Techno Nusa Mandiri: Journal of Computing and Information Technology
Vol. 22, No. 2 September 2025 | DOI https://doi.org/10.33480/techno.v22i2.6539

OPTIMIZATION OF MACHINE LEARNING ALGORITHMS IN THE CLASSIFICATION OF VECTOR-BORNE DISEASES

Sukrul Ma'mun[1]; Eni Heni Hermaliani[2*]

Magister Ilmu Komputer[1], Sistem Informasi[2]
Universitas Nusa Mandiri, Jakarta, Indonesia[1,2]
https://nusamandiri.ac.id[1,2]
14220019@nusamandiri.ac.id[1], enie_h@nusamandiri.ac.id[2*]
(*) Corresponding Author

Abstract— *Developing a predictive model is the objective of this study, focusing on vector-borne diseases using various machine learning methods, including Random Forest (RF), Logistic Regression (LR), k-nearest Neighbors (kNN), Tree (DT), and XGBoost. The main goal is to use oversampling techniques like SMOTE and Random Oversampling to correct the dataset's class imbalance. The dataset was obtained from Kaggle and literature references published in Frontiers in Ecology and Evolution (Endo and Amarasekare 2022), consisting of approximately 9,490 entries with environmental, demographic, and clinical attributes. Dengue Fever is one of the diseases that this study focuses on. Aedes aegypti mosquitoes spread it, and it is a significant health risk in tropical areas. The DT and XGBoost models had the highest accuracy, at 99.2%. Logistic Regression and Random Forest also did well, with 99.1% accuracy. KNN did well, too, but with a lower recall, at 99.0%. The ROC curve gave a complete picture of how well each model classified things. These findings indicate that when combined with proper data handling, machine learning models can significantly improve early detection of vector-borne diseases and support more accurate and timely decision-making in public health interventions.*

Keywords: *Disease Control, Disease Prediction, Machine Learning (ML), SMOTE, Vector-Borne Diseases.*

Intisari— Mengembangkan model prediktif adalah tujuan dari penelitian ini, dengan fokus pada penyakit yang ditularkan melalui vektor menggunakan berbagai metode pembelajaran mesin, termasuk Random Forest (RF), Logistic Regression (LR), k-nearest Neighbors (kNN), Tree (DT), dan XGBoost. Tujuan utamanya adalah menggunakan teknik oversampling seperti SMOTE dan Random Oversampling untuk mengoreksi ketidakseimbangan kelas dataset. Dataset diperoleh dari Kaggle dan referensi literatur yang diterbitkan dalam di Frontiers in Ecology and Evolution (Endo and Amarasekare 2022), yang terdiri dari sekitar 9.490 entri dengan atribut lingkungan, demografi, dan klinis. Demam Berdarah adalah salah satu penyakit yang menjadi fokus penelitian ini. Nyamuk Aedes aegypti menyebarkannya, dan merupakan risiko kesehatan yang signifikan di daerah tropis. Model DT dan XGBoost memiliki akurasi tertinggi, yaitu 99,2%. Logistic Regression dan Random Forest juga bekerja dengan baik, dengan akurasi 99,1%. KNN juga berhasil, tetapi dengan perolehan kembali yang lebih rendah, yaitu 99,0%. Kurva ROC memberikan gambaran lengkap tentang seberapa baik setiap model mengklasifikasikan berbagai hal. Temuan ini menunjukkan bahwa bila dikombinasikan dengan penanganan data yang tepat, model pembelajaran mesin dapat secara signifikan meningkatkan deteksi dini penyakit yang ditularkan melalui vektor dan mendukung pengambilan keputusan yang lebih akurat dan tepat waktu dalam intervensi kesehatan masyarakat.

Kata Kunci: *Pengendalian Penyakit, Prediksi Penyakit, Pembelajaran Mesin (ML), SMOTE, Penyakit Tular Vektor.*

INTRODUCTION

Vector-borne diseases, such as malaria and dengue fever, remain major public health threats, particularly in tropical regions like Indonesia. These diseases are transmitted through vectors such as mosquitoes, whose populations are heavily influenced by environmental factors, climate, and changes in human behavior (Kumar, Bauch, and Anand 2025). The increasing incidence of vector-borne diseases highlights the need for new

approaches to predicting and controlling their spread (Piscitelli and Miani 2024).

The Canters for Disease Control and Prevention have created the National Electronic Disease Surveillance System in Indonesia in an attempt to address this issue., which aims to accelerate standardized data. However, despite significant advancements in disease surveillance through information technology, major challenges remain in its implementation, particularly regarding the high cost of licensed software and limited resources in remote areas (Babawarun et al. 2024).

With the advancement of information technology, machine learning methods have emerged as a potential solution to improve disease prediction accuracy (Priyono et al. 2023). Machine learning enables more complex data analysis, allowing the identification of disease spread patterns that might be overlooked by conventional methods. However, previous research has shown significant challenges, particularly in class imbalance in vector-borne disease data, where minority cases are often ignored by predictive models. For example, a study by Mokammel Hossain Tito et al. used the same dataset ("vector-borne-disease-prediction") and achieved a maximum accuracy of 92%. However, it failed to optimally address class imbalance, limiting the potential for detecting cases in vulnerable populations (Lin et al. 2025).

This research attempts to close this gap by employing a variety of machine learning methods to create a more precise forecast model for vector-borne illnesses, including Logistic Regression (LR), Random Forest (RF), Decision Tree (DT), k-Nearest Neighbors (kNN), and XGBoost. Additionally, to address class imbalance, this study applies oversampling techniques such as SMOTE and Random Oversampling. As a result, this study not only aims to increase overall accuracy but also strengthens the model's capacity to identify underrepresented situations that were previously missed (Priyono, Ispandi, and Rusdi 2024). The study is expected to significantly improve prediction accuracy, contributing to more effective control of vector-borne disease spread both in Indonesia and globally (Manikandan et al. 2022), (Febiriana, Hassan, and Aldila 2024).

With this approach, the study aims to provide a significant new contribution to the literature, particularly in the context of applying machine learning to predict vector-borne diseases in resource-limited environments (Kamguem et al. 2025). It also demonstrates how more advanced and adaptive machine-learning techniques can be implemented to address urgent public health challenges. A literature review of other studies on this topic can be found in Table 1.

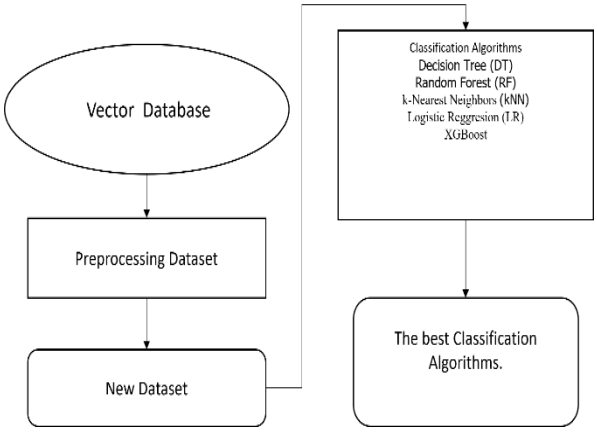
Table 1. Study of The Latest Research Literature

Study	Year	Proposal Method	Accuracy
Mokammel Hossain Tito et al,	2022	MLP, SL dan SVM	92%, 88%, 90.87 %
Micheal Olaolu Arowolo et al, (Ramadhan and Khoirunnisa 2021)	2021	SVM	98,3%
Nur Ghaniaviyanto Ramadhan et al, (Pise and Patil 2023)	2021	SVM	92%
Reshma Pise et al, (Shaikh et al. 2024)	2023	GoogLeNet	96%
Salim Gulab Shaikh et al, (Kundu and Das 2019)	2024	RNN	98.76 %

Source : (Research Results, 2024)

MATERIALS AND METHODS

To perform predictive analysis for the classification of vector-borne diseases, the steps taken to obtain results are shown in Figure 1.



Source : (Research Results, 2024)
Figure 1. Research Methodology

NumPy, Matplotlib, and Scikit-Learn were among the modules utilized in all studies, which were conducted using Python 3.9.12. The Python programming language's NumPy package contains a number of high-level mathematical operations that may be used on huge, multi-dimensional arrays and matrices. A graphing package called Matplotlib is part of NumPy, a Python programming language extension for numerical mathematics. The free machine learning package for Python is called Scikit-learn, formerly known as scikits learn and occasionally called sklearn.

a. Dataset

The initial stage of this research involved gathering literature from the Kaggle dataset: https://www.frontiersin.org/articles/10.3389/fevo.2022.758277/full#supplementary-material. These are the actions that can be documented in the study: data gathering, in which information was gathered using the Kaggle database. The issue of vector-borne disease prediction is the main focus of the study. An oversampling technique was used to solve the class imbalance., especially SMOTE, to handle the dataset's unequal class distribution. The dataset's attributes may comprise different traits or variables that are measured or seen for each sample. These attributes can provide information about the characteristics or subjects, and other relevant factors as shown in Table 2.

Table 2. Dataset Attribute Details.

Attribute Name	Description
SEXO	Refers to a person's gender as stated in the data. Typically, "M" stands for "male" (Masculino), and "F" stands for "female."
IDADE	Refers to a person's age when the information is collected. Typically expressed in years.
DATA	The date when data or notification is recorded or received. The date format usually depends on the standard used, such as DD/MM/YYYY.
REGIÃO	Indicates the geographical region where an individual or related event is located. This can include broad areas such as the East, West, North, South, or Central regions within a specific context.
UF	Abbreviation for "Unidade Federativa," which refers to the federative unit or state in Brazil. It is a two-letter code used to identify each state in Brazil, such as "SP" for São Paulo, "RJ" for Rio de Janeiro, and so on.
MUNICÍPIO	Refers to the city or municipality where an individual or event occurs. It is an administrative division under the state (Estado).
DATA:day_of_month	Date in the month.
DATA:day_of_week	1 to 7 Days in a week.
DATA:half year	1 or 2 Half-years.
DATA:year	Year.
days_diff (DATA, Today)	Difference in days from a specific date to today.

Source : (Research Results, 2024)

b. Preprocessing

When dealing with missing data, one must look for any missing numbers and use the mean value to fill them in. In the preprocessing step, data duplication management entails identifying and eliminating duplicate entries from the dataset. Converting categorical data to numerical format guarantees that Python models can process the complete dataset. Data standardization is performed to prevent certain attributes from dominating, using the Min-Max Normalization method. The data is prepared for use in the machine learning modelling process following processing.

c. New Dataset

In this study, the dataset undergoes rigorous evaluation for reliability. The analysis phase includes identifying and handling missing data. Data duplication is addressed to ensure integrity. Consistency and structure are verified to guarantee the dataset's reliability. By thoroughly implementing these steps, the resulting dataset serves as a trustworthy reference. This procedure assists in the creation of precise and useful vector-borne disease prediction models and improves knowledge of the patterns of disease transmission in the area under question. Researchers can be confident that the dataset used in this study meets high-quality standards, enabling the development of reliable and relevant predictive models for vector-borne diseases. For prediction purposes, the dataset is divided into two sets: input and output variables. Input data is stored as vector **X**, while target data is stored as **y**. The proposed model is utilized for predicting vector-borne diseases using machine learning algorithms.

d. Test Data

After the last pre-processing step is finished, data is saved in CSV (Comma-Separated Values) format and used as input for the classification stage. The data is initially divided into several subsets before the model moves on to the classification stage. Using the sklearn (sci-kit-learn) module, the dataset is separated into training and testing data (Zhang and Chen 2024). in Python 3.9.12. Consequently, we divided the data into two categories: 20% for testing and 80% for training. Because it is a straightforward and efficient method that works well with big datasets, this fraction is selected at random.

e. Machine learning classification methods

We used several well-known machine learning methods, as explained in the next subsection.

1. XGBoost.

A gradient boosting algorithm implementation that makes use of decision trees is called XGBoost. The fundamental procedures and formulas utilized are as follows: Two components make up the objective function that XGBoost minimizes: the regularization function and the loss function (Zhen et al. 2024).

$$\text{Obj}(\theta) = \sum_{i=1}^{n} \text{L}(\text{yi}, \text{y^i}) + \sum_{k=1}^{k} \Omega(\text{fk}) \quad (1)$$

L(yi,y^i) is the loss function
Ω(fk) is the regularization term, which controls model complexity to prevent overfitting.

2. K-Nearest Neighbors (k-NN)

K-NN is a popular nonparametric supervised learning technique. By taking into account its closest neighbors, it creates a boundary to categorize data. Based on their k-nearest neighbors, this method places freshly provided data items in the category with the highest incidence. The prediction depends on the hyperparameter k, which determines the nearest k-NN. Because of the flatter separation curves, models with larger k values are simpler (Dirantara and Sugandi 2025).

$$d(x,y) = \sqrt{\sum_{i=1}^{k} \text{P(H|X)} \ (x_i - y_i)^2} \quad (2)$$

P(H|X), X = proof X, and H = hypothesis = likelihood the proof X supports hypothesis H.
P(H|X) is H likelihood the past under the assumption that X.
P(H|X) the possibility that given hypothesis H, X will occur, or the probability X under the assumption that.
P(H) is proof X prior probability.
H: wT(x) + b = 0
b = the hyperplane equation intersects the hyperplane and a bias term is consistently formatted as a D-1 operator in space with dimensions of D. For example, a linear line in 2-D space is called a hyperplane (1-D).

3. Random Forest (RF)

In the RF machine learning technique, random data choices, generally referred to as "bagging," are used to create ensembles of many individual decision trees, or "random forests." Besides bagging, RF uses random feature selection and random subsets of data to build trees. Within the RF, every tree forecasts a category, and the model predicts the most popular category. (Sobari et al. 2025).

$$Entropy\ (S) = \sum - \text{Pi}\ log2\ (\ )\ n \quad (3)$$

i=1 = Number of partitions S = Set of cases n, fraction of S to S = Pi

4. Decision Tree (DT)

DT, known as a tree diagram, is a graphical depiction of a series of choices or occurrences. It is used to assist in discovering the optimum course of action based on particular conditions or criteria by visualizing the processes in a decision-making process. DT often has practical significance that can be used to aid treatment decisions by drawing reasonable medical conclusions (Saputra and Pratama 2025).

$$E(S) = \sum_{i=1}^{c} \text{Pi} - P_1 log_2 P_1 \quad (4)$$

S = stands for starting condition,
i = arrange a class on S,
Pi = likelihood or a node's share of class I

5. An outcome's probability is expressed as a logistic regression (LR)

A standard supervised machine learning classifier is used to predict the likelihood of an occurrence (in this case, categorizing a person as normal, overweight, or obese) by examining a given set of independent variables (Dirantara and Sugandi 2025).

$$log = \frac{p}{1+p} = a + \beta 1 + x2 \ldots . + \beta i + xi \quad (5)$$

p = probability of an outcome.
a = intercept.
β1 = related coefficient.
I = is the predictor variable's value.

RESULTS AND DISCUSSION

A. Accuracy Results

A predictive model for vector-borne diseases was effectively created in this study by utilizing a number of machine learning methods, such as k-Nearest Neighbors (kNN), Random Forest (RF), Decision Tree (DT), Logistic Regression (LR), and XGBoost. According to the evaluation results, the highest accuracy of 99.2% was attained by DT and XGBoost, while 99.1% was attained by LR and RF. Strong performance was also shown by other algorithms, like kNN, which had 99% accuracy. Table 3 displays the machine learning techniques' performance outcomes used in this investigation.

Table 3. Accuracy Values

Algorithm	Accuracy	Precision	Recall	F1 Score
Decision Tree	0.992	0.881	0.858	0.868
LR	0.991	0.991	0.987	0.989
Random Forest	0.991	0.913	0.884	0.897
kNN	0.99	0.877	0.787	0.828
XG Boost	0.992	0.838	0.786	0.81

Source : (Research Results, 2024)

B. Discussion

This study demonstrates that the use of machine learning algorithms can significantly improve prediction accuracy for vector-borne diseases, such as malaria and dengue fever. Various methods, including Decision Tree (DT), Random Forest (RF), k-Nearest Neighbors (kNN), Logistic Regression (LR), and XGBoost, were tested to predict these diseases. The results show that DT and XGBoost achieved exceptionally high accuracy, reaching 99.2%, representing a significant improvement compared to previous studies, which only reached a maximum accuracy of 92%. This finding suggests that these models are more reliable in predicting disease spread over time.

One important component of this study is the use of oversampling methods to balance the dataset's classes, including SMOTE and Random Oversampling. When the minority class has a substantially lower number of cases than the dominant class, class imbalance is a prevalent problem in epidemiological data. Without proper handling, machine learning models tend to overlook the minority class, leading to inaccurate predictions. By implementing oversampling techniques, this study successfully improved the model's sensitivity to the minority class, allowing for better detection of previously overlooked cases.

Furthermore, ROC Curve model evaluation shows that the used techniques are effective in differentiating between positive and negative situations, which is essential for disease prediction. These outcomes demonstrate the models' superior accuracy as well as their capacity to assist evidence-based decision-making in the management of disease.

This study also emphasizes the importance of utilizing widely accessible information technology, such as open-source software, to support the development of efficient and effective disease prediction systems. By overcoming the limitations of licensing costs and resource constraints, the proposed solution can be implemented across various settings, including resource-limited areas. Vector-borne disease datasets generally exhibit non-linear relationships between features, such as climate conditions, environmental factors, and disease cases. Decision Tree (DT) and XGBoost are highly effective at capturing these non-linear relationships, as they can partition the feature space hierarchically and dynamically. DT is naturally resistant to noise and outliers, as it only processes relevant information for splitting at each node. XGBoost, as an advanced form of boosting trees, is even more powerful because it optimizes the errors from previous trees.

XGBoost also has an internal mechanism for automatic feature selection by assigning greater weights to features that contribute most to predictive performance. This is especially beneficial when the dataset contains numerous features with varying levels of relevance. DT and XGBoost demonstrate a high level of adaptability to datasets processed through oversampling. With techniques like SMOTE or Random Oversampling applied in this study, the number of minority class instances increases, and tree-based algorithms such as DT and XGBoost can leverage these new patterns more effectively than linear algorithms like Logistic Regression.

Models built based on a specific dataset (with particular features such as geographic location, time, temperature, rainfall, and others) may not perform optimally on other datasets with significantly different characteristics. For example, a model trained on vector-borne disease data from Southeast Asia may not be accurate when applied to data from Africa or South America without proper adjustments. The accuracy of the model heavily depends on the completeness and quality of the data used. When applied in regions with incomplete or non-real-time disease reporting systems or environmental data, the model's performance may decline.

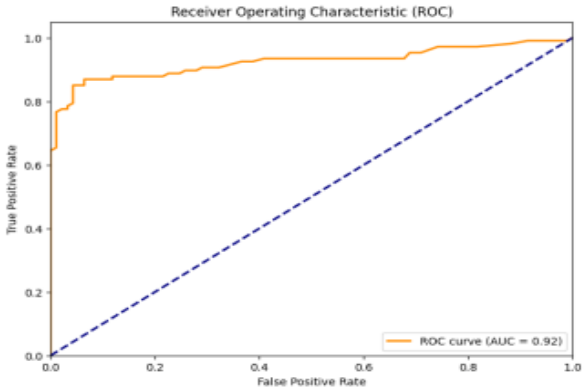
With a very high accuracy (99.2%), there is a possibility that the model is overfitted to the training data. Although validation techniques have been applied, further evaluation using independent test data (external validation) is still necessary to ensure the model's generalizability. Several features in vector-borne disease datasets, such as human mobility patterns or climate changes, are highly dynamic. The developed model may not fully account for temporal or spatial changes if the training data is not sufficiently representative.

C. ROC Curve

The Receiver Operating Characteristic (ROC) curve shows the performance of a classification model at different threshold settings. The connection between the False Positive Rate (FPR) and the True Positive Rate (TPR) at various thresholds is measured by the ROC curve. The True Positive Rate (TPR), also known as recall or sensitivity, evaluates the model's ability to identify

positive cases. The formula for TPR, where TP is True Positive and FN is False Negative, is:
TPR = TP / (TP + FN). The False Positive Rate (FPR) measures how often the model incorrectly classifies negative cases as positive. The formula for FPR is: FPR = FP / (FP + TN) where TN is True Negative, and FP is False Positive.

In this context, an AUC of 0.92 indicates that the model has a strong ability to distinguish between positive cases (vector-borne diseases) and negative cases (non-vector-borne diseases). With an AUC value of 0.92, there is strong evidence that the classification model performs exceptionally well. Therefore, an ROC curve with a high AUC value like this demonstrates good model performance, as shown in Figure 4.



Source : (Research Results, 2024)
Figure 4. ROC Curve

CONCLUSION

Using machine learning techniques, this study developed a highly accurate predictive model to combat vector-borne diseases. The model outperformed previous approaches by applying algorithms like Decision Tree (DT) and XGBoost, along with SMOTE oversampling, especially in detecting cases among at-risk populations often missed by traditional models. The results show that machine learning can fix class imbalance and make predictions more accurate, which are two big problems when looking at epidemiological data. This research helps the world fight disease by giving us a solution that can be used in places with few resources and is easy to scale up. It also stresses how important it is to have cheap, flexible technology for disease surveillance. The system made with open-source tools can be used by a lot of people without having to pay for licenses, which makes it more accessible. Even though the study has some good points, it also has some problems. The dataset is a good example, but it might not fully show how patterns of disease spread change over time and in different areas. We haven't tested how well the model works in real-time or changing conditions yet. To make research more useful and trustworthy, it should include data from the present, cover more areas, and take into account more social and environmental factors. These steps can help make predictions more accurate and interventions more effective in a given situation.

REFERENCE

Babawarun, Oloruntoba, Chioma Anthonia Okolo, Jeremiah Olawumi Arowoogun, Adekunle Oyeyemi Adeniyi, and Rawlings Chidi. 2024. "Healthcare Managerial Challenges in Rural and Underserved Areas: A Review." *World Journal of Biology Pharmacy and Health Sciences* 17 (2): 323–30. https://doi.org/10.30574/wjbphs.2024.17.2.0087.

Dirantara, Reza, and Febri Sugandi. 2025. "Prediksi Calon Kelulusan Mahasiswa Menggunakan Algoritma K-Nearest Neighbor (K-NN)." *Journal of Science and Social Research* 8 (1): 552–56. https://doi.org/10.54314/jssr.v8i1.2748.

Endo, Andrew, and Priyanga Amarasekare. 2022. "Predicting the Spread of Vector-Borne Diseases in a Warming World." *Frontiers in Ecology and Evolution* 10 (April). https://doi.org/10.3389/fevo.2022.758277.

Febiriana, Iffatricia Haura, Abdullah Hasan Hassan, and Dipo Aldila. 2024. "Enhancing Malaria Control Strategy: Optimal Control and Cost-Effectiveness Analysis on the Impact of Vector Bias on the Efficacy of Mosquito Repellent and Hospitalization." *Journal of Applied Mathematics* 2024 (March).

Kamguem, Inès Sopbué, Nathalie Kirschvink, Abel Wade, and Catherine Linard. 2025. "Determinants of Viral Haemorrhagic Fever Risk in Africa's Tropical Moist Forests: A Scoping Review of Spatial, Socio-Economic, and Environmental Factors." *Plos Neglected Tropical Diseases* 19 (1). https://doi.org/10.1371/journal.pntd.0012817.

Kumar, Athira Satheesh, Chris T. Bauch, and Madhur Anand. 2025. "Climate-Denying Rumor Propagation in a Coupled Socio-Climate Model: Impact on Average Global Temperature." *Plos One* 20 (1). https://doi.org/10.1371/journal.pone.0317338.

Lin, Tai Han, Hsing Yi Chung, Ming Jr Jian, Chih Kai Chang, Hung Hsin Lin, Chiung Tzu Yen, Sheng Hui Tang, et al. 2025. "AI-Driven Innovations for Early Sepsis Detection by Combining Predictive Accuracy with Blood Count Analysis

in an Emergency Setting: Retrospective Study." *Journal of Medical Internet Research* 27 (1). https://doi.org/10.2196/56155.

Manikandan, S., A. Mathivanan, Bhagyashree Bora, P. Hemaladkshmi, V. Abhisubesh, and S. Poopathi. 2022. "A Review on Vector Borne Diseases and Various Strategies to Control Mosquito Vectors." *Indian Journal of Entomology* 86 (1): 329–38. https://doi.org/10.55446/IJE.2023.410.

Piscitelli, Prisco, and Alessandro Miani. 2024. "Climate Change and Infectious Diseases: Navigating the Intersection through Innovation and Interdisciplinary Approaches." *International Journal of Environmental Research and Public Health*. https://doi.org/10.3390/ijerph21030314.

Pise, Reshma, and Kailas Patil. 2023. "A Deep Transfer Learning Framework for the Multi-Class Classification of Vector Mosquito Species." *Journal of Ecological Engineering* 24 (9): 183–91. https://doi.org/10.12911/22998993/168501.

Priyono, Eko, Teddy Al Fatah, Sukrul Ma'mun, and Faruq Aziz. 2023. "Tubercolusis Segmentation Based on X-Ray Images." *Journal Medical Informatics Technology* 1 (4): 101–4. https://doi.org/10.37034/medinftech.v1i4.22.

Priyono, Eko, Ispandi, and Rusdi. 2024. "Evaluating the Impact of Agricultural Technology on Greenhouse Gas Emissions Using Machine Learning." *Journal of Information Systems and Informatics* 6 (4). https://doi.org/https://doi.org/10.51519/journalisi.v6i4.870.

Ramadhan, Nur Ghaniaviyanto, and Azka Khoirunnisa. 2021. "Klasifikasi Data Malaria Menggunakan Metode Support Vector Machine." *Jurnal Media Informatika Budidarma* 5 (4): 1580–84. https://doi.org/10.30865/mib.v5i4.3347.

Saputra, Rendy Amy, and Aditya Pratama. 2025. "Implementasi Decision Tree Untuk Prediksi Harga Rumah Di Daerah Tebet." *Journal of Information System Management (JOISM) e-ISSN* 6 (2): 2715–3088. https://doi.org/10.24076/joism.2025v6i2.1928.

Shaikh, Salim Gulab, Billakurthi Suresh Kumar, Geetika Narang, and Nishant Nilkanth Pachpor. 2024. "Hybrid Machine Learning Method for Classification and Recommendation of Vector-Borne Disease." *Journal of Autonomous Intelligence* 7 (2). https://doi.org/10.32629/jai.v7i2.797.

Sobari, Syahrul, Ade Irma Purnamasari, Agus Bahtiar, and Kaslani. 2025. "Meningkatkan Model Prediksi Kelulusan Santri Tahfidz Di Pondok Pesantren Al-Kautsar Menggunakan Algoritma Random Forest." *Jurnal Informatika Dan Teknik Elektro Terapan* 13 (1). https://doi.org/10.23960/jitet.v13i1.5704.

Zhang, Dongsong, and Tianhua Chen. 2024. "Scikit-ANFIS: A Scikit-Learn Compatible Python Implementation for Adaptive Neuro-Fuzzy Inference System." *International Journal of Fuzzy Systems* 26 (6): 2039–57. https://doi.org/10.1007/s40815-024-01697-0.

Zhen, Jianing, Dehua Mao, Zhen Shen, Demei Zhao, Yi Xu, Junjie Wang, Mingming Jia, Zongming Wang, and Chunying Ren. 2024. "Performance of XGBoost Ensemble Learning Algorithm for Mangrove Species Classification with Multisource Spaceborne Remote Sensing Data." *Journal of Remote Sensing* 4 (June). https://doi.org/10.34133/remotesensing.0146.