, denoting a creator, thus coining the term VUPC to better represent a transmedia storytelling consumer. From this point on the author will call him VUPC as she believes it is the term that best suits a consumer of transmedia storytelling. The VUPC faces challenges, he requires specific knowledge and technological access to participate in transmedia experiences. This gives rise to the problem of unequal access to technology, which Jenkins calls the digital divide. However, transmedia storytelling can also enhance literacy and accessibility by spreading across multiple channels. Jenkins also states that the culture of participation could diversify content and democratize access to communication channels [10]. A significant pitfall of transmedia storytelling is redundancy resulting from improper use. Repetition of content across different channels can cause the VUPC to lose interest, leading to the potential death of the franchise. Therefore, it is crucial to provide new layers of knowledge and experiences that update the franchise and sustain the loyalty of the VUPC [12]

The authors in [13] provide further clarification about what transmedia storytelling is not. They assert that it is neither adaptation, multimedia, nor cross-media. Adaptation does not add anything new and might lead to loss of interest as it merely retells the same story in another channel. Multimedia involves using multiple communication platforms with no mutual relationship, while transmedia storytelling leverages different platforms and creates a mutual relationship between them. On the other hand, cross-media is a marketing technique aiming to market and disseminate the product across multiple platforms, whereas transmedia storytelling tells different aspects of the same story across all channels. In figure 1 we can see a comparison of these three models.

Its suggested [7] that transmedia storytelling has a nested structure, similar to Russian nesting dolls (Matryoshka). This structure can create different entry points and cater to various audience segments. Within this structure, a dedicated fan might navigate from one medium to another, whereas an occasional consumer might have a specific, isolated, and even sporadic interaction. This structure can be observed in figure 2. This author presents four strategies to expand the world of storytelling: creation of micro-link stories which may for example expand the period between seasons with online clips; creation of parallel stories which may reveal themselves at the same time as the macro-story reveals itself, may evolve and become a spin-off; creation of peripheral stories which may be more or less close to the macro-story, may also evolve and become a spin-off; and finally the creation of user-generated content [7].

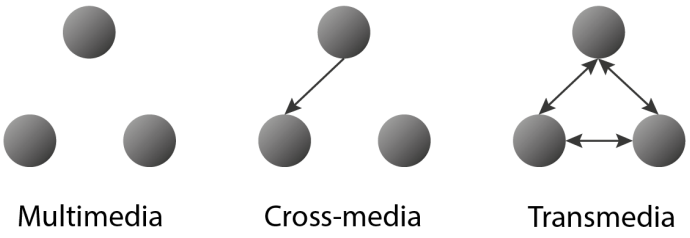


Figure 1: Comparison between Transmedia Storytelling and other storytelling models [13].

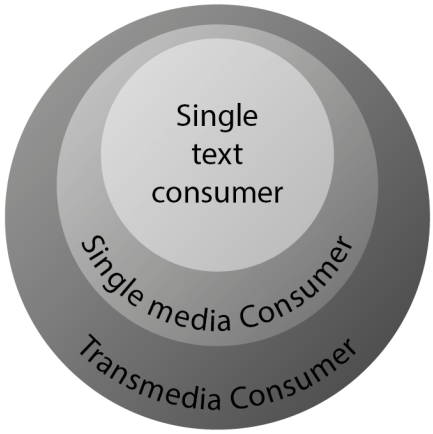


Figure 2: Transmedia World [7].

For a transmedia world to be successful, it must have a story that can extend across multiple media, and the design of the world must be thought of as independent, even of the main character. Each episode should maintain the world's narrative harmony and reveal more than one chain of events adaptable to different media [13]. In Figure 3 the authors describe how each transmedia world is composed of franchised products that contain the main story, which is transferred by "parts", enriching the story of the world through different channels; and each narrative has its own value. Concluding, [14] offers a definition of a transmedia world as abstract content systems from which a repertoire of fictional stories and characters can be actualized or derived across a variety of media forms. The audience and designers share a mental image of the worldness - a set of distinguishing features of its universe, which can be elaborated and changed over time. This world often has a cult (fan) following across media as well.

2.3. Social Art

The term “Social Art” begins by emphasizing that all art is inherently social and has a role to play in influencing societal norms and values. In the modern world characterized by a process of structural hybridization, the term "social art" gains prominence as a simplified and ambiguous definition that acknowledges the interplay between art and society [15]. In [16] it’s called this art

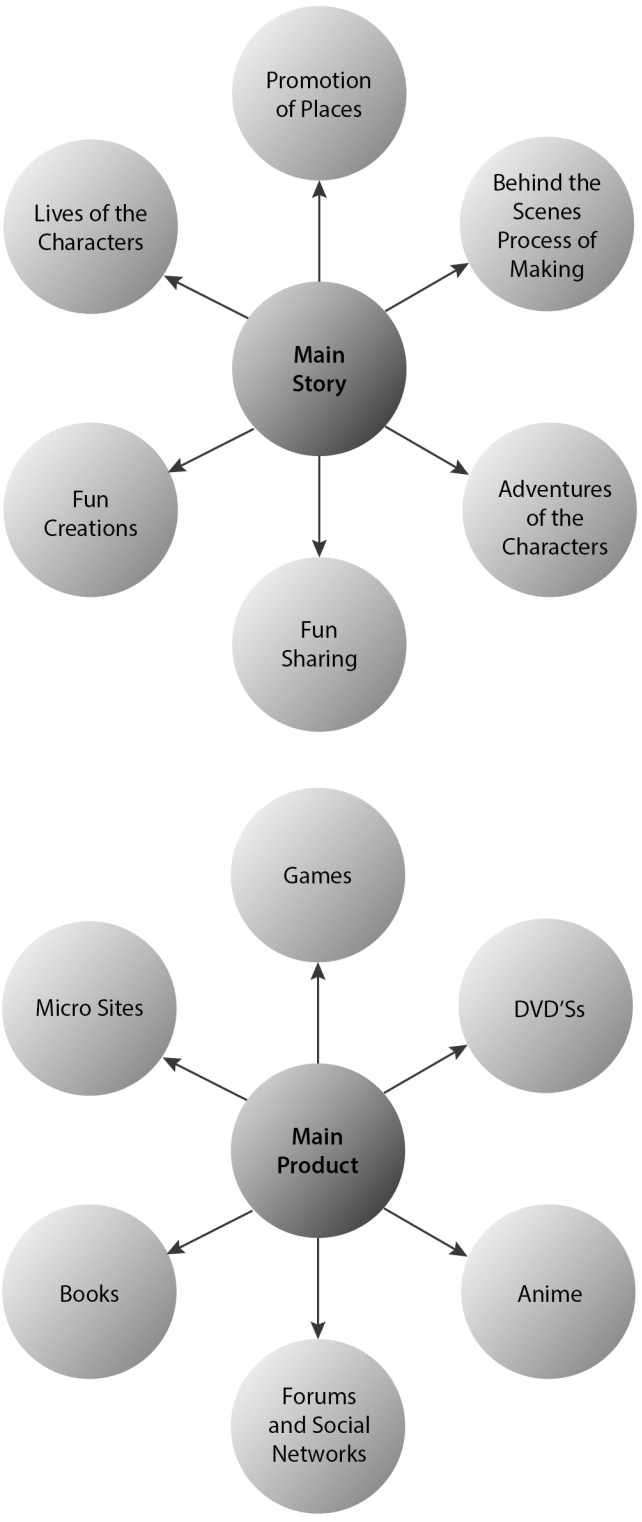


Figure 3: Nested structure [13].

with social approaches "relational aesthetics", art of social justice, social practice, happenings, interventions or community art, she also mentions that in the last years, there was an increase in the number of artists and art projects that intended to communicate with the public, engaging in a process that included careful listening, thoughtful conversation, and community organizing. Other authors [17] refer to another term – socially engaged art – an art form that believes in the responsibility of art and artists to influence social change. This art form uses a variety of tools, operates outside conventional settings, and often involves artists collaborating with community members. It can address social, political, or economic issues, but it's not a requirement. The arsenal of resources at the disposal of socially engaged artists is diverse, including conversation, community coordination, shaping and enhancing public spaces (placemaking), guiding collaborative efforts (facilitation), campaigns to raise public awareness, or policy formulation. Additionally, their mediums extend to theatrical productions, art exhibits, musical performances, participatory digital content, and spoken-word poetry, among others. The artistic process is often collaborative, involving artists, community members, other sectors or fellow artists. While the subject matter may sometimes directly address social, political, or economic issues, it doesn't necessarily have to. The mere act of cultural expression can, in itself, serve as a political act, particularly for groups whose creative voices have been stifled by factors like poverty, cultural assimilation, or systemic oppression. In [18] is traced the roots of socially engaged art back to the 1960s, notably in the incorporation of feminist education into art practice by Alan Kaprow; and the exploration of performance and pedagogy by Charles Garoian and the work of Suzzanne Lacy, among others. The author also mentions that this art falls within the conceptual tradition, safeguarding the issue of the removal of the artist from production that happens in conceptual and minimal art as, in socially engaged art, this cannot happen, the artist is a definitive element. In [19]the discussion is continued, highlighting the 1960s as a pivotal time when artists began to question whether other ways of thinking about art could open the doors that excluded so many from their domain. Although it was understood at the time as a political project, its origin, purpose and values were artistic. More recently the term socially engaged art has been replaced by social practice. This term excludes reference to art, thus avoiding evocations of the role of the modern artist as a visionary and the postmodern version of the artist as a self-conscious critical being. Choosing the term "participatory art," the author[19] sets it apart from community art, explaining that participatory art involves collaboration between professional and non-professional artists, while community art leans towards an emancipatory social commitment.

In 1991, it is introduced another term - New Genre Public Art - which generally refers to public art with an activist nature, often created outside the institutional structure and which places the artist in a direct relationship with the public, while addressing social and political issues. In this paper has been choosing to use the term Social Art as the author believes it is broader and in this way can encompass all the others. Regardless of the term used, and the form this art can take, there are according to [17] three essential components: intentions (what one wants to do and for whom); skills (both artistic and social); and ethics. However, there are works that are not made with social intention but are ultimately beneficial. There is a broad consensus among researchers that community arts have a number of benefits, for example: Living As Form: Socially Engaged Art from 1991-2011 [16] shows us several case studies, although it is only descriptive, it gives us a very comprehensive overview of projects done in different places around the world; Art and Upheaval: Artists on the World's Frontlines by Cleveland where it investigates how art can be used to heal and rehabilitate communities affected by different social problems; How Arts Impact Communities [20] which addresses the problem of how to assess positive and negative impacts on the

community; there are also several municipal reports as well as guidelines for initiating projects. For example, Making Art with Communities: a work guide [21], besides giving us information about different projects, has some guidelines to orientate the creation of a project: all projects should promote and incorporate: social inclusion and equality - respecting diversity and inclusive of differences and needs; active participation; creative collaboration; community and/or collective ownership; transparency, clear processes and honesty; a clear understanding of expectations, process and context; reciprocity - sharing, caring and generosity; respect and trust; empowerment of participants; development of skills, knowledge, capacity and capability; and a shared understanding that everyone has rights and responsibilities. They have also created a guide to help us evaluate the impact of these projects: Evaluating Community Arts and Community-Well-Being.

2.4. Intersecting Pathways: The Synergistic Alliance of Digital Media, Transmedia Storytelling, and Social Art

Over the past decade, the methods used by communities and community organizations to bring attention to their struggles and foster change have evolved rapidly. Digital technology plays a pivotal role in this progress, offering a conduit for the swift and widespread creation and propagation of stories. Transmedia storytelling has emerged as a cutting-edge practice in this context, with several projects already harnessing its potential, consciously or not. Transmedia storytelling redefines the narrative landscape and reshapes the way storytellers engage their audience. It provides platforms for groups that have been marginalized or overlooked to voice their narratives in the public sphere. This socially-driven application of transmedia storytelling has given rise to a new term - Transmedia Activism. The authors who first introduced the term [22], describe it as a fusion of story, community, and collaboration. It harnesses the power of storytelling distributed across multiple mediums, offering myriad perspectives that shed light on social, political, and cultural contexts, thus potentiating effective social action.

Let's consider some prominent projects in this field:

- WWO (World Without Oil) is an interactive game engaging participants across demographics in employing 'collective imagination' to address a real-world issue: the peril our unchecked oil consumption poses to our economy, climate, and quality of life.
- Wasteland, conceptualized by Maria Cristina Finucci, presents a series of interrelated actions across varied locations. Finucci establishes a 'state' composed of ocean-drifting plastics - The Garbage Patch State - replete with the trappings of nationhood, including a flag, constitution, citizens (plastics), a national day (April 11), a passport, and even a Facebook page. This interactive piece, offering postcards or even nationality, brings the environmental impact of plastic waste to the fore in an engaging and unconventional way.
- HIGHRISE is a collaborative documentary project spanning several years and various media formats. Since its inception in 2009, HIGHRISE has produced over 20 unique projects encompassing interactive documentaries, mobile productions, live presentations, performances, installations, site-specific intervention projects, and films.

These examples are just the tip of the iceberg, serving to underscore the real and potential synergies between transmedia storytelling and Social Art, with numerous artists and non-artists already making this collaboration a reality. Active participation is a growing concern among sponsors and artists, and the transformative potential of technological advancements to catalyse and amplify new forms of engagement cannot be overlooked. In [6] the author cites a report commissioned by Arts Council England, Arts & Business, and Museums, Libraries and Archives Council which acknowledges the paradigm-shifting power of the internet to revolutionize how we engage with, share, and create artistic content. It enriches the real-life experience without supplanting it. The report also points to the myriad opportunities now available for arts and cultural organizations to expand and intensify audience engagement via the internet, not only as a marketing and audience development tool but also as a core platform for content distribution and delivery of immersive, participatory, and fundamentally new artistic experiences. Although transmedia storytelling is not confined to the internet, the latter is a potent conduit for such narratives. The author is confident that transmedia storytelling will extend both the reach and the commitment of all involved, creating a more engaging and dynamic sphere for artistic and social expression.

It's evident that DMA and Social Art also exhibit crucial overlapping features in their practices, such as collaboration and participation. Moreover, the author posits that several primary characteristics of DMA can greatly enhance participation and expand reach. While DMA practices do not explicitly aim for social involvement, numerous DMA practices offer platforms for exchange and collaboration between users; they unite communities by making networking and exchanges easier [6]. They also serve as avenues for political activism or advocacy, leveraging internet, mobile, networking technologies, and social media. They offer media education and unrestricted software access and expertise, thereby contesting the capitalist control over proprietary platforms and 'closed' systems (for instance, open-source art pieces and art hacking workshops). Furthermore, they stimulate collective creativity by offering platforms that foster active participation and personal contributions (like community-driven, user-generated projects).

Although DMA is not inherently socially involved, it encompasses approaches and procedures that bear relevance to socially engaged practices. Political participation is intimately tied to new media participation, and there are several correlations between "new media art" and "socially engaged art" [6]. This suggests an intriguing convergence of practices and goals, highlighting how digital media can be a conduit for socially engaged art. By building on shared characteristics such as collaboration and participation, it's possible to create a more dynamic, inclusive, and impactful practice. It shows the power of digital media to bring communities together, challenge existing structures, and create new, shared experiences. Thus, while not explicitly socially engaged, DMA's emphasis on participation, collaboration, and open access makes it a natural partner for Social Art in the pursuit of collective creativity and social change.

In conclusion, the dynamic alliance between DMA, transmedia storytelling, and Social Art presents an exciting new frontier for creative expression, social activism, and community engagement. In the digital age, technology has become the enabling force for communities and organizations, allowing them to raise awareness of their causes and instigate change. Transmedia storytelling, a key innovation in this digital revolution, offers a fresh approach to

narrative that engages audiences on a deeper level and gives voice to groups who were previously marginalized. When viewed through the lens of Social Art, transmedia storytelling transforms from just a method of conveying stories across multiple platforms, to a mechanism for empowering communities and driving social change. In essence, transmedia storytelling and Social Art converge to form 'Transmedia Activism'. The birth of this term testifies to the powerful potential of this innovative approach, as it blends storytelling, community engagement, and collaboration to spark significant social action. On the other hand, DMA, while not inherently socially engaged, showcases various approaches that are strikingly relevant to social practices. The emphasis on participation, collaboration, and open access in DMA practices resonates deeply with the principles of social art.

Therefore, it's evident that digital media, transmedia storytelling, and Social Art are not just individual entities. They are interconnected threads of a larger narrative, each amplifying and enhancing the others. Together, they have the power to reshape the cultural and artistic landscape, fostering a more inclusive, participatory, and socially engaged environment. This convergence of practices and goals signifies a promising future where art is not confined to galleries or concert halls but is ingrained in the fabric of our communities. It foretells a world where stories are not just told but lived, and where art becomes a catalyst for social change. The future of art is here, and it's digital, transmedia, and profoundly social.

3. The Seven Artefact – old version

"Seven," the interactive art installation, probed the correlation between DMA and Social Art by using digital art techniques to convey a socially significant message. This message was designed to raise awareness about various contemporary, worldwide issues. The installation is crafted from a mix of materials and techniques: acrylic is utilized as the main material for its transparent quality, enabling the central content - a blend of resin, pigment, and acrylic paint - to appear as if it's gradually disappearing, thus producing a dreamy aura; around this acrylic structure is a chain of LEDs that illuminates each panel with one of the seven colours from the solar spectrum; the base of this structure houses a sensor that activates both the LEDs and audio, ensuring that these elements are only triggered when a user is within one meter of the panel. An Arduino is also incorporated into the structure to ensure the entire process of LEDs and audio activation, and their respective colour and musical note. The light from the LEDs and the sound contribute to an immersive experience. The LED lights add to the dreamy ambience, while the sound is designed to elicit a slight emotional disquiet. Positioned beneath the painting is a tag with the panel's title, representing one of the original seven deadly sins. The tag carries a retro design, juxtaposing most of the materials used. This concept of duality is omnipresent: good versus evil, old versus new, tradition versus technology, the monochrome painting against the coloured LEDs; the geometric contrasts, the organic against the inorganic, and so forth. To view a brief AR video, the user must install an app. Each video visually depicts a mortal sin; all the videos are created using 3D animation and maintain a transparent background to ensure the panel remains visible. In Figure 4, you can observe one of the panels and the tag that explains how to install the app. At the conclusion of each video, the corresponding modern social sin to the ancient one is revealed. The correlations between the ancient and modern sins, colours, and sounds can be viewed in Table 1, while Figure 5 showcases these correlations with the visual composition located at the centre of the structure.

The user must navigate the journey in the sequence presented in Table 1 and Figure 5. The final sight after viewing all the panels will be the question – WHAT WAS YOUR SIN TODAY? –

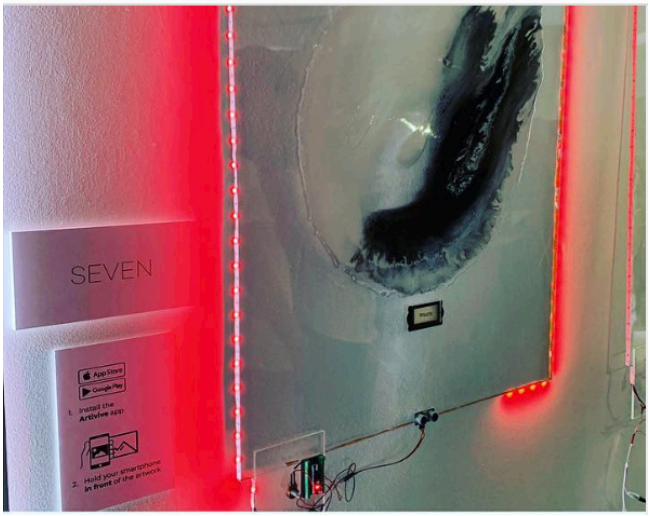


Figure 4: Detail of a panel (exhibited at “Cartografias na cidade”, Quinta da Cruz, Viseu).

intended to incite self-reflection as the culmination of the journey. It's a period of reflection where all your senses relax after the immense stimulation they have encountered thus far. This message should appear in a minimalistic, white area with dim lighting and silence, encouraging this contemplative moment. The use of the English language is to ensure the artwork is accessible to users of various nationalities. The pursuit of universality is also evident in other components such as light and sound. Different users will interact with different aspects; it isn't crucial for them to grasp all elements. For instance, a visually inclined user may focus on the painting and colours, while a user with a keen ear may discern the variety of sounds. This characteristic also amplifies the installation's accessibility.

Table 1- Sins Associations

Sin	Colour	Musical note	Modern sin
Wrath	Red	Do	Drug trafficking
Gluttony	Orange	Re	Violation of the fundamental rights of human nature
Greed	Yellow	Mi	Creating poverty
Envy	Green	Fa	Obscene wealth
Lust	Blue	Sol	Immoral scientific experimentation
Sloth	Indigo	La	Destroying the environment
Pride	Violet	Ti	Genetic manipulation

4. Seven ReImagined

Building upon the foundation laid by the original "Seven" installation, its evolution into a transmedia storytelling project titled "Seven ReImagined" represents the next logical step in its journey of digital and Social Art exploration. "Seven ReImagined" embodies the principles of DMA, Social Art, and transmedia storytelling to create an immersive, interactive, and comprehensive

narrative experience across multiple platforms and channels. The advancements implemented within "Seven ReImagined" seek to adapt to the evolving digital landscape, as well as further the reach and impact of the artwork's core message - prompting reflections on morality and our roles in contemporary global issues. In this iteration, the physical panels of "Seven" form the heart of the experience, but the narrative's reach is extended via digital platforms and interactive technologies. The goal of this evolved project is to maintain the original's ethereal and emotionally impactful essence, whilst enhancing accessibility, interaction, and reach to engage a broader audience and facilitate deeper engagement.

The first major change in "Seven ReImagined" is the introduction of virtual reality (VR). Whereas the original required users to be physically present to fully interact with the installation, the new VR world will enable a remote yet immersive engagement.

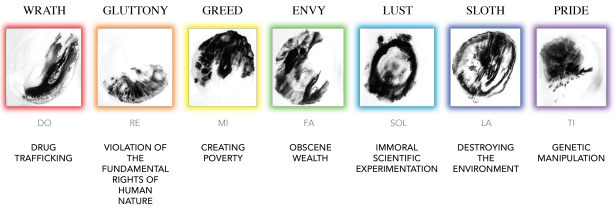


Figure 5: Visual representation of the associations described in table 1.

The project will employ the Metaverse - a collective virtual shared space that is being facilitated by the convergence of virtually enhanced physical reality and physically persistent virtual space - as its hosting platform. Metaverse is essentially a network of 3D virtual worlds focused on social connection, which fits perfectly with the Social Artt objectives of "Seven ReImagined". As a VR-enabled project, "Seven ReImagined" could transcend geographical and physical limitations, making the installation accessible to a global audience. In this VR exhibition, the users, via their avatars, can navigate the virtual gallery that houses the "Seven ReImagined" panels. The panels would be 3D digital reproductions of the original panels, retaining the design and detail, including the sense of depth provided by the acrylic, resin, pigment, and LED components. Creating this virtual exhibition would involve the use of VR development platforms such as Unity or Unreal Engine, both renowned for their VR development capabilities. The development would also require extensive use of 3D modelling tools like Blender for creating both the gallery environment and the 3D digital reproductions of the panels. This shift to VR would enhance the user's ability to engage with the exhibit, adding an extra layer of depth to the overall experience. The user would be able to physically look around the gallery, approach each panel, and engage with it as if they were physically present.

A second change will be made in the audio experience. Actually, this point was a concern since the beginning, the author was never satisfied with the audio experience it seem to be rough and break the meditation process. An immersive audio experience will be developed using binaural audio technologies to create 3D soundscapes. This technique, proven to enhance immersion in projects like BBC's The Turning Forest VR experience, directed by Oscar Raby and written by Shelley Silas, will ensure remote users also experience the unsettling yet thought-provoking sounds designed for each panel. The immersive audio environment will translate well to the VR platform as well. Spatial audio techniques could be employed to create a soundscape that adapts to the user's movements, enhancing the immersion and interactivity of the VR experience.

In addition to the primary exhibition, the VR environment could host live events such as guided tours, talks, or interactive sessions with the creators, enriching the user's understanding of the artwork. Similarly, real-time social features could be added, allowing users to visit the exhibition together with friends, discuss their interpretations, or even meet new people. Incorporating VR into the "Seven ReImagined" project aligns with the increasing trend of creating VR experiences in the art world. Major institutions like the Tate Modern and the Smithsonian American Art Museum have created VR exhibitions, revealing the potential of VR to revolutionize the way we interact with art. By embracing this technology, "Seven ReImagined" positions itself at the forefront of this digital transformation in art, enhancing its reach, accessibility, and impact.

Enhancements will be brought to the physical exhibition, transforming it into a network of eight distinct spaces. Each space will be dedicated to a particular sin, culminating in a final reflective room. The captivating centrepiece of each sin room will be the paintings from the panels of Seven, but they will stretch across a wall, operating as a meditative mandala. The atmospheric lighting in each room won't be provided by LED strips but by room-wide illumination, imbuing each space with a colour representative of the sin it is hosting. This adjustment adds depth and character to each room, immersing the viewer in a coloured ambience. For contemplation, comfortable seating will be scattered throughout each room, ensuring adequate personal space for introspection.

The concluding room will present the thought-provoking inquiry – WHAT WAS YOUR SIN TODAY? – now, the participant will have the opportunity to respond. This room will house three confessional booths where visitors can reflect on their personal transgressions and record their responses via a compact screen and keyboard. All contributions will be gathered anonymously and amalgamated on a dedicated website, fostering an anonymous community of introspection. Upon leaving the final room, guests will be encouraged to extend their engagement through the website or the digital platform. This digital extension, built within the Metaverse, will mirror the physical exhibition, including the distinctive spatial organization and the introspective confessional booths.

Given that the Augmented Reality component of the physical exhibition proved to be somewhat distracting, causing some visitors to forego its use, this component will be omitted in the remodel. Consequently, the 3D videos reliant on this technology will be similarly removed.

In order to amplify the transmedia experience, a bespoke "Seven ReImagined" website will be established as the nexus for the wider narrative. This platform will offer detailed insights into each panel, interviews with the creator, exclusive behind-the-scenes content, and a digital confessional booth for online participants to share their thoughts.

Furthermore, a podcast series will be introduced, following the concept popularized by Lance Weiler in his transmedia project "Hope is Missing." This auditory narrative will delve deeper into the exploration of each sin and its modern equivalent, featuring guest discussions from a diverse range of experts, including

psychologists, philosophers, historians, social activists, and fellow artists.

In the spirit of audience inclusivity, "Seven ReImagined" seeks to effectively infiltrate the digital landscape populated by younger audiences through a robust social media strategy. Utilizing platforms such as Instagram, TikTok, and Twitter, the project will introduce eye-catching animations and teasers, each crafted to embody one of the seven deadly sins in its modern form.

For Instagram, short, visually stimulating videos will be created to grab the audience's attention. Each sin will be represented by a unique colour scheme, symbolic imagery, and brief text overlays to explain its modern interpretation. The Instagram strategy will also include engaging stories and interviews or behind-the-scenes content.

On TikTok, we will harness the platform's algorithm and hashtag system to reach new audiences. Each sin could have its dedicated 'challenge' where users are prompted to create content related to their interpretation of that sin. For instance, for the sin 'Sloth' equated with 'Destroying the environment', the challenge could be for users to share their eco-friendly habits or initiatives.

Twitter's approach will focus on initiating a dialogue around these sins. Scheduled tweet threads detailing the thought process behind each sin and its current interpretation, engaging Twitter polls, and Q&A sessions could be a part of this strategy.

The pre-existing 3D videos from the original "Seven" can be repurposed as short video snippets or GIFs, offering the audience an immersive sneak peek into the project. They can be used as engaging visual content on any of these platforms, encouraging shares, comments, and reactions.

The audience will be motivated to participate by tagging friends, sharing their own insights, or creating related artwork using a specific project hashtag, fostering a global online community centred around the themes presented by "Seven ReImagined". Each social media platform will serve a unique function in this comprehensive strategy, harnessing its specific features and user behaviours to further the reach and impact of the project.

5. Final Considerations and Future Work

The Seven ReImagined transmedia project powerfully combines traditional and digital media, physical and virtual reality, to deliver a profound exploration of the timeless seven deadly sins and their contemporary counterparts. The project marries artistic creativity with technological innovation, effectively facilitating user interaction and engagement. The use of VR technologies in a Metaverse environment creates an immersive digital exhibition that transcends geographical boundaries. The project then seamlessly extends the interaction to the web, harnessing the power of digital platforms to cultivate a global conversation around these universal human themes. The conceptual evolution of the exhibition to encompass eight rooms, each dedicated to a particular sin, facilitates a more profound interaction with the user. The meditative environment, enhanced by colour-coded ambient lighting, encourages introspection and dialogue about these shared moral concerns. The inclusion of a confessional booth, both in the physical and virtual exhibition, is a powerful narrative device allowing for anonymous expression and communal participation. It moves beyond observation and contemplation, to action and contribution. In contrast to the original Seven, which relied on the use of AR technology, this new direction takes advantage of the growing prevalence and acceptance of VR technologies and Metaverse platforms. The decision to shift from AR to VR and Metaverse came after observing that the task of installing an AR app distracted many visitors and disrupted the immersive experience.

The dedicated website and the podcast series further contribute to a transmedia narrative that extends beyond the physical or virtual exhibition. It allows for a more nuanced exploration of the themes, the artistic process, and the societal implications. The engagement on social media platforms allows the project to reach and interact with a younger audience in a space they are most comfortable in.

Reflecting on the project's evolution and learning experiences, it is clear that "Seven ReImagined" has significant potential for further development and expansion. Firstly, future iterations could consider enhancing the VR experience through the inclusion of additional sensory stimuli. Haptic technology, for instance, could make the interaction with the artwork more tangible and immersive. Collaborations with other artists, historians, or technologists in these expansions would also create a more diverse and richer narrative. Secondly, the project could partner with educational institutions and educators, to use "Seven ReImagined" as a pedagogical tool. This could be particularly relevant for subjects such as art, philosophy, ethics, and digital media. The project can stimulate insightful discussions, enable immersive learning experiences, and inspire creative student projects. Thirdly, the confessional data collected anonymously from different platforms could be analysed and used for research purposes. It could provide valuable insights into our shared moral concerns and dilemmas, our understandings and interpretations of the sins, and how they vary across cultures and societies. Lastly, the project could experiment with alternative exhibition formats. Pop-up exhibitions or installations in public spaces could invite unexpected encounters and interactions, reaching a wider audience.

Overall, "Seven ReImagined" is not only a testament to the exciting possibilities of transmedia storytelling and digital innovation, but it is also a reminder of the enduring relevance of ancient moral wisdom in our complex modern world. By continually adapting and evolving, it can continue to inspire, provoke, and engage audiences around the globe in a deeply human conversation.

References

[1] J. Braguez, "SEVEN: a socially engaged digital media art installation," 022 Third International Conference on Digital Creation in Arts, Media and Technology (ARTeFACTo), 2022, doi:10.1109/ARTEFACTO57448.2022.10061262.

[2] A. Marcos, "Média-arte digital: arte na era do artefacto digital/computacional," InVISIBILIDADES: Revista Ibero-Americana de Pesquisa Em Educação, Cultura e Artes, **7**, 4–8, 2014.

[3] M. Tribe, R. Jana, New Media Art, Taschen, Germany, 2007.

[4] T. Laurenzo, DECOUPLING AND CONTEXT IN NEW MEDIA ART, Universidad de la República Montevideo Uruguay, 2013.

[5] M. Connolly, L. Moran, S. Byrne, WHAT IS_(New) Media Art?, Education and Community Programmes, Irish Museum of Modern Art, Jul. 2022.

[6] M. Chatzichristodoulou, "New Media Art, Participation, Social Engagement and Public Funding," Visual Culture in Britain, **14**(3), 301–318, 2013, doi:10.1080/14714787.2013.827486.

[7] C. Scolari, "Transmedia Storytelling: Implicit Consumers, Narrative Worlds, and Branding in Contemporary Media Production," International Journal of Communication, **3**(21), 586–606, 2009.

[8] A. Toschi, "The Entertainment revolution: Does transmedia storytelling really enhance the audience experience? Working paper," in Convergence Culture and Transmedia Storytelling, California State University, Fullerton, 2009.
[9] H. Jenkins, Convergence Culture: Where Old and New Media Collide, New York University Press, New York, 2006.
[10] V. Navarro, "Sites of Convergence: an interview with Henry Jenkins," CONTRACAMPO - REVISTA DO PROGRAMA DE PÓS-GRADUAÇÃO EM COMUNICAÇÃO - UNIVERSIDADE FEDERAL FLUMINENSE, **21**, 2010.
[11] C. Lacalle, "As novas narrativas da ficção televisiva e a Internet," Matrizes, **3**(2), 79, 2011, doi:10.11606/issn.1982-8160.v3i2p79-102.
[12] H. Jenkins, "Transmedia Storytelling 101," 2007.
[13] E. Gürel, Ö. Tığlı, "New World Created by Social Media: Transmedia Storytelling," Journal of Media Critiques, **1**(1), 35–65, 2014, doi:10.17349/jmc114102.
[14] L. Klastrup, S. Tosca, "Transmedial worlds - Rethinking cyberworld design," Proceedings - 2004 International Conference on Cyberworlds, CW 2004, 409–416, 2004, doi:10.1109/CW.2004.67.
[15] C. Madeira, "A (in) visibilidade da arte social," Revista Arte & Sociedade, **2**, 33–44, 2011.
[16] N. Thompson, Living As Form: Socially Engaged Art from 1991-2011, MIT Press, Cambridge, 2012.
[17] A. Frasz, H. Sidford, Mapping the Landscape of Socially Engaged Artistic Practice, Helicon Collaborative, 2017.
[18] P. Helguera, Education for Socially Engaged Art: A Materials and Techniques Handbook., Jorge Pinto Books, New York, 2011.
[19] F. Matarasso, Uma Arte Irrequieta, Fundação Calouste Gulbenkian, Lisboa, 2019.
[20] J. Guetzkow, "How the Arts Impact Communities: An introduction to the literature on arts impact studies," Working Papers, Princeton University, School of Public and International Affairs, Center for Arts and Cultural Policy Studies, **44**, 2002.
[21] Vic Health, Arts Access, "Making Art with Communities - a work guide," 2007.
[22] Lina Srivastava, The tongue-tied storyteller in Rwanda: Lina Srivastava at TEDxTransmedia 2013, 2013.