

Utilizing 3D models for the Prediction of Work Man-Hour in Complex Industrial Products using Machine Learning

Ahmet Emin Ünal^{*1,2}, Halit Boyar¹, Burcu Kuleli Pak¹, Vehbi Çağrı Güngör³

¹R&D Department, Adesso Turkey, Istanbul, 34398, Turkey

²Dept. of Comp. Engineering, Istanbul Technical University, Istanbul, 34485, Turkiye

³Dept. of Comp. Engineering, Abdullah Gül University, Kayseri, 38080, Turkiye

ARTICLE INFO

Article history:

Received: 20 September, 2024

Accepted: 04 November, 2024

Revised: 05 November 2024

Online: 18 November, 2024

Keywords:

Complex Industrial Products

Metal Sheet Stamping

Work Man-hour Prediction

Machine Learning

Gradient Boosting

ABSTRACT

The integration of machine learning techniques in industrial production has the potential to revolutionize traditional manufacturing processes. In this study, we examine the efficacy of gradient-boosting machine learning models, specifically focusing on feature engineering techniques, applied to a novel dataset with 3D product models pertaining to work man-hours in metal sheet stamping projects, framed as a regression task. The results indicate that LightGBM and XGBoost surpass other models, and their effectiveness is further enhanced by employing feature selection and synthetic data generation methods. The optimized LightGBM model exhibited superior performance, achieving a MAPE score of 10.78%, which highlights the effectiveness of gradient boosting mechanisms in handling heterogeneous data sets typical in custom manufacturing. Additionally, we introduce a methodology that enables domain experts to observe and critique the results through explainable AI visualizations.

1. Introduction

This manuscript serves as an extension of a previous study on predicting work man-hours of complex industrial products, originally presented in 2023 4th International Informatics and Software Engineering Conference (IISEC) [1].

This study aims to contribute to the application of machine learning in industrial production by focusing on enhancing efficiency, productivity, and decision-making, specifically targeting work man-hour prediction in metal sheet stamping. By addressing this challenge, our research provides insights that fit within the broader scope of machine learning advancements in manufacturing. The integration of machine learning techniques in industrial production has the potential to revolutionize traditional manufacturing processes. Predictive systems for forecasting production and operational costs are crucial in shaping the future of machine learning applications in industrial production, and this study directly contributes to this important research area by focusing on work man-hour prediction.

In the field of complex industrial product management, where a custom configuration is needed for every product, accurately predicting the work man-hour for a product is essential for ensuring successful project completion. Rapid and precise responses to cus-

tomers inquiries are crucial to maintaining competitiveness in the industry. However, given the complex and configurable nature of products, traditional methods of cost estimation may not provide the needed speed and accuracy. In the conventional approach, according to the domain knowledge of experts who shared the required dataset, they estimate the man-hour using customer requirements, 3D models, past similar projects, and a comprehensive analysis of the product. Traditional cost estimation methods have struggled to keep pace with the increased complexity and competitive environment of the industry, highlighting the need for more advanced approaches.

Recently, machine learning techniques have shown promising empirical results in improving the accuracy of various cost prediction models across many industrial sectors. This study builds upon these advancements by applying machine learning specifically to work man-hour prediction in the metal sheet stamping industry, addressing unique challenges in custom, short-run production. Recent studies have explored the application of machine learning in enhancing cost estimation in manufacturing processes. In [2], the authors applied back-propagation neural networks (BPN) and least squares support vector machines (LS-SVM) to address product life cycle cost estimation challenges, demonstrating the potential of ma-

*Corresponding Author: Ahmet Emin Ünal, Fax: +90 212 346 20 03, Phone: +90 212 346 20 02 & ahmet.unal@adesso.com.tr

chine learning in this area. Similarly, in [3], authors emphasized the importance of selecting a standard set of attributes for developing machine learning models for building project cost estimation, showcasing the advancements that machine learning offers in accurate cost estimation within the construction sector.

Research in [4] focused on explainable artificial intelligence for manufacturing cost estimation and machining feature visualization, indicating a growing interest in deep learning approaches for estimating manufacturing costs. In [5], authors proposed the use of two-dimensional (2D) and three-dimensional (3D) convolutional neural networks (CNN) for manufacturing cost estimation, highlighting the potential of deep learning methods in this context. In [6], authors explored early cost estimation in customized furniture manufacturing using machine learning, showcasing the application of machine learning for estimating costs in specific manufacturing niches. Furthermore, [7] discussed how intelligent job shop scheduling (JSS) systems, powered by machine learning and artificial intelligence solutions, aim to reduce costs based on specific cost functions, such as making span or economic cost. Additionally, [8] conducted an empirical study in the automotive industry, where they proposed machine learning as an advanced cost estimation method.

The use of neural networks in machining operations has been highlighted as advantageous in reducing uncertainties within the cost estimation process. In [9], authors emphasized the capacity of neural networks to enhance cost estimation accuracy in machining operations, showcasing the potential of machine learning in refining cost estimation models. In [10], authors compared various machine learning methods for estimating the manufacturing cost of jet engine components, displaying the effectiveness of different machine learning approaches in cost estimation for the aerospace industry. Moreover, in [11], authors developed methods for direct cost estimation in manufacturing parts, with recent studies leveraging deep learning techniques to predict manufacturing costs based on 3D CAD models. Additionally, [12] highlighted how machine learning improves prediction performance in surface generation and roughness in ultraprecision machining, emphasizing the role of machine learning in advancing automation and digitization in manufacturing processes.

Forecasting the work man-hour for producing complex industrial products poses distinct challenges. In contrast to conventional manufacturing methods that entail bulk production of identical units, often running into thousands or millions, metal sheet stamping operations are frequently tailored with short-run, tailored orders. This variability in design, materials, and processes complicates work man-hour estimations. Furthermore, the time-sensitive nature of such projects, combined with intense industry competition, demands swift and accurate cost estimates. The automotive sector serves as an example, predominantly employing manufacturing through sheet metal stamping projects [13].

The reliance on custom orders in the metal sheet stamping industry results in significant variability between projects, often leading to discrepancies in cost estimations. This variability complicates accurate cost prediction and underscores the importance of developing advanced estimation methods to mitigate financial and operational risks. An inaccurate prediction not only affects the financial bottom line but can also disrupt the broader supply chain, delay projects, and damage client relationships. In the worst cases, it may cause

the rejection of profitable projects due to overestimated costs or the acceptance of unprofitable ones due to underestimations.

In the context of sheet metal stamping, where high production rates and cost-effectiveness are crucial factors, inaccurate cost predictions can result in suboptimal decision-making regarding material selection, tooling design, and process optimization [14]. This can lead to increased scrap rates, rework, and overall inefficiencies in the production line. Moreover, inaccurate cost predictions may also affect the competitiveness of manufacturers in the market, as cost overruns can erode profit margins and hinder the ability to offer competitive pricing [15]. Furthermore, inaccurate cost predictions in metal sheet stamping can impact the overall sustainability of manufacturing operations. For instance, if the estimated costs do not align with the actual expenses incurred during the stamping process, it can lead to increased waste generation, energy consumption, and environmental impacts [16]. This can undermine efforts to improve the environmental performance of manufacturing processes and reduce the overall carbon footprint of sheet metal stamping operations. Moreover, inaccurate cost predictions can also affect the quality and reliability of stamped metal parts. Suboptimal cost estimations may result in compromises in material selection, tooling quality, or process parameters, leading to variations in part dimensions, surface finish, or mechanical properties [17]. This can ultimately impact the functionality and performance of the stamped components, leading to potential quality issues and customer dissatisfaction.

Accurate cost predictions are essential for ensuring the economic viability, operational efficiency, and sustainability of metal sheet stamping processes. Inaccuracies in cost estimations can lead to significant issues, such as poor cost control, reduced competitiveness, increased environmental impact, and compromised product quality. Accurate predictions are crucial to prevent these issues, ensuring manufacturers can make informed decisions, maintain market competitiveness, and promote sustainable practices. This study aims to address these challenges by leveraging advanced machine learning techniques to enhance cost estimation accuracy in the metal sheet stamping process. Therefore, leveraging advanced cost estimation methods, such as machine learning algorithms or finite element modeling, can help mitigate the risks associated with inaccurate cost predictions and optimize the overall performance of sheet metal stamping processes.

The integration of machine learning in cost estimation processes within the manufacturing sector has shown significant promise in enhancing accuracy, efficiency, and decision-making. From product life cycle cost estimation to customized furniture manufacturing and jet engine component cost estimation, machine learning methods have demonstrated their versatility and effectiveness in optimizing cost estimation models. As manufacturing industries continue to deploy advanced technologies, the role of machine learning in cost estimation will become even more pivotal in driving operational excellence and cost-effectiveness.

In this study, we examine the efficacy of gradient-boosting machine learning models, specifically focusing on feature engineering techniques. We apply these methods to a novel dataset related to work man-hours in metal sheet stamping projects, framing the problem as a regression task. The results indicate that LightGBM and XGBoost surpass other models, and their effectiveness is further improved by employing feature selection and synthetic data generation

techniques.

Our study utilizes gradient boosting machine learning models, known for their efficacy with tabular data, and uniquely incorporates domain-specific knowledge tailored to the metal sheet stamping industry. This integration of expert insights and historical data aims to capture the unique challenges of custom, short-run production, setting our approach apart from general-purpose cost estimation models. This approach aims to enhance the predictive accuracy by integrating insights from historical data and expert analysis, tailored specifically to the nuances of metal sheet stamping.

A significant advancement in gradient boosting is the development of the XGBoost algorithm, known for its scalability and efficiency in building tree boosting models [18]. XGBoost, an integrated learning technique utilizing the gradient boosting algorithm, has been successfully applied in diverse industrial domains. For example, it has been used in predicting power demand for industrial customers [19] and transforming the used car market by accurately predicting prices [20]. The robustness and performance of XGBoost have been demonstrated in various applications, underscoring its effectiveness in industrial cost prediction tasks. Furthermore, the application of gradient boosting in industrial contexts extends to addressing specific challenges in cost prediction and optimization. NGBost, a gradient boosting approach utilizing Natural Gradient, has been developed to tackle technical challenges in probabilistic prediction, thereby enhancing the accuracy and reliability of predictive models [21]. Additionally, diversified gradient boosting ensembles have been employed for predicting the cost of forwarding contracts, showcasing the versatility of gradient boosting methods in effectively handling regression and classification problems [22].

In the realm of energy consumption modeling, gradient boosting machines have been utilized to model the energy consumption of commercial buildings, demonstrating improved prediction accuracy compared to traditional regression models and random forest algorithms [23]. Similarly, in the context of cargo insurance frequency prediction, XGBoost has shown superior accuracy compared to other machine learning models, highlighting the efficacy of gradient boosting in diverse industrial prediction tasks [24].

The utilization of gradient boosting algorithms, such as XGBoost and LightGBM, has significantly impacted industrial cost prediction by enhancing prediction accuracy, scalability, and robustness in diverse industrial settings. From energy consumption modeling to customer attrition prediction, gradient boosting has emerged as a powerful tool for optimizing predictive models and improving decision-making processes in industrial cost estimation and optimization tasks. This study examines the effectiveness of gradient boosting machine learning models as well as feature engineering strategies on a new dataset concerning work man-hours in a metal sheet stamping project, framed as a regression task. The results indicate that LightGBM and XGBoost yield better performance compared to other models, and that feature selection along with synthetic data generation enhance the outcomes. The main aims of this research are as follows:

1. Compare the performance of different machine learning models and feature engineering techniques for work man-hour prediction in metal sheet stamping projects.
2. Identify key variables and features that contribute to the accu-

racy of work man-hour predictions.

3. Assess the integration of industry-specific knowledge into machine learning models, evaluating its impact on predictive accuracy.

The structure of this paper is outlined as follows: Section II reviews the existing research on various methodologies for forecasting work man-hours in industrial projects; Section III provides an explanation of the utilized dataset; Section IV provides a detailed account of the model experiments conducted during the study; Section V presents a discussion of the experimental findings; and Section VI offers concluding remarks.

2. Related Works

Various studies across industrial fields such as automotive, construction, and furniture manufacturing have explored the prediction of production costs, labor costs, and material costs using diverse machine learning methods. The application of these techniques varies significantly based on the industry's specifics and the nature of the available data, highlighting the need for industry-specific adaptations of general methodologies.

In [25], authors employed several machine learning models on wheel cost data of 1340 automobiles. After implementing feature selection techniques, their findings revealed that Support Vector Regression (SVR) achieved the highest R^2 value in the cross-validation set. Interestingly, Linear Regression (LR) scored better in the test set, which may suggest that simpler models can sometimes outperform more complex ones in less volatile environments. This finding is relevant to our research as it underscores the importance of evaluating model complexity in the context of specific data characteristics, which is crucial for optimizing cost prediction accuracy in our own study.

Voxelization is a fundamental process in feature extraction for cost prediction tasks, especially in industrial production settings. It involves converting 3D CAD models or point cloud data into a structured voxel grid, which is particularly important for enabling deep learning models like Convolutional Neural Networks (CNNs) to effectively process and analyze complex geometries. This process is crucial in our research as it allows us to capture detailed geometric features that directly impact cost prediction accuracy, especially in scenarios involving intricate part designs. By using voxelization, we can ensure that our models effectively learn from the geometric complexity of the industrial components, leading to more precise predictions. Various studies in computer science and point cloud processing emphasize the importance of voxelization in processing point cloud data for tasks like object detection and feature extraction [26, 27]. In the field of mechanical parts manufacturing, authors of [28] innovatively applied Convolutional Neural Networks (CNN) to predict manufacturing costs. By utilizing voxelization to transform 3D models into a trainable format, they achieved a mean absolute percentage error (MAPE) of 6.34%. This approach underscores the potential of advanced image processing techniques in enhancing feature extraction for cost prediction models. Voxel-based methods have shown particular success in the aerospace industry as well, where converting complex 3D geometries of jet engine components into voxel grids allows for more accurate cost estimation and defect

detection. Techniques like Fully Convolutional Networks (FCN) and autoencoders, as discussed in [29] and [30], further enhance voxel-based feature extraction for tasks such as object detection and image processing.

The furniture manufacturing industry also demonstrates the importance of early cost estimation due to its highly customizable nature. In [31], authors compared various algorithms, such as Extra Trees Regressors (ETR), Gradient Boosting Regressors (GBR), and Random Forest (RF) on data from 1026 products of a Lithuanian furniture manufacturer. The RF algorithm exhibited superior performance, achieving an R^2 score of 0.84, which highlights the effectiveness of ensemble methods in handling heterogeneous data sets typical in custom manufacturing.

Random Forest, as a versatile machine learning algorithm, excels at handling high-dimensional data and capturing complex relationships, making it ideal for cost prediction and optimization in industrial production. Its robust performance in classification and regression tasks supports accurate component classification and production cost prediction, essential in custom manufacturing [32]. Additionally, Random Forest has been instrumental in developing efficient predictive maintenance systems, enabling organizations to anticipate equipment failures, optimize maintenance schedules, and improve production performance [33]. Its interpretability is particularly beneficial in environments where stakeholders must understand the factors influencing costs or production outcomes, aiding decision-making processes [34]. Furthermore, Random Forest's capability to manage complex interactions and highly correlated variables makes it well-suited for settings with intricate production processes and variable interdependencies [35]. Given its flexibility and strong adaptability in real-world applications, Random Forest is a reliable choice to improve production efficiency and optimize cost prediction in the landscape of custom manufacturing [36]. These studies supports our methodology by demonstrating the value of using ensemble methods like Random Forest to effectively manage variability and complexity, similar to the challenges faced in our study of cost prediction for metal sheet stamping.

Parallel to our focus, in [37], authors developed an early cost estimation model specifically for stamping dies, employing Artificial Neural Networks (ANN), which demonstrated a deviation of 8.28% on test data. This study exemplifies the applicability of ANN in industries where data can be nonlinear and complex. ANNs excel in nonlinear cost estimation for custom manufacturing due to their ability to capture complex nonlinear relationships between variables and costs which is vital to model intricate scenarios [38, 39]. They surpass traditional methods in prediction accuracy, thus optimizing schedules and enhancing decision-making processes [39, 40]. Furthermore, ANNs adapt flexibly to changing data patterns, effectively managing the intricacies of custom manufacturing cost data [41, 42]. They are adept at extracting relevant features from complex datasets and recognizing hidden patterns, which is crucial for optimizing cost estimation models [43, 44]. Despite their complexity, efforts to enhance the interpretability of ANNs help provide transparency in the decision-making process [45, 44]. Ensemble methods involving ANN improve prediction accuracy and reduce errors, thus bolstering the robustness and reliability of models [46]. Furthermore, ANNs demonstrate excellent generalization to unseen data and maintain robust performance in diverse scenarios [47], significantly enhanc-

ing cost estimation processes, optimizing resource allocation, and supporting decision making in custom manufacturing environments. The capabilities of ANNs justify their use as a benchmark in managing the complex interactions and nonlinear relationships inherent in cost data for metal sheet stamping. By evaluating ANNs, we tried to find the best approach for achieving high prediction accuracy and reliability, thereby enhancing the efficiency of our cost estimation model.

Additional research efforts in man-hour prediction across various industries further enrich our understanding. For instance, In [48], authors targeted the power transformers manufacturing sector, comparing Support Vector Machines (SVM), Gaussian Process Regression (GPR), and Adaptive Neural Fuzzy Inference System (ANFIS) models. GPR was found to outperform the others in their dataset, potentially due to its effectiveness in managing noise and uncertainty in production data.

In [49], authors focused on the shipbuilding industry, implementing Multiple Linear Regression (MLR) and Classification and Regression Tree (CART) to predict man-hours in sub-processes. CART outperformed MLR, likely due to its superior handling of categorical and nonlinear data, which are common in such fragmented production processes.

In the construction sector, authors of [50] combined Random Forest (RF) and Linear Regression (LR) to predict Building Information Modeling (BIM) labor costs, finding that the hybrid approach outperformed individual methods. This suggests the potential benefits of methodological hybridization in enhancing prediction accuracy.

These studies collectively highlight the diverse applications and potential of machine learning in cost prediction across industries, informing our approach and methodology in the metal sheet stamping industry, where the challenges of custom, short-run production dominate.

Table 1: Comparison of Existing Studies for Man-hour Prediction of Industrial Products.

Dataset and Paper	SS	# of Features*	Best Model	Year
[37]	150	8	ANN	2014
[49]	300k	11	CART	2015
[51]	99	11	LS-SVM (PSO)	2015
[28]	400k	3D Voxels	CNN	2020
[50]	19	9	RF + LR	2020
[31]	1026	18	RF	2021
[25]	1340	>13	LR	2021
[48]	1249	9	GPR	2022
[52]	>8	4	LS-SVM (PSO)	2023
[53]	1605	10	LR	2023
[1]	4000	47	LightGBM (Optuna)	2023
Our	4890	47 + 3D	LightGBM (Optuna)	2024

SS: Sample Size, *Feature count prior to preprocessing operations

3. Data

In this research, we utilized two distinct datasets pertaining to the output of sheet metal stamping parts to forecast the operational costs involved in the manufacturing process. These datasets encompass details on the product features as well as the die characteristics necessary for each operation. The primary variable of interest is the work man-hour for each operation, referred to as 'OperationCost.' These datasets were supplied by a sheet metal stamping company.

The supplied product information data set encompasses details regarding each individual product (or part). The data set comprises information on 875 products. It includes the following features:

- **InquiryID:** Distinct identifier assigned to each individual part inquiry.
- **SheetThickness:** Thickness of the metal sheet required for manufacturing the part, represented as a floating-point number.
- **NetX and NetY:** Dimensions of the sheet metal, specified as length and width, respectively.
- **ContourSize:** Size of the contour of the final manufactured part.
- **SurfaceArea:** Total surface area of the completed part.
- **SheetTsMax:** Measure of the tensile strength of the sheet material.
- **SheetElongation:** Attribute describing the elongation capacity of the metal sheet.
- **MetalHardness:** Categorical attribute denoting the hardness of the sheet metal, classified as Soft, Medium, or Hard.
- **Year:** The calendar year in which the inquiry quotation request was received.
- **YearDay:** The specific day of the year, ranging from 0 to 365, on which the inquiry was recorded.

The data set for operations encompasses details related to the attributes of the die and the operations required for the production of each product. The dataset comprises a total of 4000 operations, wherein each product is subjected to between 2 to 8 sequential operations, with certain operations potentially being carried out concurrently. The sub-operations are expressed as natural language strings, which had to be parsed with regex. The dataset features the following attributes:

- **OperationID:** A unique identifier assigned to each row within the operations dataset.
- **InquiryID:** An identifier corresponding to the specific part for which the operation is conducted.
- **OperationOrder:** An indicator of the arrangement of the operation within the manufacturing process.
- **PressTonnage:** The tonnage required in the press during the stamping process.

- **DieX, DieY, DieZ:** The dimensions (length, width, height) of the die.
- **DieWeight:** The weight of the die measured in kilograms.
- **DieFilling:** The percentage of the die's internal space that is filled.
- **Sub-operation features:** Boolean and integer features denoting the presence or frequency of various sub-operations (such as metal sheet blanking, shearing, bending, etc.) and other configurations relevant to the die. The string comprises a series of sub-operation types accompanied by the respective frequency of their execution, with each sub-operation type separated by a comma. Through the use of regular expressions to parse this string, we obtained a frequency list of the sub-operations, which was subsequently integrated into the dataset as die directions T, R, L (booleans indicating the die direction top, right, and left respectively); and sub-operation type execution counts. The sub-operations can be described as:
 - **BLANK:** Cuts out a flat metal piece (blank) from a larger sheet, typically in the shape needed for further forming.
 - **SHEAR:** Cuts or trims metal along a straight or curved line to achieve specific dimensions or separate parts.
 - **BEND:** Deforms the sheet along a straight axis to create angles, folds, or flanges, turning flat sheets into 3D shapes.
 - **DRAW:** Pulls metal into a die cavity to form deep, hollow shapes, commonly used for creating cups or cylindrical parts.
 - **GAUGE:** Measures and controls the thickness of the sheet or part to ensure uniformity and adherence to specifications.
 - **WELD:** Joins two or more metal parts together, typically by applying heat or pressure, to create a single assembly.
 - **PROG (Progressive Stamping):** Uses multiple stations in a single die to perform a sequence of operations (like blanking, bending, and drawing) on a single part as it moves through the press.
 - **OTHER:** Other infrequent sub-operation types within the dataset, such as coining, ironing, flanging, hemming, embossing, etc., are collected under this type.
- **work man-hour:** The man-hour of operation, serving as the target variable, which we seek to forecast.

For each set of operational data, the corresponding product information from the product dataset is appended. Consequently, this merging of datasets enables machine learning models to predict the work man-hours for each operation. Subsequently, professionals can integrate these costs to establish the target cost for the operations of a given product. There are numerous suboperations, characterized by interconnections among them. Following consultations

with domain experts, suboperations were categorized into primary groups.

Despite the comprehensive nature of the datasets, several data quality issues were identified that could potentially impact model performance. The distribution of the 'OperationCost' variable, our primary target, exhibited notable skewness, predominantly featuring lower values and discrete increments, often in multiples of 50. This pattern suggests possible label noise due to rounding or estimation practices in recording work man-hours. Additionally, the inclusion of operations with costs below 250 and above 3000 introduced potential sampling bias, as these extremes may represent atypical production scenarios or the involvement of subcontractors utilizing different procedures and equipment. Inconsistencies in the numerical representation of sub-operation counts and missing values in certain features necessitated careful data preprocessing and imputation strategies. These biases and noise within the dataset could lead to challenges such as model overfitting or underfitting, adversely affecting the generalization performance of the predictive models. Nonetheless, in this study, these anomalies were retained following preliminary data cleaning, as they are essential for the model to accommodate these atypical samples to ensure robust learning outcomes.

In the previous study [1], the number of samples was lower, which increased after newly processed and provided data samples from the data source. A better way to represent 3D products and the sequential nature of operations can be suggested to increase the performance of the models. Thus, in this study, we added features to represent 3D attributes of the parts. We were able to acquire some features after processing the STL files of each part. These features are volume of the part that is calculated after voxelization of the part, surface area of the part that is calculated directly from the mesh, and number of triangles in the part file (which was correlated with the complexity of the part).

For each sub-operation within the operations, there were inconsistent numerical representations of the step count. This value is incorporated as a new numeric feature when it is available and assigned a value of 0 in its absence. Additionally, the operation dataset is adjusted to include the aggregate count of subsequent and preceding suboperations to provide temporal context to the model. Furthermore, we enhance the whole feature set through the process of feature crossing, which involves the application of multiplicative combinations of pairs of features, as well as the inclusion of squared terms of individual features. Although this method results in an exponential increase in the total number of features, subsequent feature selection is used to mitigate the overall expansion. Specifically, we retain only the features whose importance exceeds the expected importance of the original features.

Given the characteristics of the manufacturing processes, data on certain sub-operation types exhibited imbalances and sparsity in the dataset. This hindered machine learning models from effectively generalizing these operations. To address this issue, we applied the Synthetic Minority Over-sampling Technique (SMOTE) [54] to create synthetic data for these underrepresented sub-operation types. By thoroughly analyzing the dataset's intrinsic patterns and relationships, such as interactions between product dimensions, material properties, and operational parameters, we ensured that the synthetic samples accurately reflected the complexities of real-world

metal sheet stamping operations. Subsequently, we employed the Tomek-link method [55] to prune the synthetic data, thereby reducing noise and preventing potential overfitting. This approach not only augmented the dataset's diversity and volume but also led to a significant reduction in the mean absolute error (MAE) for the rare sub-process types, while the other sub-processes exhibited minimal changes. Consequently, the machine learning models were able to learn more generalized and nuanced patterns, improving their predictive performance on unseen data and enhancing their applicability in practical settings.

4. Methods

The data's label values are adjusted by a constant factor to enhance stability during training. Both categorical and numerical features have been reviewed with domain experts and tailored to suit the requirements of each algorithm. Following feature processing, new features are generated based on the data's sequential nature. These novel features encompass information on past and future operations for a single operational step. In the quotation process, experts determine sequential procedures, and similar operations may incur varying costs depending on their position within the sequence.

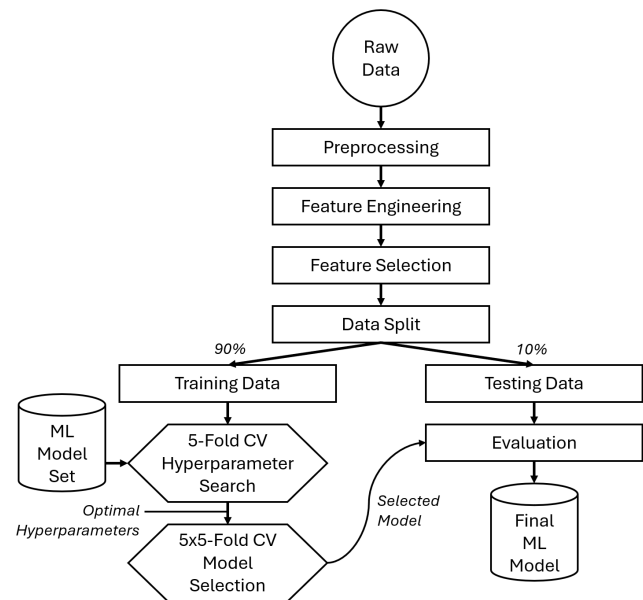


Figure 1: The flow diagram of the process.

Once pre-processing and feature engineering are completed, forward feature selection is employed to reduce the number of features for more stable training. This greedy algorithm in machine learning aims to determine the most relevant features for model prediction. It starts with an empty set and incrementally adds features that most improve model performance until no further significant enhancement is observed or all features are included. Although this method seeks to minimize redundancy and maximize relevance, it can be computationally intensive with high-dimensional datasets and may not always find the optimal subset due to possible local optima.

The data is subsequently divided, with 90% allocated to the training set and 10% to the test set. This test set is used to observe

the capability of the model in unseen data. Each machine learning model undergoes 5-fold cross-validation on the training set to identify the best hyper-parameters. The Optuna library [56] is utilized for hyper-parameter optimization.

To assess the performance of each experimented machine learning model, we conducted 5x5 cross-validation on the training set using the optimal hyper-parameters identified during the hyper-parameter tuning phase. *MAE* and *MAPE* were employed as metrics to compare the various models. Figure 1 provides a flow diagram of the architecture, illustrating how the models are trained and compared.

We conducted trials with a variety of machine learning models. Given their proven effectiveness on tabular datasets, our primary investigation centered on LightGBM [57] as we acquired best results with this method in our previous study [1], while still comparing the method with XGBoost [58]. Additionally, to establish benchmarks, we explored Random Forest, Support Vector Regression, and Multilayer Perceptron techniques. For the models' assessment, we utilized mean absolute error (MAE) and mean absolute percentage error (MAPE). MAE quantifies the average absolute deviation between the forecasted and observed values, whereas MAPE calculates the average percentage deviation between the predicted and true values.

To operationalize the predictive models developed in this study, we implemented a 3D shape-based pricing service designed to integrate seamlessly with the company's existing quotation system. This service provides a machine learning-based tool for predicting industrial product prices, specifically focusing on work-hour estimation for labor costs. It accepts inputs such as 3D models in STL format and various numerical and categorical parameters; including material thickness, type, surface area, hardness, and operation-specific details like mold dimensions and press types, all of which are elaborated in the data section.

Users interact with the service via application programming interface (API) which may be augmented into a dedicated user interface, permitting manual data input or the selection of pre-existing components from the system. The API also allows users to enable or disable the inclusion of 3D data in the predictions. Upon receiving input, the system extracts key features from the 3D model, such as triangle count, total surface area, and volume. These features, along with additional parameters, are fed into the selected machine learning model.

The predicted work hours are converted into a labor cost using a configurable multiplier, allowing the cost to be adjusted based on departmental rates or project-specific requirements. The system also calculates department-specific costs proportionally to the work hours, providing a detailed cost breakdown for activities such as CAD, CAM, 2D cutting, drilling & machining, assembly, measurement, and various CNC processes. This flexible approach enables users to make informed pricing decisions quickly, streamlining the cost estimation process.

For deployment and integration, the service is containerized using Docker and designed to run on-premises to ensure data confidentiality. It exposes a set of API endpoints for various functionalities, including data operations, model training, and prediction. These APIs allow for uploading 3D models and tabular data, configuring cost ratios, training new models, and making predictions. The ser-

vice supports both single models trained on the entire dataset and ensemble models trained using 5-fold cross-validation, providing options for balancing performance and computational efficiency. This modular and secure design facilitates easy integration with other applications and supports the scalability of the solution within the company's infrastructure.

5. Results and Discussion

In this section, we present a comprehensive analysis of our findings. We begin with an overview of the experimental setup used for hyperparameter tuning, followed by a detailed examination of model experimentation results. Subsequently, we delve into model interpretability through the utilization of SHAP values. Finally, we examine the results of software testing, with particular emphasis on usability and performance metrics.

5.1. Experimental Setup

In our experiments with machine learning models, we meticulously optimized the hyperparameters to enhance the models' performance in predicting work man-hours for metal sheet stamping projects. The hyperparameter tuning was conducted using the Tree-structured Parzen Estimator available in the Optuna library [56], which efficiently explores the hyperparameter space to identify optimal settings.

An investigation of the final acquired hyperparameters of a model, LightGBM, can provide a more profound comprehension of the obtained results. The final model employed the 'dart' boosting type, which integrates dropout techniques into the boosting process to prevent overfitting by randomly dropping trees during training. We selected a learning rate of 0.33 to accelerate convergence, allowing the model to learn efficiently from the data without excessively prolonging the training time. A maximum depth of 30 and a high number of leaves (208) were set to enable the model to capture complex nonlinear relationships inherent in the manufacturing data, accommodating the intricate interactions among numerous features.

Minimal regularization was applied, with λ_{l1} , λ_{l2} set to near-zero values (1.06×10^{-8} and 2.97×10^{-4} , respectively), indicating that strong regularization was unnecessary due to effective overfitting control by other parameters like feature and bagging fractions. To introduce randomness and promote generalization, we utilized a feature fraction of 0.5675 and a bagging fraction of 0.84 with a bagging frequency of 2. This ensured that each iteration trained on a random subset of features and data samples, reducing the risk of the model becoming too tailored to specific patterns in the training set.

We set the minimum data in a leaf to 1, allowing the model to capture rare patterns and exceptions in the data, which is crucial for accurately predicting work man hour costs associated with infrequent sub-operation types. The maximum bin was configured to 212, permitting finer discretization of continuous features and enabling the model to capture subtle variations in feature values that significantly impact the target variable. By leveraging the Optuna library's Tree-structured Parzen Estimator for hyperparameter optimization, we were able to systematically explore the hyperparameter space and identify the optimal settings that maximized the model's

predictive performance. This careful tuning was essential for developing a robust model capable of achieving high predictive accuracy and meeting our target error rate, thereby effectively supporting decision-making processes in the manufacturing workflow.

5.2. Model Experimentation Results

We utilized the test set and implemented 5x5 cross-validation on the training set to evaluate various models. For the assessment, Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE) metrics were chosen due to their interpretability and wide acceptance within the domain. The outcomes of these evaluations are presented in Table 2. As expected, LightGBM and XGBoost exhibited superior performance compared to other models, with LightGBM achieving the lowest MAE and MAPE values during cross-validation (CV). The results on the test data from the top-performing model (LightGBM) are depicted in Figure 2. Based on consultations with industry experts, a model must exhibit a maximum MAPE of 10% to be deemed valuable, which constitutes the target Key Performance Indicator (KPI) for this study. The findings suggest that the majority of samples fall within this acceptable range. Furthermore, the models were compared using a variety of KPI metrics. Figure 3 depicts the proportion of samples accurately predicted according to these selected KPI metrics across all experiments. The selected 10% error threshold KPI target is also indicated in the Figure 3 as a red dashed line. For the top-selected model, it is apparent that 75% of the samples are within this acceptable range. Therefore, given the complexity of the task, we conclude that it is feasible to develop and deploy models for predicting work man hour costs in the sheet metal stamping industry.

Table 2: Comparison table of the cross validation results

Models	Results	
	5x5 CV MAPE	5x5 CV MAE
LightGBM*	10.89%	71.49
LightGBM	10.78%	70.37
XgBoost*	11.30%	73.72
XgBoost	11.23%	72.25
KNN Regressor	20.55%	122.01
MLP	16.61%	105.25
Linear Regression	26.35%	136.99

Models that end with "*" indicates that the model is trained without the additional 3D data features.

We further assessed how feature engineering techniques affect model performance. Our findings revealed that performance notably improves for both LightGBM and XgBoost after applying feature selection and generating synthetic data for chosen features. It is important to highlight that cost prediction in any field may be constrained by data representation limitations. The features depicting the product and operations might be overly generic, lacking detailed representation for each training example. Additionally, given the typical scarcity of industrial data in cost estimation tasks, additional efforts could focus on data augmentation. This could unlock the

potential of more sophisticated machine learning models, such as Transformers and DNNs.

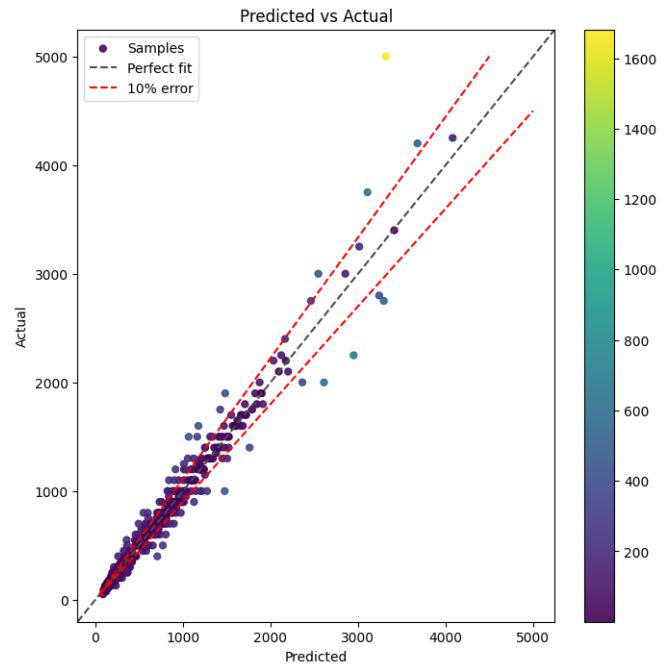


Figure 2: Evaluation of the best model on the test set.

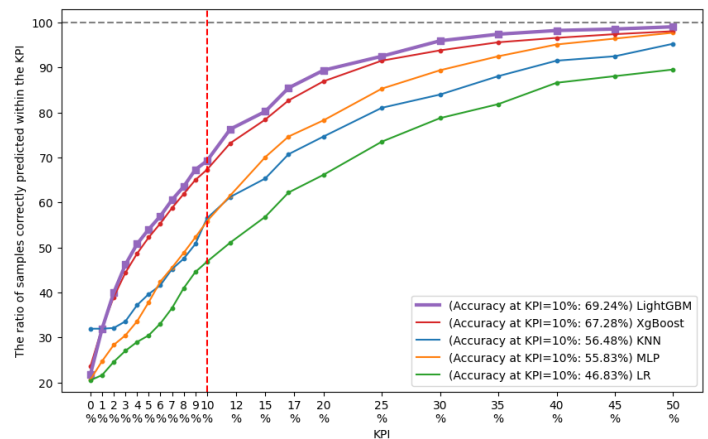


Figure 3: Change of the ratio of samples correctly predicted within different KPI metrics.

5.3. Model Interpretability

In addition, to gain deeper insights into the model's decision-making process, we employ SHAP (SHapley Additive exPlanations) values [59] to interpret the contribution of each feature to the predictions. SHAP is a model-agnostic interpretability method based on cooperative game theory, which assigns each feature an importance value by calculating its average marginal contribution across all possible feature combinations. By analyzing the SHAP values for our model, we found that press tonnage, the presence of a progressive operation, and die dimensions (DieX, DieY, DieZ) significantly influence the work man-hour estimation, as depicted in Figure 4. These features

have the highest SHAP values, indicating they contribute most to the predicted costs.

Specifically, higher press tonnage is associated with increased work man-hours, aligning with domain knowledge that higher tonnage presses require more setup time and operational complexity. The presence of a progressive operation also contributes to higher predicted man-hours due to the additional tooling and coordination required for such operations. Larger die dimensions (length, width, height) impact the prediction by indicating more substantial or complex dies, which typically necessitate more labor for handling and setup. We also observed that the high triangle count of the part 3D model, which correlates with the high part complexity, increases the work man-hours in general.

Other features, such as material properties and other sub-operation counts, have a comparatively moderate effect on the prediction. The SHAP analysis enhances the interpretability of our model by illustrating how each feature influences the output, ensuring that the model's behavior aligns with expert understanding. This transparency in the decision-making process not only validates the model's reliability but also builds trust with stakeholders by demonstrating that the predictions are based on logical and explainable factors relevant to the manufacturing context.

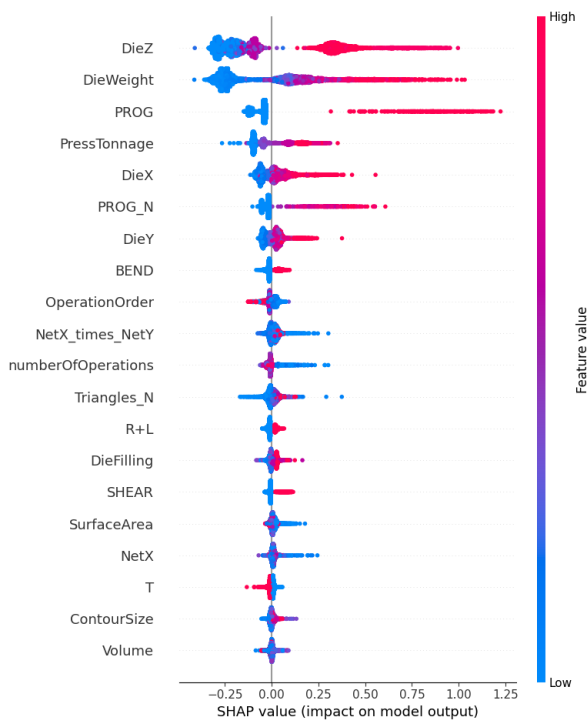


Figure 4: The SHAPley summarization of the features.

5.4. Software Test Results

In our implementation of the pricing service, LightGBM was selected due to its superior performance in predicting work hours, achieving a target error rate of less than or equal to 10%. The adoption of this machine learning model has significantly reduced the time required to generate quotations. On average, the inference time of the model is less than 15 seconds. While the analysis of the 3D part can add some time, especially if the data is not already

stored in the database, resulting in a total average inference time of approximately 2 minutes \pm 15 seconds, the overall process still represents a substantial improvement. Considering the reported time from the manufacturer, this approach reduces the time taken to respond to customer inquiries by 90%, which is crucial in industries where speed is a competitive advantage. This significant reduction in response time not only enhances operational efficiency but also provides a competitive edge in the fast-paced metal sheet stamping industry.

6. Conclusions

In this study, we extended our previous study on work-man hour forecasting in metal sheet stamping processes by conducting a comparative assessment of the efficiency of diverse machine learning algorithms. Additionally, the research investigates the influence of different feature engineering strategies on the results. This problem is formulated and analyzed within the framework of a regression model. We identified the most influential variables and features affecting work man-hours within the field. Additionally, we examined the performance of the models and outlined current limitations that require further investigation. The findings indicated that LightGBM and XGBoost achieved the highest accuracy (lowest MAPE error of 10.78%) compared to other experimented models, exhibiting commendable performance. While the initial study improved the predictive performance of the models through feature selection and synthetic data generation techniques, the present study focused on augmenting the dataset with additional real-world data and incorporating advanced feature engineering methods. With the additional improvements, most influential variables contributing to the work man-hour was similar to the previous study, as the die dimensions, the amount of press tonnage, and the presence of progressive operations, with the exception of die weight. Collaboration with domain experts proved instrumental in understanding the utilization of certain features and the overall constraints of the project. Overall, our research underscores the potential of machine learning models in the context of work man-hours for metal sheet stamping projects and emphasizes the importance of feature engineering and the incorporation of domain-specific knowledge in enhancing model performance. The main limitation of this research is the insufficient availability of real-world data, which obstructs the application of deep learning techniques that could more effectively utilize the 3D models. Future research could focus on improved data representation methods using the 3D part data, such as image renders of the 3D part. Additionally, further research may explore the deployment of deep learning methods, that are adept at leveraging voxel-based 3D data.

Conflict of Interest The authors declare that they have no conflict of interest.

Acknowledgment This work was supported by the TÜBİTAK TEYDEB Program with Project no: 9210055. The dataset and the use case problem were provided by ERMETAL A.Ş. The authors thank Cem Yıldız and Ali Erman Erten for sharing their expertise with us.

References

- [1] A. E. Ünal, H. Boyar, B. Kuleli Pak, C. Yıldız, A. E. Erten, V. C. Güngör, "Man-hour Prediction for Complex Industrial Products," 4th International Informatics and Software Engineering Conference, 2023.
- [2] T. Yeh, S. Deng, "Application of machine learning methods to cost estimation of product life cycle," *International Journal of Computer Integrated Manufacturing*, **25**(4-5), 340–352, 2012, doi:10.1080/0951192x.2011.645381.
- [3] H. Salleh, "Selecting a standard set of attributes for the development of machine learning models of building project cost estimation," *Planning Malaysia*, **21**, 2023, doi:10.21837/pm.v21i29.1359.
- [4] S. Yoo, N. Kang, "Explainable artificial intelligence for manufacturing cost estimation and machining feature visualization," *Expert Systems With Applications*, **183**, 115430, 2021, doi:10.1016/j.eswa.2021.115430.
- [5] F. Ning, Y. Shi, M. Cai, W. Xu, X. Zhang, "Manufacturing cost estimation based on a deep-learning method," *Journal of Manufacturing Systems*, **54**, 186–195, 2020, doi:10.1016/j.jmsy.2019.12.005.
- [6] O. Kurasova, V. Marcinkevičius, V. Medvedev, B. Mikulskienė, "Early cost estimation in customized furniture manufacturing using machine learning," *International Journal of Machine Learning and Computing*, **11**(1), 28–33, 2021, doi:10.18178/ijmlc.2021.11.1.1010.
- [7] L. Yang, J. Li, F. Chao, P. Hackney, M. Flanagan, "Job shop planning and scheduling for manufacturers with manual operations," *Expert Systems*, **38**(7), 2018, doi:10.1111/exsy.12315.
- [8] F. Bodendorf, J. Franke, "Application of the technology acceptance model to an intelligent cost estimation system: an empirical study in the automotive industry," 2022, doi:10.24251/hicss.2022.144.
- [9] M. Atia, J. Khalil, M. Mokhtar, "A cost estimation model for machining operations; an ann parametric approach," *Journal of Al-Azhar University Engineering Sector*, **12**(44), 878–885, 2017, doi:10.21608/aej.2017.19195.
- [10] J. Loyer, E. Henriques, M. Fontul, S. Wiseall, "Comparison of machine learning methods applied to the estimation of manufacturing cost of jet engine components," *International Journal of Production Economics*, **178**, 109–119, 2016, doi:10.1016/j.ijpe.2016.05.006.
- [11] F. Silva, V. Sousa, A. Pinto, L. Ferreira, M. Pereira, "Build-up an economical tool for machining operations cost estimation," *Metals*, **12**(7), 1205, 2022, doi:10.3390/met12071205.
- [12] K. Manjunath, S. Tewary, N. Khatri, K. Cheng, "Monitoring and predicting the surface generation and surface roughness in ultraprecision machining: a critical review," *Machines*, **9**(12), 369, 2021, doi:10.3390/machines9120369.
- [13] H. Y. Gong, J. Y. Wang, Z. H. Zhao, "Study on the springback characteristics of cr340la steel during the typical auto part stamping process," *Advanced Materials Research*, **322**, 98–101, 2011, doi:10.4028/www.scientific.net/amr.322.98.
- [14] R. Zeng, L. Huang, J. Li, "Fracture prediction in sheet metal stamping based on a modified ductile fracture criterion," *Key Engineering Materials*, **639**, 543–550, 2015, doi:10.4028/www.scientific.net/kem.639.543.
- [15] S. Zvonov, Y. Klochkov, "Computer-aided modelling of a latch die cutting in deform - 2d software system," *Key Engineering Materials*, **685**, 811–815, 2016, doi:10.4028/www.scientific.net/kem.685.811.
- [16] S. Kokare, "Toward cleaner space explorations: a comparative life cycle assessment of spacecraft propeller tank manufacturing technologies," *The International Journal of Advanced Manufacturing Technology*, **133**(1-2), 369–389, 2024, doi:10.1007/s00170-024-13745-y.
- [17] M. Moghadam, C. Nielsen, N. Bay, "Analysis of the risk of galling in sheet metal stamping dies with drawbeads," *Proceedings of the Institution of Mechanical Engineers, Part B: Journal of Engineering Manufacture*, **234**(9), 1207–1214, 2020, doi:10.1177/0954405420911307.
- [18] T. Chen, "Xgboost: a scalable tree boosting system," 2016, doi:10.48550/arxiv.1603.02754.
- [19] F. Wang, "Extended-window algorithms for model prediction applied to hybrid power systems," *Technologies*, **12**(1), 6, 2024, doi:10.3390/technologies12010006.
- [20] S. Guo, "Revolutionizing the used car market: predicting prices with xgboost," *Applied and Computational Engineering*, **48**(1), 173–180, 2024, doi:10.54254/2755-2721/48/20241349.
- [21] T. Duan, A. Avati, D. Ding, K. Thai, S. Basu, A. Ng, et al., "Ngboost: natural gradient boosting for probabilistic prediction," 2019, doi:10.48550/arxiv.1910.03225.
- [22] D. Ruta, M. Liu, L. Cen, Q. Vu, "Diversified gradient boosting ensembles for prediction of the cost of forwarding contracts," 2022, doi:10.15439/2022f291.
- [23] S. Touzani, J. Granderson, S. Fernandes, "Gradient boosting machine for modeling the energy consumption of commercial buildings," *Energy and Buildings*, **158**, 1533–1543, 2018, doi:10.1016/j.enbuild.2017.11.039.
- [24] P. Panjee, "A generalized linear model and machine learning approach for predicting the frequency and severity of cargo insurance in Thailand's border trade context," *Risks*, **12**(2), 25, 2024, doi:10.3390/risks12020025.
- [25] F. Bodendorf, J. Franke, "A machine learning approach to estimate product costs in the early product design phase: a use case from the automotive industry," *Procedia CIRP*, **100**, 643–648, 2021.
- [26] L. Liu, E. Chen, Y. Ding, "Tr-net: a transformer-based neural network for point cloud processing," *Machines*, **10**(7), 517, 2022, doi:10.3390/machines10070517.
- [27] Z. Yang, Y. Sun, S. Liu, X. Shen, J. Jia, "Std: sparse-to-dense 3d object detector for point cloud," in *Proceedings of the IEEE International Conference on Computer Vision*, 204, 2019, doi:10.1109/iccv.2019.00204.
- [28] F. Ning, Y. Shi, M. Cai, W. Xu, X. Zhang, "Manufacturing cost estimation based on a deep-learning method," *Journal of Manufacturing Systems*, **54**, 186–195, 2020.
- [29] X. Zhang, "Multiattention mechanism 3d object detection algorithm based on rgb and lidar fusion for intelligent driving," *Sensors*, **23**(21), 8732, 2023, doi:10.3390/s23218732.
- [30] B. Huang, Y. Feng, T. Liang, "A voxel generator based on autoencoder," *Applied Sciences*, **12**(21), 10757, 2022, doi:10.3390/app122110757.
- [31] O. Kurasova, V. Marcinkevičius, V. Medvedev, B. Mikulskienė, "Early cost estimation in customized furniture manufacturing using machine learning," *International journal of machine learning and computing*, **11**(1), 28–33, 2021.
- [32] V. Svetnik, A. Liaw, C. Tong, J. Culberson, R. Sheridan, B. Feuston, "Random forest: a classification and regression tool for compound classification and qsar modeling," *Journal of Chemical Information and Computer Sciences*, **43**(6), 1947–1958, 2003, doi:10.1021/ci034160g.
- [33] H. Zermane, A. Drardja, "Development of an efficient cement production monitoring system based on the improved random forest algorithm," *The International Journal of Advanced Manufacturing Technology*, **120**(3-4), 1853–1866, 2022, doi:10.1007/s00170-022-08884-z.
- [34] G. Pan, "Xgboost and random forest algorithm for supply fraud forecasting," 2022, doi:10.1117/12.2641948.
- [35] C. Strobl, A. Boulesteix, T. Kneib, T. Augustin, A. Zeileis, "Conditional variable importance for random forests," *BMC Bioinformatics*, **9**(1), 2008, doi:10.1186/1471-2105-9-307.
- [36] A. Ziegler, I. König, "Mining data with random forests: current options for real-world applications," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, **4**(1), 55–63, 2013, doi:10.1002/widm.1114.
- [37] B. Özcan, A. Fiğlalı, "Artificial neural networks for the cost estimation of stamping dies," *Neural computing and applications*, **25**, 717–726, 2014.

- [38] Z. Leszczyński, T. Jasiński, "An artificial neural networks approach to product cost estimation: the case study for electric motor," *Informatyka Ekonomiczna*, **1**(47), 72–84, 2018, doi:[10.15611/ie.2018.1.06](https://doi.org/10.15611/ie.2018.1.06).
- [39] B. Waziri, K. Bala, S. Bustani, "Artificial neural networks in construction engineering and management," *International Journal of Architecture Engineering and Construction*, **6**(1), 2017, doi:[10.7492/ijaec.2017.006](https://doi.org/10.7492/ijaec.2017.006).
- [40] M. Meharie, W. Mengesha, Z. Gary, R. Mutuku, "Application of stacking ensemble machine learning algorithm in predicting the cost of highway construction projects," *Engineering Construction & Architectural Management*, **29**(7), 2836–2853, 2021, doi:[10.1108/ecam-02-2020-0128](https://doi.org/10.1108/ecam-02-2020-0128).
- [41] I. Peško, V. Mučenski, M. Šešljija, N. Radović, A. Vujkov, D. Bibić, et al., "Estimation of costs and durations of construction of urban roads using ann and svm," *Complexity*, 1–13, 2017, doi:[10.1155/2017/2450370](https://doi.org/10.1155/2017/2450370).
- [42] S. Magdum, A. Adamuthe, "Construction cost prediction using neural networks," *Ictact Journal on Soft Computing*, **8**(1), 1549–1556, 2017, doi:[10.21917/ijsc.2017.0216](https://doi.org/10.21917/ijsc.2017.0216).
- [43] O. Durán, N. Rodríguez, L. Consalter, "Neural networks for cost estimation of shell and tube heat exchangers," *Expert Systems With Applications*, **36**(4), 7435–7440, 2009, doi:[10.1016/j.eswa.2008.09.014](https://doi.org/10.1016/j.eswa.2008.09.014).
- [44] M. Bouabaz, M. Hamami, "A cost estimation model for repair bridges based on artificial neural network," *American Journal of Applied Sciences*, **5**(4), 334–339, 2008, doi:[10.3844/ajassp.2008.334.339](https://doi.org/10.3844/ajassp.2008.334.339).
- [45] K. Kim, I. Han, "Application of a hybrid genetic algorithm and neural network approach in activity-based costing," *Expert Systems With Applications*, **24**(1), 73–77, 2003, doi:[10.1016/s0957-4174\(02\)00084-2](https://doi.org/10.1016/s0957-4174(02)00084-2).
- [46] M. Juszczak, "Early fast cost estimates of sewerage projects construction costs based on ensembles of neural networks," *Applied Sciences*, **13**(23), 12744, 2023, doi:[10.3390/app132312744](https://doi.org/10.3390/app132312744).
- [47] A. Zouidi, F. Fnaiech, K. Al-Haddad, "A multi-layer neural network and an adaptive linear combiner for on-line harmonic tracking," 2007, doi:[10.1109/wisp.2007.4447612](https://doi.org/10.1109/wisp.2007.4447612).
- [48] K. Işık, S. E. Alptekin, "A benchmark comparison of Gaussian process regression, support vector machines, and ANFIS for man-hour prediction in power transformers manufacturing," *Procedia Computer Science*, **207**, 2567–2577, 2022.
- [49] M. Hur, S. K. Lee, B. Kim, S. Cho, D. Lee, D. Lee, "A study on the man-hour prediction system for shipbuilding," *Journal of Intelligent Manufacturing*, **26**, 1267–1279, 2015.
- [50] C. H. Huang, S. H. Hsieh, "Predicting BIM labor cost with random forest and simple linear regression," *Automation in Construction*, **118**, 103280, 2020.
- [51] T. Yu, H. Cai, "The prediction of the man-hour in aircraft assembly based on support vector machine particle swarm optimization," *Journal of Aerospace Technology and Management*, **7**, 19–30, 2015.
- [52] S. Guo, T. Jiang, "Cost prediction of equipment system using LS-SVM with PSO," in *2007 International Conference on Wireless Communications, Networking and Mobile Computing*, 5285–5288, IEEE, 2007.
- [53] X. Hu, M. Lu, S. AbouRizk, "BIM-based data mining approach to estimating job man-hour requirements in structural steel fabrication," in *Proceedings of the Winter Simulation Conference 2014*, 3399–3410, IEEE, 2014.
- [54] N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, **16**, 321–357, 2002.
- [55] R. M. Pereira, Y. M. Costa, C. N. Silla Jr., "MLTL: A multi-label approach for the Tomek Link undersampling algorithm," *Neurocomputing*, **383**, 95–105, 2020, doi:<https://doi.org/10.1016/j.neucom.2019.11.076>.
- [56] T. Akiba, S. Sano, T. Yanase, T. Ohta, M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 2623–2631, 2019.
- [57] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, et al., "Lightgbm: A highly efficient gradient boosting decision tree," *Advances in neural information processing systems*, **30**, 2017.
- [58] T. Chen, C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 785–794, 2016.
- [59] S. M. Lundberg, S. I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, **30**, 2017.

Copyright: This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY-SA) license (<https://creativecommons.org/licenses/by-sa/4.0/>).