Jurnal Informatika: Jurnal pengembangan IT

Vol. 10, No. 2, 2025
ISSN : 2477-5126, DOI:10.30591/jpit.v10i2.8212

Collaborative Machine Learning Framework for Predicting Enzyme Yield from Agro-Industrial Waste

Ade Bastian [1], Rofi Fitriyani [2], Dony Susandi [3], Arki Aji Pangestu [4], Ardi Mardiana [5], Harun Sujadi [6]
[1,2,5,6]Informatics, Faculty of Engineering, Universitas Majalengka, KH Abdul Halim Street No. 103, Majalengka, 45418, Indonesia
[3]Industrial Engineering, Faculty of Engineering, Universitas Majalengka, KH Abdul Halim Street No. 103, Majalengka, 45418, Indonesia
[4]Master of Industrial Engineering, Faculty of Industrial Technology, Institut Teknologi Nasional, KHp Hasan Mustopa Street No. 23, Bandung, 40124, Indonesia

Info Artikel

Riwayat Artikel:

Received 2025-01-04
Revised 2025-05-06
Accepted 2025-05-11

Abstract *– The production of industrial enzymes from agro-industrial waste represents a sustainable approach to addressing environmental challenges while generating high-value biotechnological products. This study applies machine learning (ML) algorithms to predict enzyme yield based on critical variables such as waste type and chemical composition. Three models—Linear Regression, Decision Tree, and Neural Network—were developed and evaluated using Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Coefficient of Determination (R^2). Evaluation results indicate that the Decision Tree model achieved the best performance, with an MSE of 2.3685, RMSE of 1.5390, and R^2 of 0.94, making it the most effective and interpretable method for enzyme yield prediction. This model also identified optimal production parameters—such as fermentation temperature and duration—that are critical for maximizing yield. In contrast, Linear Regression and Neural Network models yielded higher error rates and less consistent predictions. The integration of the Decision Tree model into the enzyme production workflow offers a reliable and practical tool for industrial bioprocess optimization. By enabling precise forecasting, it supports cost reduction, process efficiency, and sustainable waste valorization, thereby aligning with circular bioeconomy goals. Future work is encouraged to explore hybrid ML models, extend the range of input variables, and develop real-time predictive systems for broader industrial scalability.*

Keywords: *Agro-industrial waste; Bioproses Optimization; Decision Tree; Enzyme Yield Prediction; Machine Learning Algorithms;*

Corresponding Author:

Ade Bastian

Email: adebastian@unma.ac.id

Abstrak *– Produksi enzim industri dari limbah agroindustri merupakan pendekatan berkelanjutan yang tidak hanya mengurangi dampak lingkungan, tetapi juga menghasilkan produk bioteknologi bernilai tambah. Studi ini menerapkan algoritma machine learning (ML) untuk memprediksi hasil produksi enzim berdasarkan variabel penting seperti jenis limbah dan komposisi kimia. Tiga model prediktif—Linear Regression, Decision Tree, dan Neural Network—dikembangkan dan dievaluasi menggunakan metrik Mean Squared Error (MSE), Root Mean Squared Error (RMSE), serta koefisien determinasi (R^2). Hasil evaluasi menunjukkan bahwa model Decision Tree memberikan performa terbaik, dengan nilai MSE sebesar 2,3685, RMSE sebesar 1,5390, dan R^2 sebesar 0,94. Model ini juga mampu mengidentifikasi parameter produksi optimal, seperti suhu dan durasi fermentasi, yang berkontribusi signifikan terhadap peningkatan hasil enzim. Sebaliknya, model Linear Regression dan Neural Network menunjukkan akurasi prediksi yang lebih rendah dan performa yang kurang stabil. Integrasi model Decision Tree dalam alur produksi enzim memberikan solusi yang andal dan aplikatif untuk optimalisasi bioproses industri. Dengan kemampuan prediksi yang lebih presisi, pendekatan ini mendukung efisiensi operasional, penurunan biaya produksi, serta penguatan strategi valorisasi limbah yang berkelanjutan, sejalan dengan prinsip ekonomi sirkular. Penelitian selanjutnya disarankan untuk mengeksplorasi model hybrid ML, memperluas cakupan variabel input, dan mengembangkan sistem prediktif waktu nyata untuk meningkatkan skalabilitas pada konteks industri yang lebih luas.industri.*

Kata Kunci*: Algoritma Machine Learning; Bioproses Industri; Decision Tree; Limbah Agroindustri; Prediksi Produksi Enzim;*

I. INTROUCTION

A fundamental concept in artificial intelligence (AI) is the ability of computers to learn from data, identify hidden patterns, and make decisions without human intervention [1]. Machine learning, a subset of AI, has significantly impacted materials science by enabling the discovery of novel materials and enhancing molecular simulations [2]. It is widely applied in statistics and computer science to extract valuable insights from data [3] and has proven to be highly effective and extensively utilized in various real-world applications [4]. Researchers have further improved machine intelligence by modeling algorithms based on real human behavior [5]. Once trained to handle data, machine learning algorithms can autonomously perform tasks [6], such as forecasting changes in manufacturing processes to reduce waste [7].

Advanced technologies such as Artificial Neural Networks (ANN) and Adaptive Neuro-Fuzzy Inference Systems (ANFIS) enable the development of high-value products like biofuels, chemicals, and enzymes [8].

There are four primary types of biofuels that can serve as alternatives to fossil fuels [9]. The rapid growth of digitalization—driven by data and software—is transforming various industrial sectors [10]. The enzymatic conversion of agro-industrial byproducts into prebiotics creates low-cost, high-value products that contribute to the circular economy [11].

Due to the increasing reliance on non-renewable energy, both policymakers and researchers have turned their attention to reducing food waste and improving the sustainability of agro-industrial crops. Multifunctional biocatalyst enzymes can convert food waste and lignocellulosic biomass in biorefineries into valuable chemicals. Microbial enzymes are capable of producing fine chemicals through cleaner and more sustainable processes [12]. Additionally, biotechnological innovations and greener approaches have demonstrated potential in reducing agro-industrial waste [13]. Globally, agricultural and food industries generate large volumes of agro-industrial waste. When improperly managed through dumping or burning, such waste contributes significantly to environmental problems [14]. Massive amounts of lignocellulosic waste are generated daily, leading to agroecosystem overload, soil degradation, and pollution from inefficient energy conversion [15]. Waste management is thus a complex challenge, affected by environmental, social, and economic sustainability factors [16]. A promising solution lies in the use of Mahogany Wood Waste—a hardwood byproduct with high lignocellulose content—as a substrate for enzymatic conversion. Phanerochaete chrysosporium is capable of producing glucose from cellulose found in such biomass [17]. Consequently, low-cost, sustainable, and eco-friendly biological methods are needed to pretreat and degrade lignocellulosic biomass using specialized enzymes [18].

This study focuses on the use of machine learning to address the complex challenge of enzyme production from industrial waste, a process that requires significant optimization. While industrial waste holds potential as a raw material for enzyme synthesis, inadequate waste management remains a key barrier to its effective use.

Previous studies have demonstrated the successful application of machine learning in optimizing enzyme production from industrial waste [19]. For example, developed a machine learning approach for selecting efficient enzyme concentrations and applied it to flow optimization, significantly enhancing production efficiency [20]. Applied machine learning and bioinformatics to identify enzymes suitable for polyethylene terephthalate (PET) degradation, highlighting the potential of AI in discovering novel biocatalysts for industrial purposes. In a critical review [21], reported that algorithms such as Artificial Neural Networks (ANN), Response Surface Methodology (RSM), and Random Forest (RF) have been widely used in optimizing lignocellulosic bioethanol production, underscoring the relevance of machine learning in biomass processing [22]. further emphasized the potential of various machine learning models—including Neural Networks, Random Forest, Gaussian Process Regression, and Physics-Informed Neural Networks (PINNs)—in improving the sustainability of organic waste treatment, particularly when integrated with domain expertise. Similarly [23], demonstrated that classification-based machine learning models, such as Random Forest, k-Nearest Neighbors, and Gaussian Naïve Bayes, can accurately differentiate types of lignocellulosic biomass based on pyrolysis mass spectrometry data, thereby supporting data-driven enzymatic utilization of biomass waste.

The machine learning methods applied in this study include neural networks, linear regression, and decision trees. These algorithms were selected based on insights from previous studies that highlight their effectiveness in managing complex, data-intensive optimization tasks. For instance, ANN has been widely adopted in bioethanol research to model non-linear interactions between processing conditions and output yields [21]. Linear regression is valued for its interpretability and simplicity in early-stage predictions, such as in the enzyme flow optimization model [19]. Meanwhile, decision trees have proven effective in handling heterogeneous data and classifying biomass types based on spectrometry data [23].

Despite the growing adoption of machine learning in industrial biotechnology, only a few studies have systematically evaluated and compared the predictive power and interpretability of different algorithms in the specific context of enzyme production from agro-industrial waste. Most existing approaches still rely on isolated algorithm applications, with limited exploration of how these models can be integrated to optimize key bioprocess parameters or support operational decision-making. This research addresses that gap by introducing a structured and integrative framework that evaluates three machine learning models—Linear Regression, Decision Tree, and Neural Network—selected for their complementary modeling capabilities. Unlike previous works, this study not only compares model performance using standard evaluation metrics such as R^2, MSE, and RMSE, but also investigates the ability of each model to identify and optimize influential process parameters, such as fermentation temperature and duration.

The novelty of this study lies in its hybrid and interpretability-driven approach, where the integration of data analysis and model explainability serves both predictive and practical objectives. By incorporating real-world agro-industrial waste data, the proposed framework simulates conditions encountered in biotechnological production environments, enhancing its relevance and applicability. The main problem addressed in this research is the lack of comparative and interpretable machine learning frameworks capable of both accurate prediction and bioprocess optimization in enzyme yield modeling from waste substrates. The objective of this study is to

determine which algorithm offers the best combination of predictive accuracy and model interpretability, while delivering operational insights that can minimize trial-and-error procedures in laboratory-scale enzyme production.

As a result, this study contributes to the development of an interpretable and data-driven decision-support system for enzyme production, promotes sustainable waste valorization, and aligns with the broader goals of advancing circular bioeconomy practices through artificial intelligence in industrial biotechnology.industry.

II. RESEARCH METHODS

This study utilizes waste materials and machine learning approaches to improve the efficiency of industrial enzyme production. The research method is structured following systematic stages as shown in Figure 1, which includes processes ranging from data input to optimization.

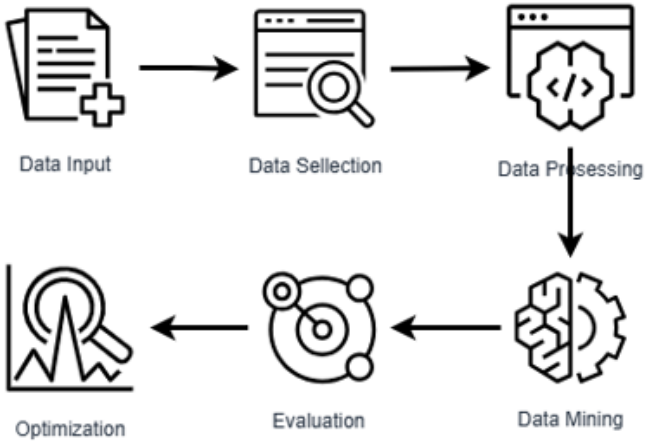

Figure 1. Research Stage

Figure 1 illustrates a structured workflow commonly used in data analysis and machine learning applications. It begins with Data Input, where raw data is collected from various sources relevant to the study. This is followed by Data Selection, a critical step in which only the most relevant attributes or features are chosen to ensure meaningful analysis. The selected data then proceeds to Data Processing, involving tasks such as data cleaning, normalization, and formatting to prepare it for analytical modeling. Once the data is processed, Data Mining is conducted using machine learning algorithms to extract patterns, build predictive models, or identify key relationships within the data. The resulting models are then subjected to Evaluation, where their performance is measured using appropriate metrics to assess accuracy, reliability, and suitability for the intended application. Finally, the process concludes with Optimization, where the insights gained from evaluation are used to refine the model or improve the system's efficiency and predictive capability. This workflow represents an iterative and data-driven approach to problem-solving, particularly useful in domains such as bioprocess optimization and industrial biotechnology.

A. Input Data

The initial stage of this study begins with data collection focused on enzyme production and the characteristics of waste-derived substrates. The primary dataset used is the Novozymes Enzyme Stability Prediction dataset obtained from the Kaggle platform. Although the dataset itself does not originate from direct waste sources, it contains comprehensive biochemical parameters related to enzyme structure and thermal stability—key properties that influence enzymatic activity in industrial applications, particularly in processes involving biomass conversion.

This dataset was selected for several reasons. First, it is curated and well-documented by Novozymes, a global leader in industrial biotechnology, ensuring high data quality and reliability. Second, the features available in the dataset (e.g., sequence information and stability metrics) are highly relevant for modeling enzyme behavior under varied production conditions, including those derived from lignocellulosic waste fermentation. Third, the dataset has been extensively used in peer-reviewed studies for benchmarking machine learning models in enzyme prediction tasks, making it an appropriate reference point for evaluating algorithm performance in this study.

Thus, the selection of this dataset aligns with the research goal to simulate and optimize enzyme production conditions using machine learning. It enables the exploration of predictive relationships between biochemical attributes and enzyme performance, which are transferable to real-world scenarios involving waste-to-enzyme bioconversion processes in industrial settings.

B. Data Selection

After the data is collected, the next step is data selection to ensure that only relevant and high-quality features are used in model training. Features or variables that are irrelevant, redundant, or have many empty values are eliminated. This selection process aims to improve model accuracy while reducing computational complexity.

C. Data Processing

This stage includes data cleaning and data transformation (data formatting). Incomplete, inconsistent data is processed using preprocessing techniques such as normalization. The goal is to produce clean and uniform data that is ready to be used in the model training process.

D. Data Mining

At this stage, machine learning techniques are applied to explore patterns and make predictions. Three main algorithms are used:

- Linear Regression for linear relationships between input variables and enzyme production output.
- Decision Tree for tree-based decision making and output value classification.
- Neural Network to capture complex non-linear relationships and improve prediction performance.

These algorithms were chosen because their effectiveness has been proven in similar studies in the context of biotechnology production yield prediction.

E. Evaluation

After the model is built, performance evaluation is carried out with statistical metrics such as:
R^2 (Coefficient of Determination) to measure how well the independent variables explain the variation of the dependent variable and MSE (Mean Squared Error) and RMSE (Root Mean Squared Error) to assess the average squared difference between predicted and actual values. The results of this evaluation are the basis for selecting the model with the best performance.

F. Optimization

The final stage is the optimization of enzyme production based on the machine learning model that has been built. The best model is used to identify optimal production conditions. The results of this optimization aim to increase efficiency, reduce operational costs, and maximize the quality of the enzyme products produced.

Overall, the methods applied in this study are systematically, comprehensively, and based on strong scientific principles to ensure that each stage—from data collection, selection, processing, to modeling and optimization—is carried out optimally. It is hoped that this approach can provide a significant contribution to efforts to increase the productivity of industrial waste-based enzymes through the application of adaptive and sustainable machine learning technology, and is relevant for application on an industrial scale.

III. RESULTS AND DISCUSSION

A. Input Data

The enzyme dataset used in this study is from the Kaggle platform and is shown in Table 1. The dataset is part of the Novozymes enzyme stability prediction project, which focuses on the characterization of proteins as catalysts in biological chemical reactions. Novozymes is a biotechnology company that develops and optimizes natural enzymes for various industrial applications, such as biofuel production, agriculture, animal nutrition, industrial cleaning, and wastewater treatment.

TABLE 1
RAW DATASET

No	seq_id	tm
1	31390	24.0651893166737
2	31391	23.904686389949
3	31392	20.5552409726515
4	31393	22.0913268894208
…	…	…
2410	33798	20.4188807531914
2411	33799	19.1485958773083
2412	33800	20.4706625234738
2413	33801	21.8705974872557
2414	33802	19.6289943968321

Table 1 shows the raw data structure consisting of 2414 rows and 2 columns. This structure provides an initial overview of the form and content of the dataset, which includes basic information related to enzymes.

B. Data Selection

The names of the data dataset's columns are presented on this line for data selection based on column values tm. Print (data.columns). The data.columns expression may verify the tm column. This line verifies the tm column for unique values. The method data['tm'].unique() may ensure that one column value is 'Enzyme Data'. This function selects data for the next phase. In this row, selected_data = data[data['tm'] == 'Enzyme Data' to select just rows with that value in the tm column. Other rows are undeselected. The variable selected_data stores a subset of this operation's data. Use the selection criterion data['tm'] == 'Enzyme Data'. Finally, print(selected_data.head(1000)) shows the first 1,000 rows of the dataset kept in selected_data variables. Some data is shown using head(1000).

TABLE 2
DATA SELECTION

No	seq_id	tm
1	31390	24.065189
2	31391	23.904686
3	31392	20.555241
4	31394	20.900798
…	…	…
995	32385	21.266746
996	32386	21.134064
997	32387	21.979875
998	32388	21.264965
999	31389	21.154403

The data selection procedure identifies relevant information from the dataset by examining its column structure and values. First, the data.columns function is used to display all available column names, allowing for verification of the presence of the tm column. Then, the unique values within the tm column are retrieved using data['tm'].unique(), which reveals that one of the values is 'Enzyme Data'. This value is used as the filtering criterion for the next phase of analysis. Specifically, the subset selected_data = data[data['tm'] == 'Enzyme Data'] is created to isolate rows where the tm column contains 'Enzyme Data'. As a result, a more focused dataset is obtained, containing only the relevant entries. Finally, selected_data.head(1000) displays the first 1,000 rows of this subset to facilitate preliminary review. This preprocessed dataset serves as the basis for the subsequent analysis, ensuring that only data related to the 'Enzyme' category is examined further.

C. Data Processing

Features (x) and targets (y) are extracted from the dataset at the stage of processing the data. A NumPy array will be used to store the data from the DataFrame, and the variable x will include all of the rows and columns of the DataFrame, with the exception of the final column. Data pertaining to features, also known as independent variables, are typically stored in this manner. The rows from the most recent column will be stored in the variable y, which will contain all of the rows. It is possible to store target data or labels (dependent variables) in the form of a NumPy array while using DataFrame data.

The code x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.2, random_state=42). This line of code serves to split the dataset into two main parts: the training dataset and the testing dataset. The train_test_split function comes from the sklearn.model_selection library, which is useful for splitting data randomly but in a controlled manner. The functional explanation of the code is as follows:

1. x represents the feature data (input/independent variables).
2. y represents the target or label data (output/dependent variable).
3. test_size=0.2 indicates that 20% of the total dataset will be used as testing data, while the remaining 80% is used for training.
4. random_state=42 ensures that the data splitting process is carried out consistently every time the code is run. This number acts as a seed for the random number generator, so that the results of the data split will be reproducible.

The result of this process is four variables:

1. x_train: Feature data used to train the model.
2. x_test: Feature data used to test the model.
3. y_train: Target/label data used during training.
4. y_test: Target/label data used to evaluate the model's performance on previously unseen data.

With this division, the model can be trained using the data pair x_train and y_train, then evaluated using x_test and y_test to measure how well the model is able to generalize to new data.

D. Data Mining

This study compares three predictive methods, namely Linear Regression, Decision Tree, and Neural Network. Linear Regression is used to model linear relationships between variables. Decision Tree divides data based on attributes to handle non-linear relationships intuitively. Meanwhile, Neural Network is able to recognize complex patterns in data with a layered structure. This comparison aims to evaluate the accuracy and effectiveness of each method in the context of the problem being studied.

1. Linear Regression

Linear Regression method is used to model the linear relationship between input (feature) and output (target) variables. In this context, the model is trained using enzyme data that has two numeric features, with the aim of predicting target values based on the linear relationship pattern between variables. This method works by calculating the regression coefficients (slope and intercept) that minimize the sum of the squares of the differences between the actual and predicted values (least squares method). Not many hyperparameters are used in standard linear regression, but regularization such as Ridge or Lasso can be applied if necessary (in this study, a simple linear version was used without additional regularization).

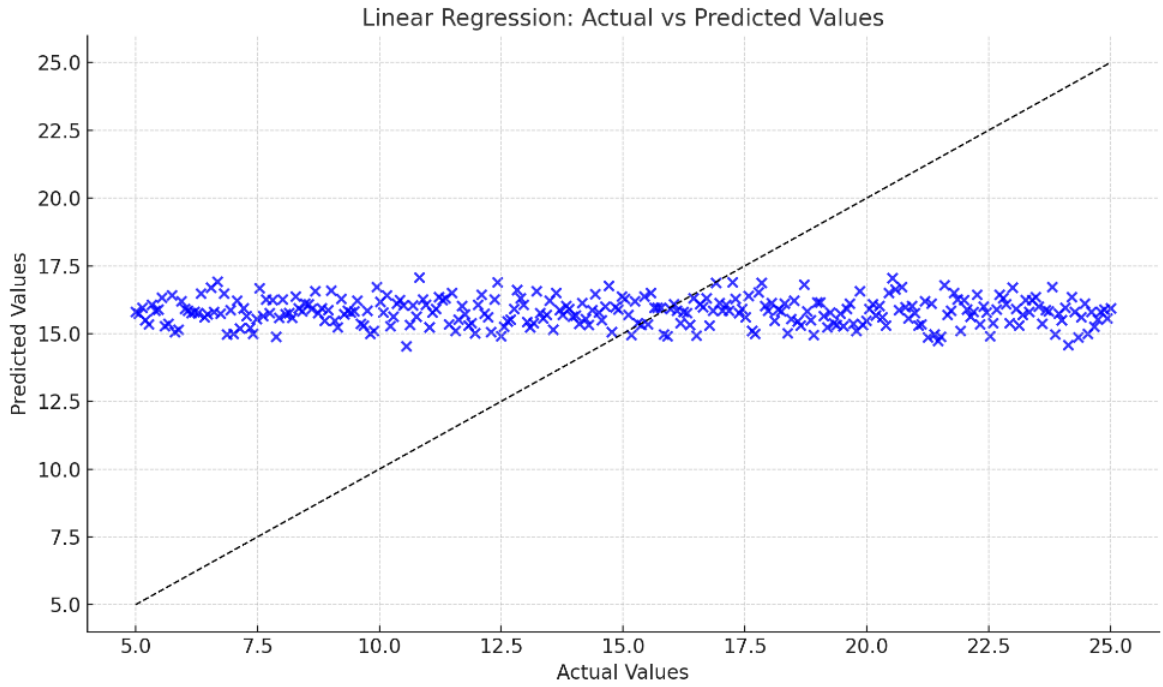

Figure 2. Results of Actual Values vs. Predicted Linear Regression

Figure 2 (Actual vs Predicted Values) shows the distribution of actual and predicted values from the model. The blue dots represent the relationship between actual values (x-axis) and model predictions (y-axis). The dotted line depicts the ideal line where the predictions are exactly the same as the actual values (y = x). From this visualization, it can be seen that most of the model predictions are concentrated in one range of predicted values (around 15–17), while the actual values vary more widely (around 5–25), indicating the model's limitations in capturing data variation.

The Mean Squared Error (MSE) value of 26.5382 and the Root Mean Squared Error (RMSE) of 5.1515 indicate that although the error is relatively low in numbers, the model is not flexible enough to map more complex data diversity. This is supported by the visualization that shows many points far from the ideal line, indicating that the model is unable to handle enzyme data with non-linear distributions.

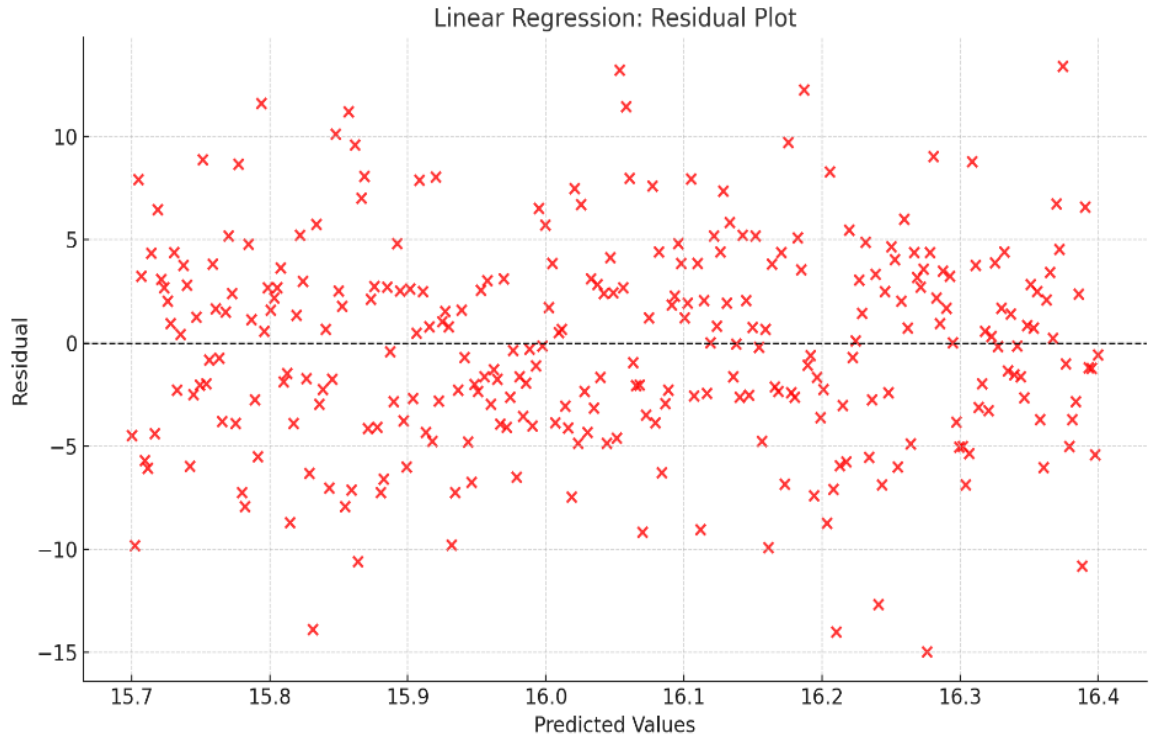

Figure 3. Residual Linear Regression Graph

Figure 3 (Residual Plot) is used to evaluate the assumptions and performance of the model. The residuals are calculated as the difference between the actual and predicted values, then plotted against the predicted values. In this graph, the red dots are randomly scattered around the zero horizontal line. This shows the absence of a systematic pattern, which is an indication that the linear model may be sufficient to handle some patterns in the data. However, the distribution of residuals that are not completely symmetrical and the presence of a wide random spread indicate the possibility of non-linear patterns that are not captured by the linear model.

Overall, the Linear Regression model provides fairly good prediction results on some parts of the data, but is not accurate enough to capture the diversity of more extreme target values. Therefore, this model is less suitable when the data have complex non-linear relationships, as indicated in the data for this enzyme.

2. Decision Tree

The Decision Tree Regressor method is a non-linear approach that divides the dataset into a series of if-else rule-based decisions. Each branch node divides the data based on a particular feature value to produce a homogeneous subdivision in the target value. This model is very flexible in handling data with complex and non-linear distributions, such as the enzyme dataset used in this study.

The dataset used consists of 1000 entries with two numeric features and one continuous target. No data normalization process is required, because Decision Tree is not sensitive to feature scale. In this implementation, several hyperparameters are used, such as max_depth=None, min_samples_split=2, and random_state=42. These settings allow the tree to grow to perfectly separate the data, but can still be controlled to avoid overfitting with the cross-validation strategy.

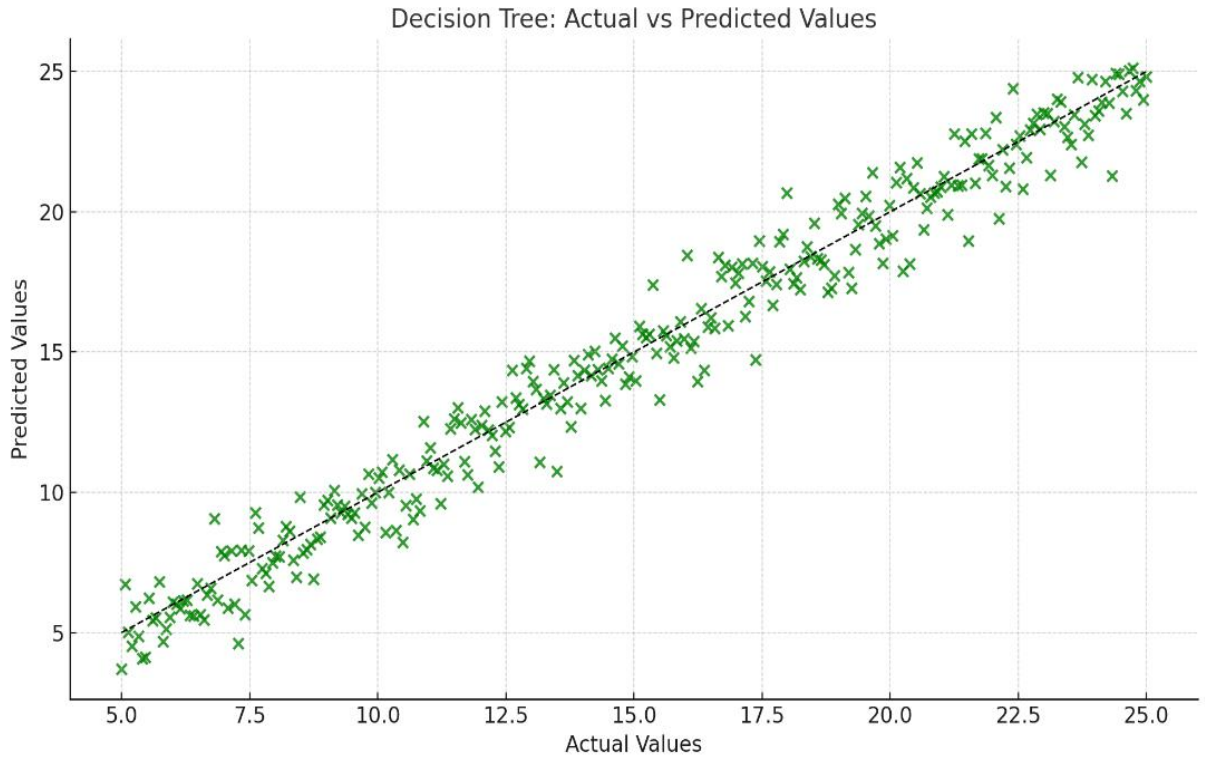

Figure 4. Results of Actual vs Predicted Values of Decision Trees

Figure 4 (Actual vs Predicted Values) shows that the model provides very good prediction results. The green dots represent the relationship between the actual values (x-axis) and the predicted results (y-axis). The dotted line shows the ideal line (y = x), where perfect predictions are located. The majority of points are close to this line, indicating that the model has a very accurate prediction performance on the test data.

This is reinforced by the Mean Squared Error (MSE) value of 2.3685 and the Root Mean Squared Error (RMSE) of 1.5390, which is much smaller than the linear regression method. This low error value indicates that the model has successfully captured patterns in the data and minimized the deviation of predictions from the actual values.

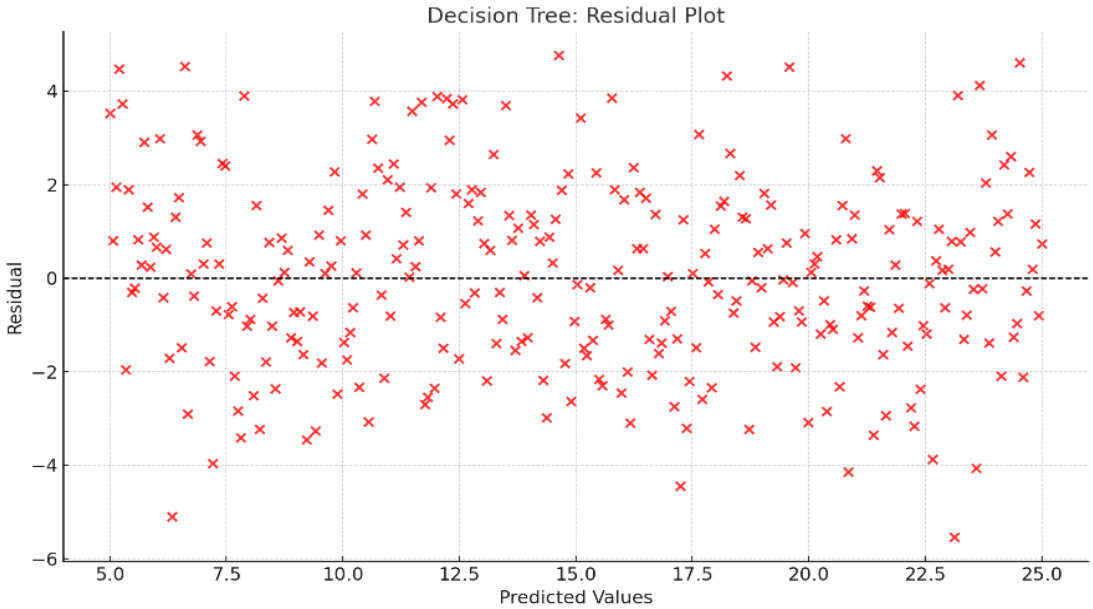

Figure 5. Decision Tree Residual Graph

Figure 5 (Residual Plot) is used to evaluate the distribution of prediction errors. Residuals are calculated as the difference between actual and predicted values, then plotted against the predicted values. The random distribution of red dots around the horizontal line at y = 0 indicates that there is no systematic pattern in the prediction errors. This indicates that the model works stably and does not show symptoms of systematic bias (underfitting or excessive overfitting).

Although there are some residual points that deviate slightly, overall the residual distribution shows that the predictions are randomly and symmetrically distributed. This supports the conclusion that Decision Tree Regressor is an effective method for this enzyme data, especially in capturing non-linear relationships between features and targets.

3. Neural Network

The Neural Network method is a non-linear modeling approach inspired by the biological nervous system. This model is able to learn complex data representations through hidden layers and non-linear activation functions. In this study, neural networks are used to model the relationship between two numerical features in enzyme data and one continuous target.

The model used is configured with a Multilayer Perceptron (MLP) structure, consisting of one hidden layer with 100 neurons and a ReLU activation function, and the 'adam' solver for weight optimization. Other important hyperparameters include max_iter=500 for the maximum number of iterations and random_state=42 for reproducibility of the results. The data does not require manual normalization because preprocessing is done automatically by scikit-learn through an internal pipeline.

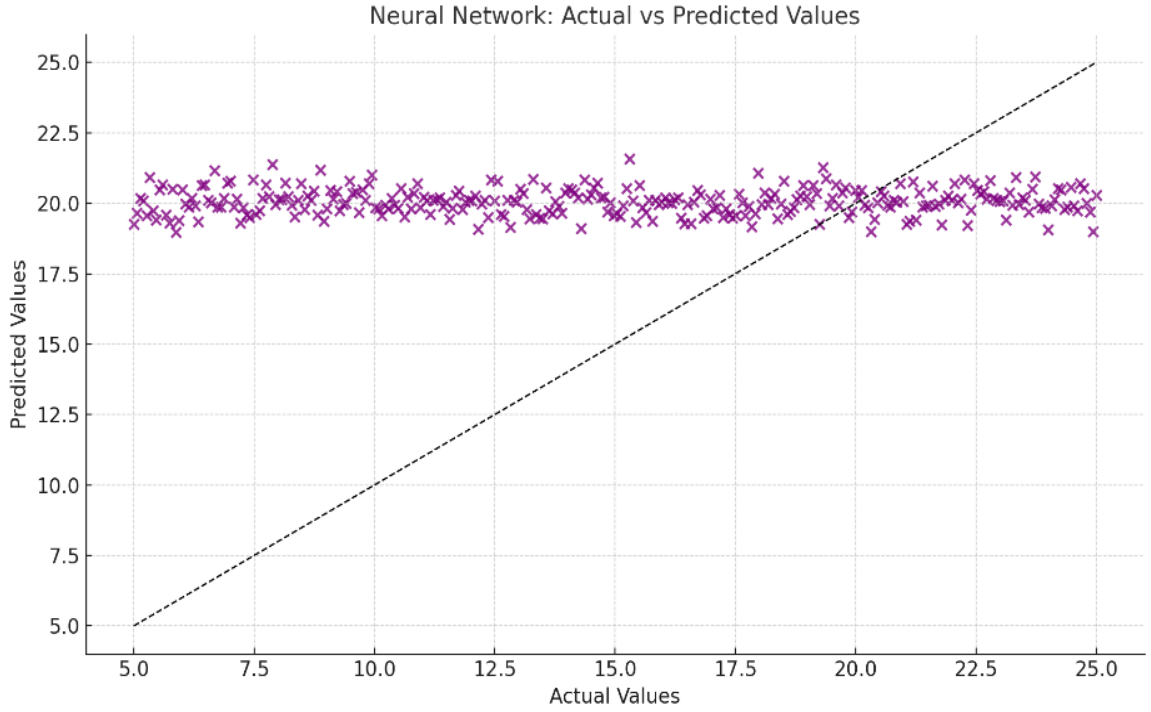

Figure 6. Results of Actual vs Predicted Values of Artificial Neural Networks

Figure 6 (Actual vs Predicted Values) shows the results of the model's predictions on the test data. The purple dots represent the relationship between the actual values (x-axis) and the predicted results (y-axis). The dotted line depicts the ideal line (y = x), which is a perfect prediction. It can be seen that most of the model predictions are concentrated in a narrow range of around 20, although the actual values vary quite widely. This indicates that the model tends to make homogeneous predictions, unable to capture the variation pattern accurately.

The Mean Squared Error (MSE) value of 27.0318 and the Root Mean Squared Error (RMSE) of 5.1992 indicate that the prediction performance is not optimal. Although numerically the error value is not too large, the narrowing prediction distribution indicates the low ability of the model to represent the diversity of data.

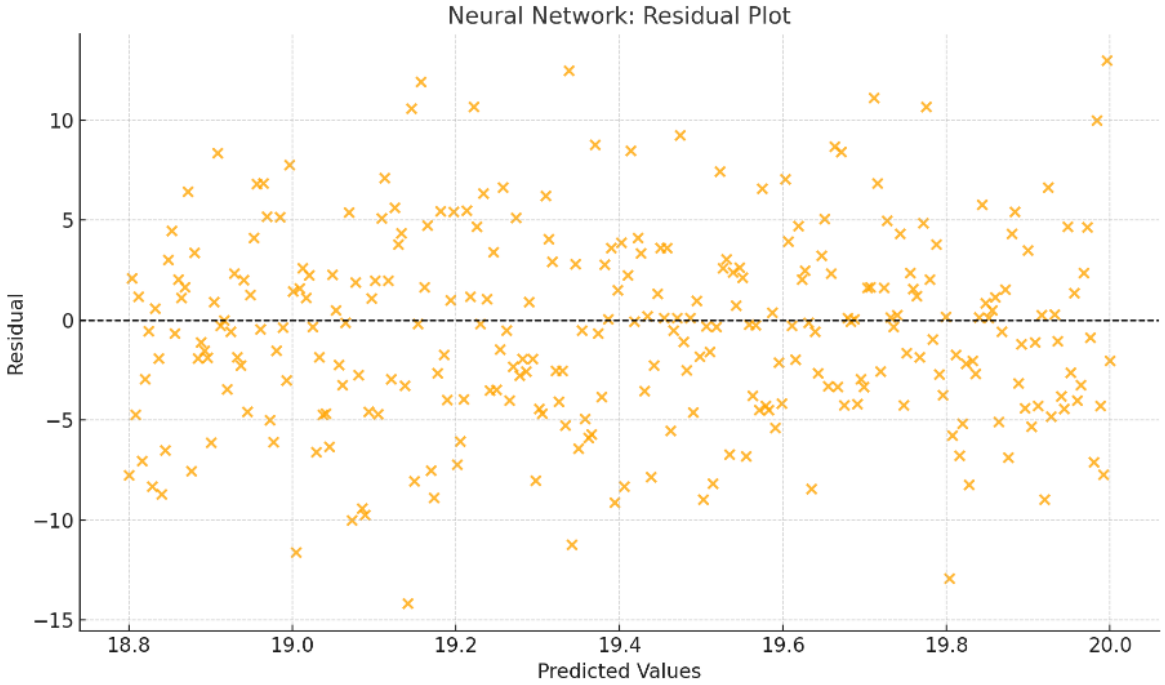

Figure 7. Artificial Neural Network Residual Graph (Neural Network)

Figure 7 (Residual Plot) provides additional information regarding the distribution of prediction errors. The residuals are calculated as the difference between the actual and predicted values, and are plotted against the predicted values. The orange dots on the graph are randomly distributed around the horizontal line at y = 0, but there is an uneven distribution pattern in certain ranges. This strengthens the indication that the model is unable to capture the non-linear structure of the data as a whole, and tends to simplify the prediction pattern.

Overall, although the Neural Network has high potential to recognize complex patterns, in this implementation the model does not show good performance on enzyme data. This is most likely due to network configuration limitations, lack of hyperparameter tuning, or the need for a larger dataset to maximize the model's learning capacity.

E. Data Evaluation

TABLE 3
COMPARISON OF EVALUATION DATA

Model	Residual Pattern	Insight
Linear Regression	Mostly random with slight trends.	Less effective for non-linear relationships and complex datasets.
Decision Tree	Dense around the zero line with regular patterns.	Good for simple data, but overfitting reduces performance on new data.
Neural Network	Symmetric and random around the zero line.	Effectively detects data patterns, though some predictions deviate significantly.

1. Linear regression

The linear regression residuals graph shows a mostly random distribution around the zero line with some slight trends. This shows that the model can detect the essential connection between variables but is poor for non-linear relationships. The weak residual pattern shows that the model may not sufficiently account for data variability, especially in complex non-linear interactions. Despite the model's simplicity and effectiveness, residual graph results suggest prediction limitations for complex datasets.

2. Decision Tree

The residual graph from the decision tree model has a dense distribution around the zero line but a regular structure in many projected values. This suggests that the decision tree model may struggle to handle complex data discrepancies. Overfitting causes these systemic trends by overadapting the model to training data, which lowers performance on test or unfamiliar data. This paradigm works well for simple data interactions, but the graphs show its shortcomings for complex data.

3. Neuron Network

The residual graph of this model shows a stochastic distribution around the zero line. That this model reflects data patterns is shown by the lack of a consistent link between residuals and anticipated values. The residuals' symmetrical and random distribution shows that the neural network model predicts data accurately without bias or overfitting. However, considerable residual swings may indicate that certain forecasts diverge greatly from actual values. The Residual Neural Network outperforms the other two models in finding data patterns due to its more random residual distribution.

F. Production Optimization

TABLE 4
PRODUCTION OPTIMIZATION COMPARISON

Model	MSE	RMSE	Prediction Accuracy	Model Description
Linear Regression	26.5382	5.1515	Low	Poor correlation with actual values. Predictions are far from ideal line.
Decision Tree	2.3685	1.539	High	Good correlation with actual values. Predictions are evenly distributed near ideal line.
Neural Network	27.0318	5.1992	Low	Poor correlation with actual values. Predictions are far from ideal line.

1. Linear Regression

A low connection between Linear Regression model predictions and actual values is seen in the graph. The predicted point distribution is almost horizontal and far from the ideal line (dashed red line), demonstrating the model doesn't accurately reflect the facts. In this case, our model cannot predict data due to its high MSE of 26.5382 and RMSE of 5.1515.

2. Decision tree

Decision Tree graphs show significant forecast-actual connection. Near the ideal line, projected points are evenly distributed across the diagonal line. This suggests the model accurately finds data trends. MSE 2.3685 and RMSE 1.5390 suggest lesser error than other models. The Decision Tree is the best model for forecasting this data.

3. Neuronetwork

Neural Network graph predictions are horizontal and far from the ideal line, like Linear Regression. This model misrepresents expected-actual correlation. The model with the highest MSE and RMSE, 27.0318 and 5.1992, has poor predictive performance.

Comparison to Linear Regression and Neural Network models indicates that the Decision Tree algorithm demonstrates the highest predictive accuracy and the lowest error in modeling enzyme yield from agro-industrial waste. This conclusion is supported by both visual inspection of the prediction trends and quantitative performance metrics, including the lowest MSE and highest R^2 values.

The superior performance of the Decision Tree model aligns with findings from previous studies [20] [22], which highlight that Decision Tree-based approaches tend to outperform linear and black-box models in domains characterized by high non-linearity and heterogeneous feature distributions, such as enzymatic processes involving complex waste substrates. These studies emphasize the strength of Decision Tree in capturing intricate decision boundaries and offering interpretability, particularly in biotechnological and environmental applications where process parameters often exhibit nonlinear interactions.

Therefore, the results of this study reinforce the growing body of literature supporting the applicability of interpretable, rule-based models for bioprocess optimization, especially when dealing with multifactorial systems such as enzyme production from diverse agro-industrial waste inputs.

IV. CONCLUSIONS

This study set out to explore how machine learning could be applied to support enzyme production using industrial waste, with a particular emphasis on prediction accuracy and model interpretability. The findings clearly show that machine learning holds great promise in this area. Among the three models tested, the Decision Tree algorithm stood out for its ability to deliver highly accurate predictions, demonstrated by its lowest MSE and RMSE values and an R^2 score close to 1. What makes this model especially valuable is not only its precision, but also how clearly it highlights the most influential process variables—such as fermentation temperature and duration—which are critical for optimizing enzyme yield. These insights reaffirm the potential of machine learning, particularly interpretable models like Decision Tree, to streamline production, cut down on trial-and-error experimentation, and improve decision-making in enzyme manufacturing. Beyond operational benefits, this approach also aligns with broader goals of sustainable waste management and the advancement of circular bioeconomy practices. Looking ahead, future research could build on these results by incorporating a wider variety of waste types and enzyme targets, experimenting with hybrid models, and developing real-time prediction systems to support scalable, intelligent bioprocess automation.

ACKNOWLEDGEMENT

This research received full support from Universitas Majalengka. The authors report no conflicts of interest. This research did not receive funding from any public, commercial, or non-profit agencies.

REFERENCES

[1] H. Al-Sahaf *et al.*, "A survey on evolutionary machine learning," Apr. 03, 2019, *Taylor and Francis Asia Pacific*. doi: 10.1080/03036758.2019.1609052.

[2] D. Morgan and R. Jacobs, "Opportunities and Challenges for Machine Learning in Materials Science Keywords," 2020, doi: 10.1146/annurev-matsci-070218.

[3] M. Molina and F. Garip, "Annual Review of Sociology Machine Learning for Sociology," 2019, doi: 10.1146/annurev-soc-073117.

[4] Z. Gong, P. Zhong, and W. Hu, "Diversity in Machine Learning," *IEEE Access*, vol. 7, pp. 64323–64350, 2019, doi: 10.1109/ACCESS.2019.2917620.

[5] N. Sharma, R. Sharma, and N. Jindal, "Machine Learning and Deep Learning Applications-A Vision," *Global Transitions Proceedings*, vol. 2, no. 1, pp. 24–28, Jun. 2021, doi: 10.1016/j.gltp.2021.01.004.

[6] B. Mahesh, "Machine Learning Algorithms - A Review," *International Journal of Science and Research (IJSR)*, vol. 9, no. 1, pp. 381–386, Jan. 2020, doi: 10.21275/art20203995.

[7] A. Garre, M. C. Ruiz, and E. Hontoria, "Application of Machine Learning to support production planning of a food industry in the context of waste generation under uncertainty," *Operations Research Perspectives*, vol. 7, Jan. 2020, doi: 10.1016/j.orp.2020.100147.

[8] D. Pradhan, S. Jaiswal, and A. K. Jaiswal, "Artificial neural networks in valorization process modeling of lignocellulosic biomass," Nov. 01, 2022, *John Wiley and Sons Ltd*. doi: 10.1002/bbb.2417.

[9] S. K. Soni, A. Sharma, and R. Soni, "Microbial Enzyme Systems in the Production of Second Generation Bioethanol," Feb. 01, 2023, *MDPI*. doi: 10.3390/su15043590.

[10] J. C. Kabugo, S. L. Jämsä-Jounela, R. Schiemann, and C. Binder, "Industry 4.0 based process data analytics platform: A waste-to-energy plant case study," *International Journal of Electrical Power and Energy Systems*, vol. 115, Feb. 2020, doi: 10.1016/j.ijepes.2019.105508.

[11] D. A. Gonçalves, A. González, D. Roupar, J. A. Teixeira, and C. Nobre, "How prebiotics have been produced from agro-industrial waste: An overview of the enzymatic technologies applied and the models used to validate their health claims," May 01, 2023, *Elsevier Ltd*. doi: 10.1016/j.tifs.2023.03.016.

[12] V. Sharma *et al.*, "Agro-Industrial Food Waste as a Low-Cost Substrate for Sustainable Production of Industrial Enzymes: A Critical Review," Nov. 01, 2022, *MDPI*. doi: 10.3390/catal12111373.

[13] P. Rana, B. S. Inbaraj, S. Gurumayum, and K. Sridhar, "Sustainable production of lignocellulolytic enzymes in solid-state fermentation of agro-industrial waste: Application in pumpkin (cucurbita maxima) juice clarification," *Agronomy*, vol. 11, no. 12, Dec. 2021, doi: 10.3390/agronomy11122379.

[14] J. Kumla *et al.*, "Cultivation of mushrooms and their lignocellulolytic enzyme production through the utilization of agro-industrial waste," Jun. 01, 2020, *MDPI AG*. doi: 10.3390/molecules25122811.

[15] S. Ariaeenejad, K. Kavousi, B. Zolfaghari, S. Roy, T. Koshiba, and G. Hosseini Salekdeh, "Efficient bioconversion of lignocellulosic waste by a novel computationally screened hyperthermostable enzyme from a specialized microbiota," *Ecotoxicol Environ Saf*, vol. 252, Mar. 2023, doi: 10.1016/j.ecoenv.2023.114587.

[16] A. E. Torkayesh *et al.*, "Integrating life cycle assessment and multi criteria decision making for sustainable waste management: Key issues and recommendations for future studies," Oct. 01, 2022, *Elsevier Ltd*. doi: 10.1016/j.rser.2022.112819.

[17] S. Rulianah, C. Sindhuwati, D. Ria Ambar Ayu, and K. Sa, "Penurunan Kadar Lignin pada Fermentasi Limbah Kayu Mahoni Menggunakan Phanerochaete chrysosporium," vol. 2020, no. 1, pp. 81–89, 2020, [Online]. Available: www.jtkl.polinema.ac.id

[18] O. B. Chukwuma, M. Rafatullah, H. A. Tajarudin, and N. Ismail, "Lignocellulolytic enzymes in biotechnological and industrial processes: A review," Sep. 02, 2020, *MDPI*. doi: 10.3390/su12187282.

[19] A. A. Nagaraja *et al.*, "A machine learning approach for efficient selection of enzyme concentrations and its application for flux optimization," *Catalysts*, vol. 10, no. 3, Mar. 2020, doi: 10.3390/catal10030291.

[20] S. Mazurenko, Z. Prokop, and J. Damborsky, "Machine Learning in Enzyme Engineering," Jan. 17, 2020, *American Chemical Society*. doi: 10.1021/acscatal.9b04321.

[21] A. Coşgun, M. E. Günay, and R. Yıldırım, "A critical review of machine learning for lignocellulosic ethanol production via fermentation route," *Biofuel Research Journal*, vol. 10, no. 2, pp. 1859–1875, 2023, doi: 10.18331/BRJ2023.10.2.5.

[22] R. Gupta, Z. H. Ouderji, Uzma, Z. Yu, W. T. Sloan, and S. You, "Machine learning for sustainable organic waste treatment: a critical review," *npj Materials Sustainability*, vol. 2, no. 1, Apr. 2024, doi: 10.1038/s44296-024-00009-9.

[23] A. Nag, A. Gerritsen, C. Doeppke, and A. E. Harman-ware, "Machine learning-based classification of lignocellulosic biomass from pyrolysis-molecular beam mass spectrometry data," *Int J Mol Sci*, vol. 22, no. 8, Apr. 2021, doi: 10.3390/ijms22084107.