Ensemble Imputation Method for Forecasting Indonesia Sugar Dataset using Machine Learning

Mas Syahdan Filsafan
Department of Informatics, Faculty of Intelligent Electrical and Informatics Technology
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
syahdanfilsafan58@gmail.com

Riyanarto Sarno
Department of Informatics, Faculty of Intelligent Electrical and Informatics Technology
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
riyanarto@if.its.ac.id

Bernadetta Raras I.R.
Department of Informatics, Faculty of Intelligent Electrical and Informatics Technology
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
rarasbernadetta@gmail.com

Agus Tri Haryono
Department of Informatics, Faculty of Intelligent Electrical and Informatics Technology
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
7025231020@student.its.ac.id

2023 1st International Conference on Advanced Engineering and Technologies (ICONNIC) | 979-8-3503-0648-4/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICONNIC59854.2023.10467599

***Abstract*—Imputing missing values in the Indonesian sugar dataset is crucial to mitigate biased predictions. In this study, we employ KNN imputation with N equal to 5 to address this issue and achieve optimal results. Multiple machine learning regression models are utilized to estimate sugar imports effectively, and a comparison is made with the deep learning method to determine the best-performing approach. After handling missing values, our research demonstrates improved accuracy and minimized errors when implementing machine and deep learning methods. Through KNN imputation, missing values are effectively handled, and the LSTM model is utilized to estimate sugar imports accurately. The LSTM model achieves a Mean Squared Error (MSE) of 0.0303 and Mean Absolute Error (MAE) of 0.1376 on the testing dataset. Additionally, KNN imputation reaches the highest average percentage value of 94.734, indicating the closest similarity to the original data.**

Keywords—Forecasting, Sugar Import, Machine Learning, Deep learning.

I. INTRODUCTION

Granulated sugar plays an important role as a basic necessity in people's daily lives, and is an irreplaceable main ingredient for food and beverage producers and sellers. Granulated sugar has an important role for Indonesian society and the global community. Apart from being used as a basic necessity, sugar is also the main sweetener in the food and beverage industry. This makes the import for sugar not only to meet direct consumption needs, but also for production needs. As reported by sucden.com, sugar consumption reached eight million tons at the beginning of the 20th century with a population of 1.6 billion. Currently, the world population reaches more than seven billion people and sugar consumption reaches more than 170 million tons. This means that there is an increase in sugar consumption not only from an increase in population but also from consumption patterns. The increase in per capita consumption reached between 4-5 times over that period, with 70% of world sugar production coming from sugar cane and the rest from beets. In 2008, Indonesia imported 2.3 million tonnes of sugar equivalent to raw sugar consisting of white sugar, granulated sugar and raw sugar [1].

Based on data released by AGI, the main conclusion is that the largest sugar imports will occur in 2023, reaching 3,181,710 tons. In addition, sugar import fluctuate from year to year. There are also several years where there are no sugar imports, which can be attached in Fig. 1.

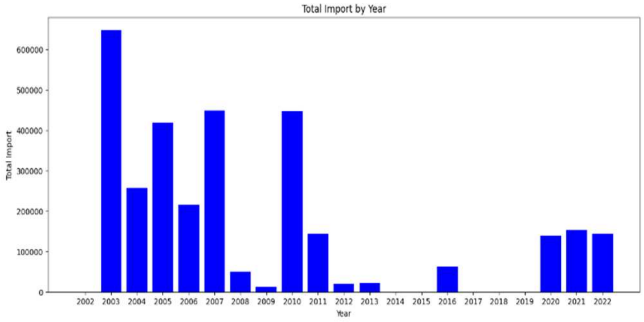

Fig. 1. Sugar Import Volume 2002-20223 From AGI

Gega Indah Arastika's research results in 2016 showed that the variables sugar production, sugar consumption, domestic sugar prices, international sugar prices, the Rupiah exchange rate against the US Dollar, gross domestic product, and import tariffs significantly influenced sugar imports in Indonesia. High sugar production and low domestic sugar prices reduce sugar imports, while high sugar consumption and high international sugar prices increase sugar imports. The Rupiah exchange rate, gross domestic product and import tariffs also affect sugar imports in a similar way [2]. This research has not used a sugar forecasting method for the future, this is needed to anticipate policy makers to prepare for the sugar needs of Indonesian society. One method that can be used is LSTM with variable values of up to 17 variables [2].

Machine Learning is a method for extracting knowledge from previously provided data [3]. It is incapable of overcoming problems such as image classification because the dataset required is very large for image classification. Therefore to resolve this problem, using one of machine learning that is capable of processing very large datasets, namely Deep Learning.

Deep Learning is a part of Machine Learning, which has algorithms like the human brain works. Algorithm of Deep Learning is usually called an Artificial Neural Network (ANN). Data completeness is essential in the Machine Learning and Deep Learning process. Data from the Badan

Pusat Statistik (BPS) Indonesia often contains empty values (missing values). To complete this data a special algorithm is needed. One proven effective method is the KNN Imputer with K-5 values, which uses nearest neighbour values to estimate and fill empty values in the dataset. By applying the KNN Imputer and using K-5 values, empty values can be filled effectively. This produces a more complete and reliable dataset for further analysis [4]. Recently, an article discussed the Voice Activity Detection (VAD) method, which utilizes speech features and speaker embeddings as input to directly predict each speaker's activity in each frame. In this research, the LSTM algorithm approach performs well in detecting speaker voice activity [17].

This research aims to analyze the factors that influence sugar imports in Indonesia over the last 20 years. The method used in this research is secondary data in the form of Panel Data from 2002 to 2023. Researchers will also apply Machine Learning and Deep Learning methods to forecast future sugar imports based on data from previous research. In this research, we will compare the forecasting results from Panel Data with the Machine Learning method. It is hoped that this research can provide a deeper understanding of the factors that influence sugar imports in Indonesia and provide an overview of forecasting sugar imports in the future.

II. Background Study

The literature review is used as a reference to provide a brief summary of research that has been carried out previously and is close in topic to this research. reference for previous research and is used as a reference for the design of this research for existing gaps. Train and test represent the accuracy of the training data and test data respectively in percent.

Missing values pose a common problem across all fields, including healthcare, where their presence can be particularly challenging, specifically to help healthcare by decision making [5]. Significant progress has been made in the field of imputation since 1998, thanks to the advancements in machine learning. The controversy surrounding imputation has been a longstanding topic of discussion, but the emergence of machine learning has contributed to notable advancements in this area [6]. As mentioned, imputation techniques for handling missing data can be approached through statistical methods and machine learning, each with strengths and limitations. However, evidence indicates that statistical imputation techniques exhibit bias in estimating missing values and suffer from information loss [7].

This study aims to integrate Deep Learning methods into Data Panel analysis for future sugar forecasting. While previous research has utilized Data Panel analysis, there needs to be more studies incorporating Deep Learning methods in this context. By adopting this approach, the study seeks to enhance the accuracy and reliability of sugar forecasting by harnessing the power and potential of Deep Learning [2].

The authors conducted a comprehensive analysis of forecasting techniques for linear and stationary processes and advancements in nonlinear and non-stationary time series forecasting models [18]. They specifically focused on real-world applications in manufacturing and health informatics. Their findings revealed a growing interest in nonparametric models for forecasting nonlinear and non-stationary time series. They suggested that incorporating principles from nonlinear dynamic systems into nonparametric modelling approaches can offer a promising path for further advancements in forecasting accuracy [8].

A. Machine Learning

Machine learning is a subfield of artificial intelligence that empowers computers to acquire knowledge and make predictions or decisions by analyzing data without relying on explicit programming. It involves the creation of algorithms that enable systems to identify patterns, extract meaningful insights, and enhance their performance through experience. Machine learning comprises two primary approaches: supervised learning, where models are trained on labelled data to predict future outcomes, and unsupervised learning, which uncovers concealed patterns and structures within unlabeled data. This field finds widespread utility, ranging from image and speech recognition applications to medical diagnosis and financial forecasting. Through data-driven learning, machine learning is pivotal in addressing intricate problems [9].

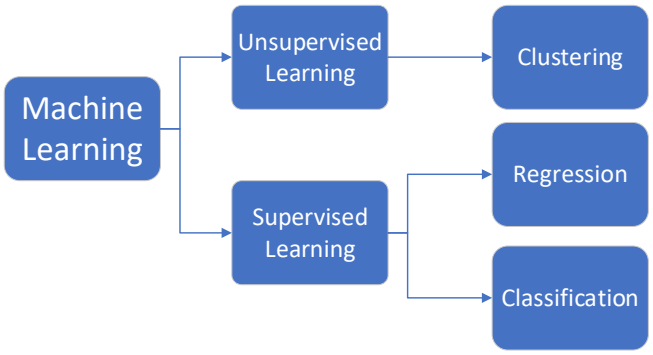

Fig. 2. Depths of Machine Learning Algorithms

B. Data Collection Techniques

The researchers employ documentation techniques to gather data from diverse sources, including Badan Pusat Statistik (BPS) Indonesia, dataindonesia.id, the Indonesian Sugar Association (AGI), literature reviews, statistical data on Java, online sources, and other pertinent references. This approach enables the researchers to acquire preexisting data for analysis and discourse.

Fig. 3. Dataset Sugar Import.

III. Proposed Method

The literature review serves as a reference point, offering a concise overview of previous research in a related field. It assists in identifying gaps in existing knowledge and informs the current research design. The terms "train" and "test" refer

to the accuracy of the training and test data, respectively, expressed as a percentage.

The data modelling process workflow begins with data preparation and collection, after which the data undergoes cleansing to remove any inconsistencies or errors. The cleaned data is divided into training, validation, and testing sets. Subsequently, deep learning modelling (specifically LSTM) and machine learning modelling techniques (including ExtraTree Regressor, Adaboost Regressor, Bagging Regressor, Random Forest Regressor, and Gradient Boosting Regressor) are employed to generate models using the prepared data. The performance of these models is evaluated using metrics such as Mean Absolute Error (MAE) and Mean Squared Error (MSE). Once the evaluation phase is completed, the data modelling process concludes.

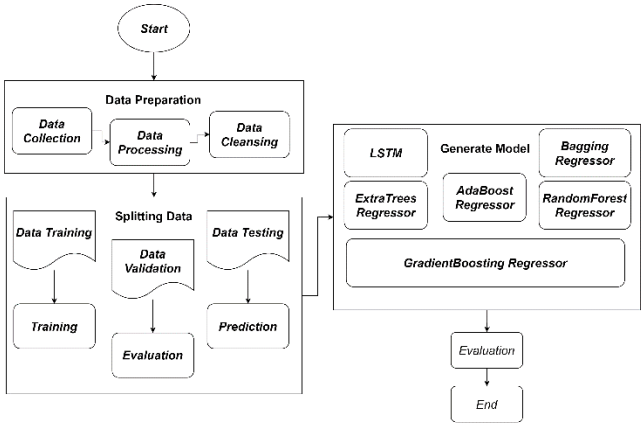

Fig. 4. Flowchart Model Building

A. Data Processing

The data is processed by converting the dataset into a format suitable for predictive analysis, enabling the extraction of insights regarding the potential for optimizing the supply chain in the national sugar ecosystem [19].

B. Data Cleaning

At the data cleansing stage, KNN Imputed was chosen as the method for filling in empty data because it has a high level of similarity, namely 94.734%, with the original data, KNN Imputed shows better performance in dealing with missing values with higher MSE and MAE values. smaller. As a result, this method was deemed feasible for handling missing data.

C. Splitting Data

To perform a random sampling without replacement for splitting the dataset into a training set and a testing set, it is common to randomly select around 60% of the rows and assign them to the training set. Approximately 20% of the rows are allocated to the testing set, while the remaining data is used for validation. The "Features," "Target," and "Validation" sections can be used as indicators to determine the assignment of data to variables like "X_train," "X_test," "X_val," "y_train," "y_test," and "y_val" for the specific division of training and testing data.

D. Handling Missing Value

The dataset underwent missing data imputation techniques using statistical algorithms and Machine Learning methods to fill in the missing values. To evaluate the performance of different approaches for handling missing data in Machine Learning problems, two commonly used criteria, Mean Absolute Error (MAE) and Root Mean Square Error (RMSE), can be employed. Table 1 demonstrates that the K-Nearest Neighbors (KNN) method outperformed other methods, exhibiting superior results regarding both MAE and RMSE.

TABLE I. COMPARISON RESULT HANDLING MISSING VALUE

Metrics	**Algorithm**			
	Median	***Most***	***Constant***	***KNN***
MAE	492.77	493.20	678.82	504.54
MSE	1.87	1.87	2.64	1.86
R^2	0.77	0.77	0.68	0.78

The authors have introduced additional methods to assess the similarity between the original data and the data after performing comparisons. This method calculates the level of accuracy with the following formula:

$$acc = (actual - predicted) / predicted * 100$$

Use of this method provides deeper insight into the extent to which predictions from various approaches to dealing with missing data are close to the actual values. With the combination of MAE, RMSE, and accuracy rate measurements, the authors were able to provide a more holistic understanding of the performance of various missing data handling techniques in the context of the Machine Learning problems discussed in this research.

E. Long Short-Term Memory

LSTM (Long Short-Term Memory) is an extension of the RNN (Recurrent Neural Network) method that incorporates LSTM cells into the RNN architecture. LSTM has effectively tackled various problems, including handwriting recognition, speech recognition, handwriting generation, and image captioning [11]. Unlike RNN, LSTM enables Machine Learning to retain pattern information in data or calculation weights for extended periods. This capability makes LSTM more adept at handling sequences with longer dependencies, thanks to its self-recurrent node known as a cell [10].

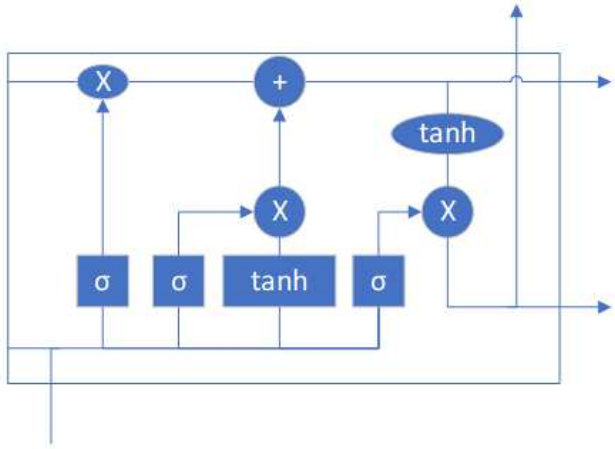

Fig. 5. Long Short-Term Architecture

F. Generate Model use Ensemble Method

Ensemble decision-making is relevant both in our daily lives and machine learning. In real-life situations, we seek diverse opinions to make informed decisions. Similarly, in

machine learning, ensemble learning combines weak classifiers to create a stronger one, using techniques like voting or averaging. In upcoming experiments, various ensemble methods will be explored to demonstrate this concept further [11].

The AdaBoost Regressor is an ensemble learning algorithm that combines weak regressors, typically decision trees, to create a more accurate predictor. It iteratively adjusts instance weights and aggregates predictions from multiple regressors to handle outliers and complex regression problems effectively [12].

The Bagging Regressor is an ensemble method that improves forecast models by incorporating additional training data through irregular sampling. It creates new training data sets by replacing data from the main set, replicating specific observations through replacement sampling. The final forecast in regression is obtained by averaging predictions from multiple models. BR uses a decision tree with twenty sub-models to optimize the output value [13].

The ExtraTrees Regressor is an ensemble learning algorithm that leverages the power of multiple decision trees to enhance prediction accuracy. It belongs to the random forest family but distinguishes itself by not employing bootstrapping. In this approach, each tree within the ExtraTrees ensemble is trained on the complete dataset while utilizing a random subset of the features. This technique enables the model to capture diverse patterns and make robust predictions.

The GradientBoosting Regressor is a versatile and powerful machine learning technique that iteratively utilizes decision trees to make predictions. It progressively fits new trees to the residuals of the preceding tree until a specific stopping criterion is met. This approach is efficient for regression tasks involving intricate relationships between the features and the target variable [11].

The Random Forest Regressor is an algorithm in machine learning that utilizes an ensemble of decision trees to generate predictions. It addresses the overfitting issue by training each decision tree on a different sample of the training data and allowing them to split only a random subset of the available features. This approach enhances the model's ability to generalize to unseen data and improves its robustness [14].

G. Evaluation Method

MSE (Mean Squared Error) and MAE (Mean Absolute Error) are commonly utilized metrics for evaluating forecast accuracy. MSE quantifies the average squared difference between predicted and actual values, assigning greater importance to more significant errors. It can be minimized using gradient descent, Newton's method, or Bayesian optimization. In contrast, MAE treats all errors equally and computes the average absolute difference between predicted and actual values. The optimal model parameters are determined by minimizing MAE, enabling their use in predicting new data. Additionally, this study incorporates mean percentage as an evaluation metric. This metric calculates the average percentage difference between predicted and actual values, offering insights into how the model predictions closely match or deviate from the exact values in percentage form [15].

H. Presentation of Data

Data separation is a crucial step in data pre-processing, particularly when developing data-based models. This technique plays a significant role in enhancing the accuracy of machine learning models by ensuring that the data used for modelling is appropriately organized and distinct [16].

Used dataset:

1. Obtain a dataset that includes information about the import of white sugar circulating on the market.
2. Fill in empty columns using the KNN imputation method to handle missing values.
3. To split the datasets into training and testing sets, use random sampling without replacement. Around 60% of the rows are randomly selected and assigned to the training set, while the remaining 20% are allocated to the testing set. The remaining data can be used for validation.
4. The target column for prediction is the import column.

IV. RESULT AND DISCUSSION

Before proceeding with the machine learning and deep learning models, the author conducted experiments on handling missing values using different algorithms. The following table summarizes the Mean Percentage values obtained from each algorithm:

TABLE II. MEAN PERCENTAGE VALUES FOR HANDLING MISSING VALUES

Algorithm	***Mean Percentage***
KNN	94.734
Most Frequent	94.715
Median	94.701
Constant	93.990

The mean percentage values represent the degree of similarity between the imputed values and the original data. The KNN algorithm achieved the highest mean percentage value of 94.734, indicating that it provided the closest similarity to the original data. Therefore, the author selected KNN as the imputation algorithm for handling missing values, ensuring the integrity of the dataset for subsequent modeling steps.

After handling missing values, the author proceeded to implement ensemble machine learning methods and deep learning models for sugar import prediction. The performance results in terms of MSE and MAE values for the test set are presented in the following table:

TABLE III. PERFORMANCE EVALUATION OF MACHINE LEARNING AND DEEP LEARNING MODELS

Model	***MSE***	***MAE***
AdaBoostRegressor	0.044	0.165
BaggingRegressor	0.048	0.166
ExtraTreesRegressor	0.050	0.164
GradientBoostingRegressor	0.044	0.157
RandomForestRegressor	0.045	0.158
LSTM	0.030	0.137

Among the evaluated models, LSTM outperforms the others with the lowest MSE and MAE values on the test set. This demonstrates the effectiveness of LSTM in capturing long-term dependencies and non-linear patterns in the import data derived from production and import records. The percentage decrease in MSE between AdaBoostRegressor and LSTM is approximately 32.1%, while the decrease in MAE is approximately 16.8%. These results highlight the superior predictive performance of LSTM over the ensemble machine learning models.

In conclusion, the author first handled missing values using the KNN algorithm, ensuring the integrity of the dataset. Subsequently, ensemble machine learning models and LSTM were implemented for import prediction. Experimenting with missing value handling and evaluating different models contribute to the overall understanding of the data and the selection of appropriate techniques for accurate predictions in the import domain.

V. CONCLUSION

In this research, a thorough examination of import predictions was carried out using machine learning and deep learning models. Before applying the model, missing values in the data set are handled using various imputation algorithms. Among these algorithms, the KNN approach was chosen based on its ability to provide the closest similarity to the original data.

After addressing missing values, we assessed various ensemble machine-learning models. LSTM outperformed the rest, demonstrating the test set's lowest Mean Squared Error (MSE) and Mean Absolute Error (MAE) values. These results indicate that LSTM's ability to capture long-term dependencies and non-linear patterns in sequential data surpasses that of machine learning ensemble models. Compared to AdaBoostRegressor, LSTM reduced approximately 32.1% in MSE and 16.8% in MAE. Furthermore, the missing-value experiments underscored the significance of selecting appropriate imputation algorithms. The KNN algorithm exhibited the highest average percentage value, highlighting its effectiveness in maintaining the integrity of the dataset.

Overall, this study highlights the effectiveness of LSTM in accurately depicting complex temporal patterns for import prediction. It also emphasizes the importance of pre-processing steps, such as missing value handling, in improving the performance of prediction models. These findings contribute to overall advances in the understanding and application of machine learning and deep learning techniques in import.

REFERENCES

[1] W. R. Susila and B. M. Sinaga, "Analisis kebijakan industri gula Indonesia," Jurnal Agro Ekonomi, vol. 23, no. 1, pp. 30–53, 2005.

[2] G. I. Arastika, "Analisis faktor-faktor yang mempengaruhi Impor gula di indonesia", Accessed: Jul. 02, 2023. [Online]. Available: https://repository.uinjkt.ac.id/dspace/handle/123456789/52507

[3] A. C. Muller and S. Guido, Introduction to Machine Learning with Python, 1st ed. O'Reilly Media, Inc., 2017.

[4] M. Jupri and R. Sarno, "Taxpayer compliance classification using C4.5, SVM, KNN, Naive Bayes and MLP," in 2018 International Conference on Information and Communications Technology (ICOIACT), Mar. 2018, pp. 297–303. doi: 10.1109/ICOIACT.2018.8350710.

[5] O. F. Ayilara, L. Zhang, T. T. Sajobi, R. Sawatzky, E. Bohm, and L. M. Lix, "Impact of missing data on bias and precision when estimating change in patient-reported outcomes from a clinical registry," Health Qual. Life Outcomes, vol. 17, no. 1, pp. 1–9, 2019.

[6] M. Mohri, A. Rostamizadeh, and A. Talwalkar, Foundations of Machine Learning. MIT Press, 2012.

[7] N. Solomon, Y. Lokhnygina, and S. Halabi, "Comparison of regression imputation methods of baseline covariates that predict survival outcomes," J. Clin. Transl. Sci., vol. 5, no. 1, 2021.

[8] C. Cheng, A. Sa-Ngasoongsong, O. Beyca, T. Le, H. Yang, Z. Kong, and S. Bukkapatnam, "Time Series Forecasting for Nonlinear and Nonstationary Processes: A Review and Comparative Study," IIE Transactions, Sep. 2015, doi: 10.1080/0740817X.2014.999180.

[9] I. A. Sabilla, Z. A. Cahyaningtyas, R. Sarno, A. Al Fauzi, D. R. Wijaya, and R. Gunawan, "Classification of Human Gender from Sweat Odor using Electronic Nose with Machine Learning Methods," in 2021 IEEE Asia Pacific Conference on Wireless and Mobile (APWiMob), 2021, pp. 109–115. doi: 10.1109/APWiMob51111.2021.9435205.

[10] P. Zhou, Z. Qi, S. Zheng, J. Xu, H. Bao, and B. Xu, "Text Classification Improved by Integrating Bidirectional LSTM with Two-dimensional Max Pooling," in Proc. COLING 2016 - 26th Int. Conf. Comput. Linguist. Tech. Pap., 2016, pp. 3485–3495.

[11] D. Sunaryono, R. Sarno, J. Siswantoro, D. Purwitasari, S. I. Sabilla, R. I. Susilo, and A. A. Garibaldi, "Enhanced Gradient Boosting Machines Fusion based on the Pattern of Majority Voting for Automatic Epilepsy Detection," International Journal of Advanced Computer Science and Applications, vol. 13, no. 7, 2022, doi: 10.14569/IJACSA.2022.0130771.

[12] E. Suganya and C. Rajan, "An AdaBoost-modified classifier using stochastic diffusion search model for data optimization in Internet of Things," Soft comput, pp. 1–11, 2019, doi: 10.1007/s00500-019-04206-9.

[13] S. E. Roshan and S. Asadi, "Improvement of Bagging performance for classification of imbalanced datasets using evolutionary multi-objective optimization," Eng Appl Artif Intell, vol. 87, p. 103319, 2020, doi: 10.1016/j.engappai.2019.103319.

[14] S. Seifert, "Application of random forest-based approaches to surface-enhanced Raman scattering data," Sci Rep, vol. 10, pp. 1–11, 2020, doi: 10.1038/s41598-020-62671-4.

[15] V. C. Sugiarto, R. Sarno, and D. Sunaryono, "Sales forecasting using Holt-Winters in Enterprise Resource Planning at sales and distribution module," in 2016 International Conference on Information & Communication Technology and Systems (ICTS), 2016, pp. 8–13. doi: 10.1109/ICTS.2016.7910264.

[16] R. Sarno and et al., Machine Learning Dan Deep Learning-Konsep Dan Pemrograman Python. Andi Publisher, 2022.

[17] N. Aafaq, U. Qamar, S. A. Khan, and Z. H. Khan, "Multi-Speaker Diarization using Long-Short Term Memory Network," in 2023 3rd International Conference on Artificial Intelligence (ICAI), Islamabad, Pakistan, 2023, pp. 164–169. doi: 10.1109/ICAI58407.2023.10136670.

[18] C. Wang, M. Sun, Y. Cao, K. He, B. Zhang, Z. Cao, and M. Wang, "Lightweight Network-Based Surface Defect Detection Method for Steel Plates," Sustainability, vol. 15, no. 4, 2023, doi: 10.3390/su15043733.

[19] John McCarthy, "What Is Artificial Intelligence?", Computer Science Department Stanford, 2007