CENTIVE
Conference on Electrical Engineering, Informatics,
Industrial Technology, and Creative Media
ISBN: 978-602-53004-0-0

Optimization of Random Forest Model with Correlation-Based Feature Selection for Enhanced Forest Health Prediction

Singgih Setia Andiko[1], Singgih Briandoko[2], Bayu Rizkya Pratama[3], Muhammad Akbar Setiawan[4], Eldas Puspita Rini [5]
[1] STMIK Widya Utama
[1]Jl. Sunan Kalijaga, Purwokerto Selatan Banyumas, Jawa Tengah, 53146, Indonesia
* singgih904@swu.ac.id, briandokosinggih@swu.ac.id , bayurizkyapratama@swu.ac.id, akbar@swu.ac.id, eldas@swu.ac.id

Abstract — Forest health serves as a key indicator for maintaining ecosystem sustainability and biodiversity. This study aims to predict forest health status using a Random Forest algorithm integrated with Correlation-Based Feature Selection (CFS). The dataset comprises 1,000 samples with 18 attributes—including Disturbance_Level, Fire_Risk_Index, Tree_Height, and Menhinick_Index—along with health status labels categorized into four classes: Unhealthy, Sub-Healthy, Healthy, and Very Healthy. The research methodology encompassed data preprocessing, feature selection using CFS, Random Forest model construction, and performance evaluation. Feature selection identified four key attributes that significantly contributed to forest health prediction. The model was trained on 70% of the data and tested on the remaining 30%, achieving an accuracy of 92%. Further analysis revealed an average precision of 91%, recall of 90%, and F1-score of 90%. The confusion matrix indicated accurate predictions across most categories, though some misclassification occurred in the Sub-Healthy class. This study demonstrates that the CFS-based Random Forest approach is effective for forest health prediction, offering a valuable analytical tool to support conservation efforts and damage risk mitigation.

Keywords: Forest Health Prediction, Random Forest, Feature Selection, Correlation-Based Feature Selection, Ecological Informatics, Predictive Modeling

I. INTRODUCTION

The preservation of ecological diversity and forest integrity is fundamental to environmental sustainability and ecosystem functionality. Forests in optimal health serve dual critical roles: they act as biodiversity reservoirs by supporting diverse species populations, and function as climate regulators through carbon sequestration processes. Conversely, forest degradation precipitates ecosystem service loss, endangers species survival, and intensifies climate change impacts. A defining attribute of resilient forests is their adaptive capacity to withstand and recover from ecological disturbances. Contemporary research has recorded pervasive declines in forest health across biomes, attributing these trends to compound stressors including climatic extremes [1], pathogen vulnerabilities [2], wildfire-induced tree mortality [3], and anthropogenic pressures like defaunation and selective logging in tropical systems. Elevated tree mortality rates have been established as a primary indicator of forest health deterioration [4].

The development of predictive frameworks for forest health assessment has consequently emerged as a priority in conservation science. Precise health evaluation empowers stakeholders—ranging from policymakers to conservation practitioners—to implement targeted interventions against forest degradation. Through integrative analysis of multidimensional parameters (including edaphic properties, dendrometric measurements, and fire risk indicators), preemptive identification of vulnerable forest zones becomes feasible. This enables responsive actions such as soil enrichment protocols or protective zoning when early warning signs manifest. Enhanced prediction accuracy facilitates precision conservation, optimizing resource allocation toward critical areas while maximizing ecological outcomes. Ultimately, data-driven forest management approaches contribute to prolonged ecosystem viability, biodiversity maintenance, and socio-ecological resilience for forest-dependent communities.

Advancements in information technology have positioned data mining and machine learning as

transformative methodologies for forest health analytics. The paradigm of Big Data has enabled the extraction of meaningful patterns from complex ecological datasets through statistical and computational approaches [6]. Such analytical capabilities provide conservation entities with actionable insights regarding determinant factors influencing forest conditions [7], thereby revolutionizing traditional monitoring practices [5].

This investigation employs a integrated analytical framework combining Correlation-Based Feature Selection (CFS) with Random Forest classification to evaluate forest health states within the "Forest Health and Ecological Diversity" dataset. The methodology operates through two constitutive phases: CFS-mediated identification of predictor attributes most strongly associated with health status, followed by Random Forest model construction using optimized feature subsets. Model validation incorporates confusion matrix analysis under a 70:30 data partitioning scheme for training and testing respectively.

The study's dual objectives comprise: (1) quantitative evaluation of CFS efficacy in enhancing Random Forest prediction accuracy, and (2) critical analysis of model performance characteristics and constraints within forest health classification contexts.

II. LITERATURE REVIEW

Existing literature highlights several pivotal considerations in methodological design. Primarily, implementing optimized data preprocessing techniques remains imperative [8]. Equally critical is the strategic selection of relevant features, complemented by the judicious choice of machine learning algorithms tailored for forest health modeling [7].

Demonstrated the application of Naïve Bayes integrated with Correlation-Based Feature Selection (CFS) to predict academic performance. Their investigation identified key determinants of student achievement, establishing that rigorous data preprocessing and CFS-mediated feature selection substantially enhanced predictive accuracy by preserving optimal attributes [10].

In a complementary study, evaluated Modified Balanced Random Forest (MBRF) for handling class imbalance in mental disorder classification. Their comparative analysis revealed MBRF's superiority over conventional Random Forest across accuracy, precision, recall, and F1-score metrics. Interestingly, CFS implementation did not yield significant improvements; instead, MBRF without feature selection achieved superior precision-F1 balance, suggesting inherent model efficiency without feature reduction [11].

III. RESEARCH METHOD

A. Research Stages

The research methodology in this study consists of several key stages, beginning with data collection. This is followed by data preprocessing, which includes data cleaning, labeling, transformation, and feature selection using the Correlation-Based Feature Selection (CFS) method. Once the data is prepared, a predictive model is constructed using the Random Forest algorithm. This model is then evaluated using a Confusion Matrix, followed by result analysis and conclusion drawing. The research workflow is illustrated in Figure 1.

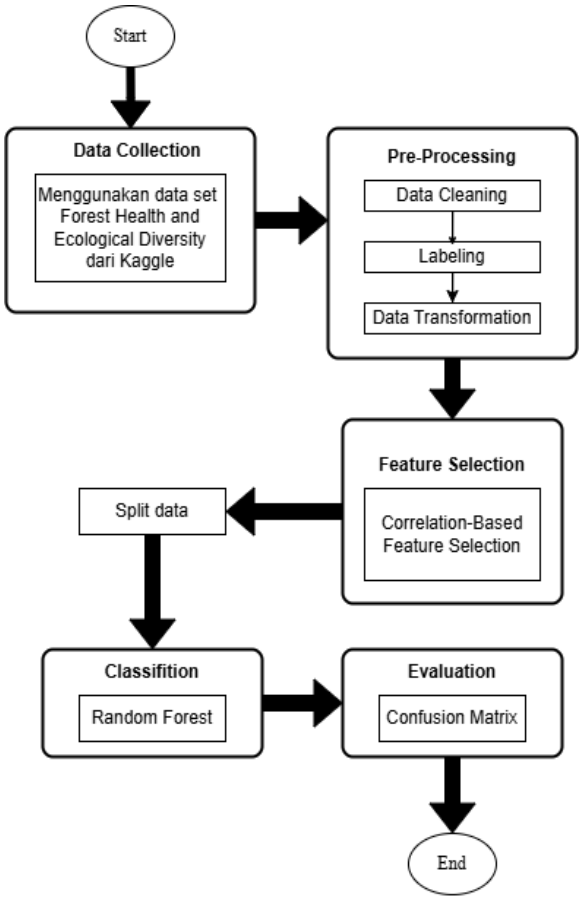

Figure 1. Research Methodology

B. Data Collection

The data collected will be used to train the model implemented in this study. The required dataset encompasses various factors influencing forest health. This data was obtained from Kaggle.com as a publicly available dataset. The dataset comprises four classification categories: Unhealthy (581 instances), Sub-Healthy (314 instances), Healthy (89 instances), and Very Healthy (16 instances), totaling 1,000 data points.

Table 1 Presents The Attribute Information Of The Dataset That Will Be Utilized For Classification.

Atribut	Deskripsi
Plot_ID	Unique identifier for each measurement plot
Latitude	Geographic latitude of the plot in degrees (indicating north-south position)
Longitude	Geographic longitude of the plot in degrees (indicating east-west position)
DBH	Tree diameter at breast height (1.3m above ground), measured in centimeters
Tree_Height	Total tree height from base to crown, measured in meters
Crown_Width_North_South	Crown width measured along north-south direction, in meters
Crown_Width_East_West	Crown width measured along east-west direction, in meters
Slope	Steepness of the terrain in degrees
Elevation	Plot elevation above sea level, measured in meters
Soil_TN	Total Nitrogen concentration in soil (g/kg)
Soil_TP	Total Phosphorus concentration in soil (g/kg)
Soil_AP	Available Phosphorus content in soil (g/kg)
Soil_AN	Available Nitrogen content in soil (g/kg)
Menhinick_Index	Species diversity index reflecting species richness in the area
Gleason_Index	Diversity index accounting for species abundance and richness
Disturbance_Level	Categorical variable indicating ecological disturbance level (0: Low, 1: Medium, 2: High)
Fire_Risk_Index	Probability of fire occurrence based on environmental conditions (scored 0-1)
Health_Status	Categorical variable indicating tree health classification: Very Healthy, Healthy, Sub-Healthy, Unhealthy

C. Preprocessing Data

Data preprocessing was conducted on the data obtained from the data collection stage. The objective of this preprocessing is to prepare the data for model processing, enabling effective classification [11].

a) Data Cleaning

The initial step involved data cleaning, which aims to eliminate duplicate entries and address missing values within the dataset [12]. During this phase, the "Plot_ID" column was removed as it solely contained plot identification numbers and was not utilized for subsequent analytical processes.

b) Data Labelling

Following data cleaning, label encoding was applied to convert categorical data into numerical values, ensuring compatibility with the model. Overall, label encoding was implemented to transform the categorically formatted Forest Health and Ecological Diversity dataset into numerically analyzable data while preserving data structure efficiency and relevance [13]. Specifically, the "Health_Status" attribute was designated as the label for the dataset. This attribute describes forest health categories—Unhealthy, Sub-Healthy, Healthy, and Very Healthy—as outlined in Table 2.

Table 2. Data Lebeling

Atribute	Label
Health_Status	Unhealthy
	Sub-Healthy
	Healthy
	Very Healthy

c) Data transformation

Data transformation is a crucial step to modify data formats to meet the requirements of data mining methods [14], After data labeling, categorical data types are converted into numerical types through Label Encoding. Label Encoding is a technique that converts categorical or ordinal data into numerical values by assigning numeric codes to each category or level. All text-based data tables will be transformed into numerical representations, where text values are expressed as numbers indicating variations or hierarchies [15]. This process ensures the data is prepared for analysis and predictive modeling. The results of the data transformation are shown in Table 3.

Table 3. Data Transformation

Atribut	Label	Transformation
Health_Status	Unhealthy	1
	Sub-Healthy	2
	Healthy	0
	Very Healthy	3

d) Correlation Based Feature Selection

Feature selection is a critical step in data analysis, where the most relevant and useful attributes are selected from all available attributes to construct a model. This study proposes a correlation-based approach to reduce the high number of attributes, accelerate computational processes, and identify the optimal feature combinations to enhance model performance during training and evaluation [16]. Correlation measurement is used to examine the relationship between two variables

in a dataset. If two features are unrelated, the correlation value approaches zero; whereas if a relationship exists, the value approaches ±1 [17]. This correlation measurement is essential as it helps understand relationships between variables and identify patterns that may influence analytical outcomes. The correlation between an attribute (x) and target (y) is calculated using Equation 1:

$$Corr_{xy} = \frac{\sum(Xi - \bar{X})(Yi - Y)}{\sqrt{\sum(Xi - \bar{X})^2 (Yi - \bar{Y})^2}} \quad (1)$$

where $Corr_{xy}$ represents the correlation between variables X and Y, and X and Y denote the mean values of X and Y respectively.

e) Data Spliting

In machine learning, data splitting involves dividing the dataset into two main subsets: training data and testing data. The training data is used to train the model using known outcomes, while the testing data evaluates the model's ability to classify data accurately [18]. The proportion between training and testing data is crucial in determining model accuracy; incorrect partitioning can adversely affect accuracy (Musu, Ibrahim, & Heriadi, 2021). In this study, the data was split into 70% for training and 30% for testing.

f) Prediction Model

Random Forest is an ensemble method comprising a collection of decision trees constructed through random processes [19]. The steps in building a Random Forest model include:

1. Drawing *n*-tree bootstrap samples from the dataset.
2. Constructing a decision tree for each bootstrap sample. At each node of the tree, variables are randomly selected for splitting, and the tree is grown until each terminal node meets a specified minimum case threshold.
3. Aggregating the results from all *n*-trees to predict new data, for instance, using majority voting for classification.
4. Calculating the out-of-bag (OOB) error rate based on data not included in the bootstrap samples.

This method was developed through the integration of multiple classifier algorithms and feature extraction techniques.

g) Model Evaluation

Evaluation of the classification method in this study was conducted using a Confusion Matrix. From this matrix, several performance metrics were calculated, including accuracy, precision, recall, and F1-Score. The formulas used for these calculations are provided in Equations 2-5 as follows [20].

$$\text{Accuracy} = \frac{(\text{TP}+\text{TN})}{(\text{TP}+\text{FP}+\text{FN}+\text{TN})}\ 100\% \quad (2)$$

$$\text{Precission} = \frac{\text{TP}}{(\text{TP}+\text{FP})}\ 100\% \quad (3)$$

$$\text{Recall} = \frac{\text{TP}}{(\text{TP}+\text{FN})}\ 100\% \quad (4)$$

$$\text{F1-Score} = 2\frac{\text{Precission x Recall}}{\text{Precission} + \text{Recall}}\ 100\% \quad (5)$$

IV. RESULT

Data analysis in this study was conducted using the Python programming language and the Google Colab platform. The analytical stages included data preprocessing, feature selection, predictive model development, and performance evaluation of the resulting model. At each stage, the study utilized various Python libraries and frameworks, such as Pandas for data manipulation, Scikit-Learn for feature selection and model creation, as well as Matplotlib and Seaborn for data visualization. Google Colab was employed as a cloud-based platform supporting code execution, enabling researchers to access GPU/TPU resources and accelerate large-scale data processing. The use of these tools facilitated an efficient and in-depth data analysis process to achieve optimal research outcomes.

A. Data Preprosesing

During the data preprocessing stage, the Forest Health and Ecological Diversity dataset consisted of 1000 data entries. The initial step involved data cleaning, during which the "Plot_ID" column was removed as it solely contained plot identification numbers where measurements were taken and was not used for subsequent analysis. No duplicate data entries were found at this stage, so no further deletions were required. Following data cleaning, the next step was data labeling, where all 1000 entries were assigned labels—Unhealthy, Sub-Healthy, Healthy, and Very Healthy—under the Health_Status attribute. The distribution was as follows: Unhealthy (581 entries), Sub-Healthy (314 entries), Healthy (89 entries), and Very Healthy (16 entries). The percentage distribution of Health_Status data is illustrated in Figure 2.

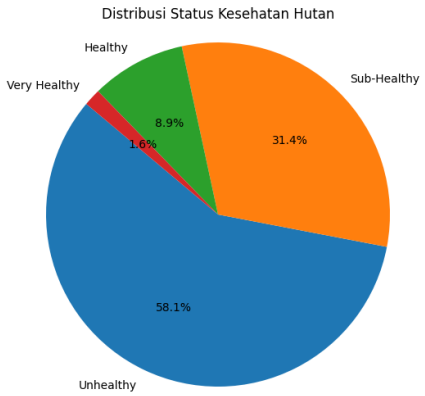

Figure 2. The Percentage Distribution of Health

After preprocessing, feature selection was performed using the correlation-based feature selection method with a threshold of 0.1. From the initial 18 attributes, 4 attributes exhibited correlations above the set threshold with the target variable, as shown in Figure 3.

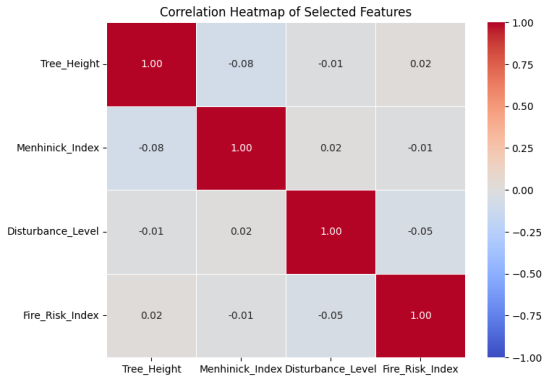

Figure 3. Corelation Heatmap of Selected Features

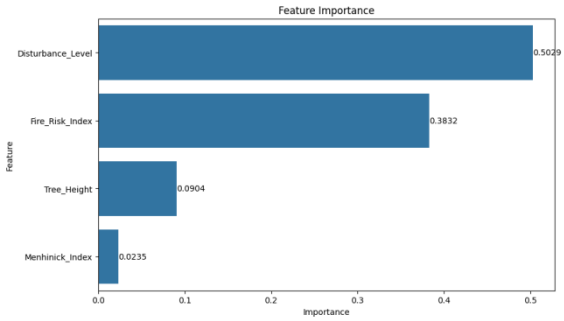

Figure 4. Features Importance

Among these, Disturbance_Level demonstrated the highest positive linear correlation (0.50), indicating a strong relationship with the target variable. This was followed by Fire_Risk_Index (0.38), Tree_Height (0.09), and Menhinick_Index (0.02), as depicted in Figure 4.

Prior to building the predictive model, the data was divided into two subsets: 70% for training data and 30% for testing data. The detailed distribution of data across categories is presented in Table 4.

Table 4. Total of Training and Testing Data

Data	Health_Status			
	Unhealthy	Sub-Healthy	Healthy	Very Healthy
Data Training	401	228	61	10
Data Testing	180	86	28	6

The table shows the class distribution in the training data, consisting of 401 Unhealthy entries, 228 Sub-Healthy entries, 61 Healthy entries, and 10 Very Healthy entries. The testing data comprised 180 Unhealthy entries, 86 Sub-Healthy entries, 28 Healthy entries, and 6 Very Healthy entries.

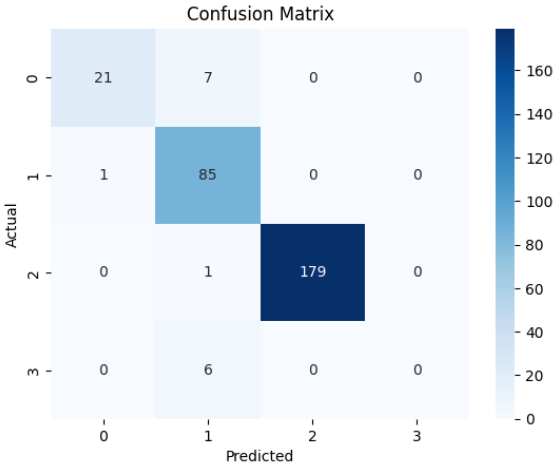

Figure 5. Confusion Matrix Before Feature Selection

Figure 5 displays the confusion matrix of the Random Forest predictions before feature selection. The results indicate that the model performed well for Class 1 and Class 2, with a high number of correct predictions. However, Class 0 was frequently misclassified as Class 1, and Class 3 was poorly predicted, with all instances incorrectly classified as Class 1.

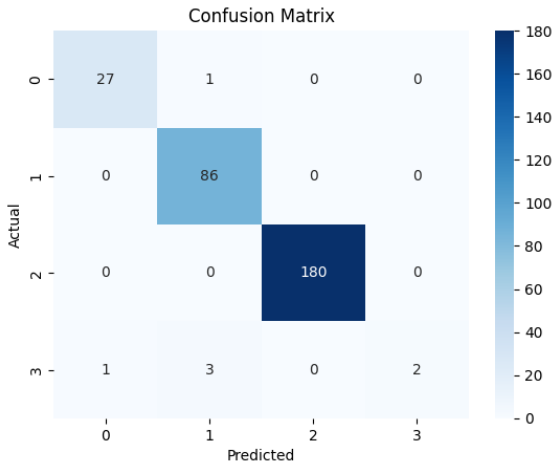

Figure 6. Confusion Matrix With Feature Selection

Figure 6 presents the confusion matrix of the Random Forest predictions after feature selection. The model demonstrated significantly improved performance for Class 1 and Class 2, with nearly all predictions being correct. For Class 0, only one misclassification occurred. Class 3, although still challenging to predict, showed some correct classifications, indicating partial improvement. Overall, the model exhibits strong performance but requires further refinement for Class 3.

Table 5. Comparison of Random Forest Evaluation Metrics Before and After Feature Selection

Feature Selection	**Random Forests**			
	Accuracy	Recall	Precission	F1-Score
Before	95%	95%	94%	94%
After	98,33%	98%	98%	98%

Table 5 presents a comparative analysis of the predictive model's evaluation metrics—accuracy, recall, precision, and F1-score—before and after feature selection. The results demonstrate a significant improvement in model performance following feature selection. Model accuracy increased from 95% to 98.33%. Additionally, other metrics also showed positive improvements: recall increased from 95% to 98%, precision from 94% to 98%, and F1-score from 94% to 98%. Thus, it can be concluded that the Correlation-Based Feature Selection (CFS) method is highly effective in enhancing the performance of the predictive model.

V. CONCLUSIONS

This study successfully identified and predicted forest health status using a Random Forest algorithm enhanced with Correlation-Based Feature Selection (CFS). Through preprocessing stages—including data cleaning, labeling, data transformation, and relevant feature selection—the constructed model demonstrated significant performance improvements. Evaluation results revealed that after feature selection, accuracy, recall, precision, and F1-score increased substantially, each rising from 95% to 98.33%. These findings indicate that correlation-based feature selection effectively enhances prediction accuracy and can be utilized to improve forest health monitoring and preventive conservation planning.

VI. RECOMMENDATION

Recommendations derived from this study include: expanding dataset diversity to enhance model generalizability; employing alternative machine learning algorithms such as Gradient Boosting or XGBoost for comparative analysis with Random Forest; and implementing oversampling techniques like SMOTE to address class imbalance. The integration of spatial data (e.g., satellite imagery or LiDAR) and the inclusion of additional features such as weather patterns or human disturbance indicators could further improve prediction accuracy. Model validation with field data is essential to ensure real-world applicability, while collaboration with stakeholders may facilitate the integration of research findings into forest conservation policies. Future studies could also explore the effectiveness of CFS compared to other feature selection methods, as well as develop real-time forest health prediction systems leveraging IoT sensors and intuitive user interfaces to enhance practical usability.

REFERENCES

[1] C. I. Millar and N. L. Stephenson, “Temperate forest health in an era of emerging megadisturbance,” *Science*, vol. 349, no. 6250, pp. 823–826, Aug. 2015, doi: 10.1126/science.aaa9933.

[2] S. Gauthier, P. Bernier, T. Kuuluvainen, A. Z. Shvidenko, and D. G. Schepaschenko, “Boreal forest health and global change,” *Science*, vol. 349, no. 6250, pp. 819–822, Aug. 2015, doi: 10.1126/science.aaa9092.

[3] M. J. Wingfield, E. G. Brockerhoff, B. D. Wingfield, and B. Slippers, “Planted forest health: The need for a global strategy,” *Science*, vol. 349, no. 6250, pp. 832–836, Aug. 2015, doi: 10.1126/science.aac6674.

[4] S. Trumbore, P. Brando, and H. Hartmann, “Forest health and global change,” *Science*, vol. 349, no. 6250, pp. 814–818, Aug. 2015, doi: 10.1126/science.aac6759.

[5] G. Feng, M. Fan, and Y. Chen, “Analysis and Prediction of Students’ Academic Performance Based on Educational Data Mining,” *IEEE Access*, vol. 10, pp. 19558–19571, Jan. 2022, doi: 10.1109/access.2022.3151652.

[6] Y. Mardi, “Data Mining : Klasifikasi Menggunakan Algoritma C4.5,” *Edik Informatika*, vol. 2, no. 2, pp. 213–219, Feb. 2017, doi: 10.22202/ei.2016.v2i2.1465.

[7] M. Sudais, M. Safwan, M. A. Khalid, and S. Ahmed, “Students’ Academic Performance Prediction Model Using Machine Learning,” *Research Square (Research Square)*, Jan. 2022, doi: 10.21203/rs.3.rs-1296035/v1.

[8] C. S. K and K. S. Kumar, “Data Preprocessing and Visualizations Using Machine Learning for Student Placement Prediction,” *2022 2nd International Conference on Technological Advancements in Computational Sciences (ICTACS)*, Oct. 2022, doi: 10.1109/ictacs56270.2022.9988247.

[9] P. T. P. P, G. Lumacad, and R. Catrambone, “Predicting Student Performance Using Feature Selection Algorithms for Deep Learning Models,” *2021 XVI Latin American Conference on Learning Technologies (LACLO)*, vol. 28, pp. 1–7, Oct. 2021, doi: 10.1109/laclo54177.2021.00009.

[10] T. Gori, A. Sunyoto, and H. A. Fatta, "Preprocessing Data dan Klasifikasi untuk Prediksi Kinerja Akademik Siswa," *Jurnal Teknologi Informasi Dan Ilmu Komputer*, vol. 11, no. 1, pp. 215–224, Feb. 2024, doi: 10.25126/jtiik.20241118074.

[11] N. Arsad, A. H. Muhammad, and T. Hidayat, "Classification of Mental Disorders Using Modified Balanced Random Forest And Feature Selection," *Jurnal Teknologi Informasi Universitas Lambung Mangkurat (JTIULM)*, vol. 9, no. 2, pp. 45–54, Oct. 2024, doi: 10.20527/jtiulm.v9i2.320.

[12] L. F. Kholig, S. Supriadi, M. Andri, T. Erviyanti, and V. Oktavianti, "Pembinaan Kesehatan Mental Remaja Di MTS Ngalaban Desa Bendet Kecamatan Diwek Jombang," *Jurnal Pengabdian Masyarakat Darul Ulum*, vol. 1, no. 1, pp. 45–51, Jan. 2022, doi: 10.32492/dimas.v1i1.522.

[13] M. K. Dahouda and I. Joe, "A Deep-Learned Embedding Technique for Categorical Features Encoding," *IEEE Access*, vol. 9, pp. 114381–114391, Jan. 2021, doi: 10.1109/access.2021.3104357.

[14] H. Henderi, "Comparison of Min-Max normalization and Z-Score Normalization in the K-nearest neighbor (kNN) Algorithm to Test the Accuracy of Types of Breast Cancer," *IJIIS International Journal of Informatics and Information Systems*, vol. 4, no. 1, pp. 13–20, Mar. 2021, doi: 10.47738/ijiis.v4i1.73.

[15] J. K. Lubis and I. Kharisudin, "Metode Long Short Term Memory dan Generalized Autoregressive Conditional Heteroscedasticity untuk Pemodelan Data Saham," Feb. 23, 2021. https://journal.unnes.ac.id/sju/index.php/prisma/article/view/44897

[16] E. S. Alomari *et al.*, "Malware Detection Using Deep Learning and Correlation-Based Feature Selection," *Symmetry*, vol. 15, no. 1, p. 123, Jan. 2023, doi: 10.3390/sym15010123.

[17] H. Zulfiqar, Q.-L. Huang, H. Lv, Z.-J. Sun, F.-Y. Dao, and H. Lin, "Deep-4mCGP: A Deep Learning Approach to Predict 4mC Sites in Geobacter pickeringii by Using Correlation-Based Feature Selection Technique," *International Journal of Molecular Sciences*, vol. 23, no. 3, p. 1251, Jan. 2022, doi: 10.3390/ijms23031251.

[18] N. B. N. Azmi, N. A. Hermawan, and N. D. Avianto, "Analisis Pengaruh Komposisi Data Training dan Data Testing pada Penggunaan PCA dan Algoritma Decision Tree untuk Klasifikasi Penderita Penyakit Liver," *JTIM Jurnal Teknologi Informasi Dan Multimedia*, vol. 4, no. 4, pp. 281–290, Feb. 2023, doi: 10.35746/jtim.v4i4.298.

[19] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, Jan. 2001, doi: 10.1023/a:1010933404324.

[20] N. A. Riska, N. Purnawansyah, H. Darwis, and W. Astuti, "Studi Perbandingan Kombinasi GMI, HSV, KNN, dan CNN pada Klasifikasi Daun Herbal," *Indonesian Journal of Computer Science*, vol. 12, no. 3, Jun. 2023, doi: 10.33022/ijcs.v12i3.3210.

[21] A. Y. Prayoga, A. I. Hadiana, and F. R. Umbara, "Deteksi Hoax pada Berita Online Bahasa Inggris Menggunakan Bernoulli Naïve Bayes dengan Ekstraksi Fitur Tf-Idf," *Jurnal Syntax Admiration*, vol. 2, no. 10, pp. 1808–1823, Oct. 2021, doi: 10.46799/jsa.v2i10.327.