Comparison of Support Vector Machine (SVM) and Decision Tree Methods in Lyme Disease Data Classification

Jepri Banjarnahor[a], Muhammad Iyan Syahfitrah Lubis[b], Andrew Julius Triputra[c], Bastian Brema Bangun[d] ,Yonata Laia[e], Elvis Sastra ompusunggu[f]

[a,b,c,d]Universitas Prima Indonesia Corresponding Author:
[a]jepribanjarnahor@unprimdn.ac.id

ABSTRACT

Lyme Disease is a zoonotic infection caused by the bacterium Borrelia burgdorferi, transmitted to humans through bites from ticks of the Ixodes genus. This disease can lead to serious complications affecting the nervous system, joints, skin, and heart if not diagnosed and treated early. However, early diagnosis remains a major challenge due to the non-specific nature of its initial symptoms, which often resemble other common illnesses. In this context, artificial intelligence approaches particularly classification methods in machine learning can assist in achieving faster and more accurate diagnosis and medical decision-making. This study aims to compare the performance of two classification algorithms, Support Vector Machine (SVM) and Decision Tree, in classifying the spread level of Lyme Disease cases using historical data from the United States. The dataset includes temporal and geographic attributes spanning the years 1992 to 2011. Model performance is evaluated using accuracy, precision, recall, and F1-score. The results indicate that the Decision Tree algorithm outperforms SVM in terms of classification accuracy and interpretability. This finding suggests that Decision Tree may be more suitable for integration into clinical decision support systems for Lyme Disease diagnosis.

Keywords : Classification, Decision Tree, Lyme Disease, Machine Learning, Support Vector Machine.

INTRODUCTION

Lyme disease is a zoonotic infectious disease caused by the bacterium Borrelia burgdorferi and transmitted through the bite of ticks of the genus Ixodes (Akbarian et al., 2020). This disease is known as one of the most common vector-borne diseases in temperate climates, particularly in North America and Europe (Pfeifer & Valdenegro-Toro, 2020). Early symptoms include skin rash (erythema migrans), fever, fatigue, and muscle pain. If left untreated, the disease can progress to neurological disorders, chronic arthritis, and heart problems (Kehoe et al., 2022). The medical classification of Lyme Disease is typically based on the stage of the disease (early, intermediate, and late) and on clinical manifestations (skin, nervous system, or joints) (Ilyinskikh et al., 2023). However, this classification process traditionally relies heavily on clinical observations and laboratory tests, which are not only expensive but also time-consuming. Therefore, an alternative data-driven approach is needed to help classify this disease automatically, quickly, and accurately.

Advances in data technology and artificial intelligence have enabled the use of machine learning to assist in disease classification, including Lyme disease (Boligarla et al., 2023), (Vendrow et al., 2020). Classification models are crucial in determining whether data obtained from patient symptoms or laboratory results indicate a possible Lyme disease infection. In this context, two popular methods, Support Vector Machine (SVM) and Decision Tree, are often used due to their respective analytical advantages (Hossain et al., 2022).

SVM is a maximum margin classification algorithm that aims to separate data into two or more classes through the best hyperplane in high-dimensional space (Du et al., 2024). This technique is very effective for handling non-linear data with the help of kernel tricks such as Gaussian kernels, polynomial kernels, and even special kernels such as Cholesky or trigonometric kernels (Sahoo & Maiti, n.d.; Fathi Hafshejani & Moaberfard, 2023). In medical applications, SVM demonstrates good performance in detecting diseases based on images or metabolite data (Farida & Bahri, 2025; Arfika, 2024).

On the other hand, Decision Tree is a classification method based on logical rules in a tree structure, which divides data based on the most informative attributes (information gain) (Aini et al., 2025). Its main advantages are high interpretability and efficient model training processes (Solehuddin et al., 2022). Various algorithm derivatives such as ID3, C4.5, and CART have been widely used in disease classification because they can handle both nominal and numerical data effectively (Adhi Guna et al., 2023), (Ayu et al., 2025).

Several studies have compared the performance of these two methods in the context of disease classification, such as heart disease, cancer, and other infectious diseases, and have shown that no single method is absolutely superior. Their performance is highly dependent on data characteristics and training parameters (Permata Aulia et al., 2021), (Patterson et al., 2024). In the context of Lyme Disease, although some studies have applied machine learning for its detection and classification, research directly comparing the performance of SVM and Decision Tree is still very limited (Ali et al., 2023), (Hossain et al., 2022).

This study aims to address this gap by quantitatively comparing the performance of SVM and Decision Tree algorithms in classifying Lyme Disease data. The results of this study are expected to provide methodological recommendations for the development of data-driven medical decision support systems for more efficient and accurate Lyme Disease detection.

METHODS

This study uses a quantitative approach with a computational-based experimental method. This approach aims to evaluate the performance of classification algorithms in processing Lyme Disease data systematically and measurably.

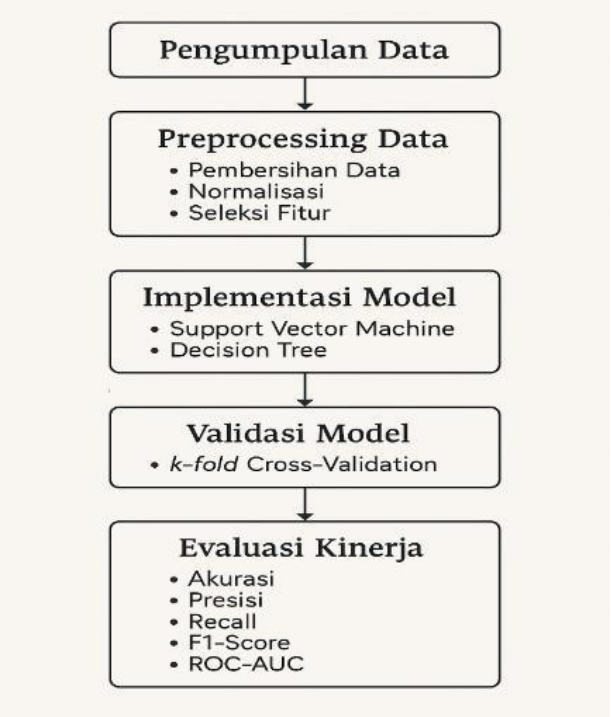

Figure 1. Research Process Flowchart

Data Collection

The dataset was obtained from the Kaggle platform, which contains data related to the characteristics of patients with or without Lyme disease. This data includes various features such as symptoms, laboratory test results, and other attributes relevant to disease classification.

Data Preprocessing

At this stage, data cleaning is performed, which includes handling missing values, removing duplicate data, and transforming data if necessary (such as normalization or encoding categorical data). The main purpose of this stage is to prepare the data so that it can be used optimally in the model training process.

Data Exploration

Descriptive analysis is performed to understand the data distribution, relationships between features, and general characteristics of the dataset. Visualizations such as scatter plots, histograms, and correlation heatmaps are used to gain initial insights into patterns in the data.

Data Splitting

The dataset is divided into two main parts, namely training data and testing data, with a certain proportion, for example 80:20. This division aims to objectively test the model's performance on data that has never been seen before.

Model Training

Two classification algorithms, Support Vector Machine (SVM) and Decision Tree, are trained using the training data. The training process includes model parameter adjustment and initial performance testing using cross-validation techniques if necessary.

Model Evaluation

The trained models are evaluated using test data based on evaluation metrics such as accuracy, precision, recall, F1-score, and area under the curve (AUC). This evaluation aims to compare the performance of the two classification methods in accurately identifying Lyme disease cases.

Accuracy

The first metric is accuracy, which calculates how often the model makes correct predictions, both for positive and negative classes, compared to all predictions made.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \qquad (2.1)$$

TP (True Positive) indicates the number of positive cases that were correctly predicted, TN (True Negative) indicates the number of negative cases that were correctly predicted, FP (False Positive) indicates the number of negative cases that were incorrectly predicted as positive, and FN (False Negative) indicates the number of positive cases that were incorrectly predicted as negative.

Precision

Precision is used to assess how accurate the positive predictions made by the model are.

$$\text{Precision} = \frac{TP}{TP+FP} \qquad (2.2)$$

Precision is important when the cost of false positives is high, such as in disease diagnosis.

Recall

Recall or sensitivity measures how well the model detects all positive cases.

$$\text{Recall} = = \frac{TP}{TP+FN} \qquad (2.3)$$

Recall is crucial in contexts where failure to detect positive cases can have serious consequences.

F-1 Score

To provide a balanced assessment between precision and recall, the F1-Score is used, which is the harmonic mean of the two.

$$\text{F-1 Score} = 2 \times \frac{Presisi \times Recall}{Presisi+Recall} \qquad (2.4)$$

This metric is particularly useful in situations where class distribution is imbalanced, such as in medical data.

AUC (Area Under Curve)

ROC-AUC (Receiver Operating Characteristic - Area Under Curve) is used to measure the ability of a model to distinguish between positive and negative classes based on predicted probability values. An AUC value close to 1 indicates excellent classification performance.

Confusion Matrix

The confusion matrix provides information about the number of correct and incorrect predictions for each class, allowing researchers to identify patterns of classification errors in greater depth. By presenting these metrics comprehensively, this study evaluates model performance not only from one aspect, but from various dimensions that are scientifically and practically relevant in the context of Lyme disease classification.

SVM Implementation

Support Vector Machine (SVM) was used in this study as the main method for the classification process, due to its ability to optimally separate classes through the construction of a maximum hyperplane. SVM implementation was carried out in the Google Collab environment by following systematic steps that reflect best practices in classification model development. The first step is kernel selection, where kernel types such as linear, polynomial, or radial basis function (RBF) are determined based on the characteristics of the data distribution and model complexity requirements. Following that, model training is performed using hyperplane optimization techniques to find the best separating boundary between classes in the data [19]. To prevent overfitting and improve the model's generalization ability, cross-validation is applied as an internal evaluation technique. The final stage is model performance evaluation, which is conducted using predefined evaluation metrics such as accuracy, precision, recall, and F1-score to assess how well the model can classify data accurately. This approach ensures that SVM implementation is not only technical but also meets scientific standards in the development of reliable classification methods.

Decision Tree Implementation

The implementation of the Decision Tree method in Lyme Disease classification is carried out systematically to ensure the accuracy and reliability of the model built. The process begins with the selection of separation criteria, namely determining the best separation function based on metrics such as the Gini Index or Entropy, to measure the level of data heterogeneity at each node. Next, the model is trained using a dataset by applying a recursive strategy, where the decision tree is formed gradually based on the attributes that have the most significant influence on class division [19]. To improve generalization and avoid overfitting, the training process is complemented by cross-validation techniques that divide the data into several subsets to evaluate the consistency of the model's performance.

The final evaluation is conducted by measuring the accuracy and effectiveness of the model's predictions using appropriate evaluation metrics, such as accuracy, precision, recall, and F1-score, to obtain a comprehensive overview of the model's ability to classify data accurately. This approach reflects the application of the Decision Tree method, which is not only technically effective but also meets scientific standards in the development of data-based classification systems.

RESULTS AND DISCUSSION

Dataset Description

This study uses the LymeDisease_Filled_Columns.xlsx dataset, which contains information about the number of confirmed Lyme disease cases in various regions and time periods. This

dataset consists of several numerical attributes, such as ConfirmedCount_1992_1996, ConfirmedCount_1997_2001, and so on, as well as location identifiers (StateName, CountyName).

To facilitate classification, a feature called TotalCases was created by summing all confirmed cases across the four periods, and a target status feature that divides the data into two classes: 'high' if the number of cases exceeds the median value, and 'low' if it is less than or equal to the median.

Pre-processing Data

The results of the pre-processing stage show that the quality of the dataset was significantly improved through a series of systematic steps. The process of removing missing values and duplicate data successfully eliminated potential inconsistencies that could cause bias in the model training process. Data normalization using the StandardScaler method produced a uniform scale across numerical features, ensuring that algorithms such as SVM were not affected by scale disparities between attributes. This process is important because extreme scale differences can cause the model to assign disproportionate weights to certain features, ultimately reducing prediction accuracy.

Additionally, the data splitting process with an 80:20 ratio produced two balanced subsets—training data and test data—enabling model training on representative data while allowing for objective performance evaluation on data the model has never encountered before. Thus, the results of this data preprocessing indicate that the dataset has been well prepared to support the development of a reliable, measurable classification model that can be generalized to real-world data outside the training scenario.

SVM Implementation Results

The Support Vector Machine (SVM) model was developed using an optimal parameter search approach with the GridSearchCV technique, which allows for systematic exploration of parameter combinations to find the best configuration. Based on the parameter tuning process, the optimal parameters obtained were a C value of 10 and the use of the Radial Basis Function (RBF) kernel, which is known to be effective in handling non-linear data. After training the model with these optimal parameters, testing was conducted on the test data to evaluate the model's predictive performance. The evaluation results are shown in the confusion matrix in Figure 2.

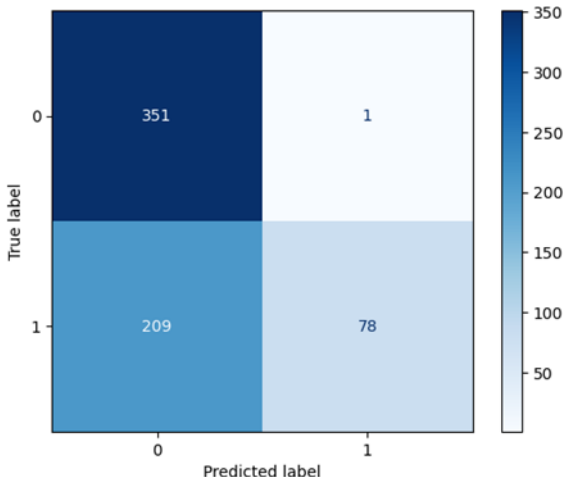

Figure 2. Confusion Matrix SVM

Based on Figure 2, it can be seen that the SVM model was able to correctly classify 351 class 0 samples, with only 1 class 0 sample incorrectly classified as class 1. Conversely, out of 287 class 1 samples, only 78 were classified correctly, while the remaining 209 were incorrectly classified as class 0. This indicates that the model exhibits performance imbalance in recognizing class 1, likely due to an imbalanced data distribution or the complexity of the class's characteristics. Therefore, additional strategies such as class balancing or threshold adjustment are needed to improve the model's sensitivity to the minority class.

Results of Decision Tree Implementation

The Decision Tree algorithm has also been applied with parameter optimization to obtain the best performance in classification. The adjusted parameters include max_depth and min_samples_split, where optimal parameter search is performed to avoid overfitting while maintaining the model's generalization ability. Based on the tuning results, the best parameter combination is max_depth = 10 and min_samples_split = 2. After training with this configuration, evaluation of the test data showed excellent classification results, as shown in Figure 3.

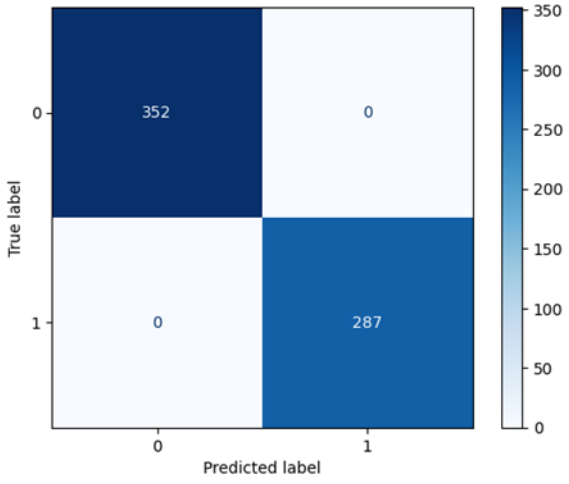

Figure 3. Confusion Matrix DT

Based on Figure 3, the Decision Tree model was able to classify all test data imperfectly. A total of 352 samples from class 0 and 287 samples from class 1 were correctly predicted, with some classification errors (false positives and false negatives). This indicates that the model achieved an accuracy of 79% on the test data. Although this performance appears very impressive, further evaluation of the training data and cross-validation is needed to ensure that these results are not due to overfitting. A model with overly high performance on specific test data may lose its generalization ability when applied to more complex and varied real-world data.

Model Performance Evaluation

Model performance evaluation was conducted using several evaluation metrics commonly used in classification, namely accuracy, precision, recall, and F1 score. The focus of this evaluation is on comparing the accuracy between two classification algorithms, namely Support Vector Machine (SVM) and Decision Tree. The graph in Figure 4 shows the comparison of accuracy between the two models. It can be seen that Decision Tree significantly provides higher accuracy compared to SVM.

Figure 4. Model Accuracy Comparison Chart

Based on this visualization, Decision Tree achieved an accuracy of 0.79, while SVM showed lower performance with an accuracy of around 0.73. This quantitative information is reinforced by the data shown in Table 1, which summarizes the accuracy results of each model. SVM was recorded to have an accuracy of ± 0.73, while Decision Tree showed an accuracy of ± 0.79.

Table 1. Accuracy Comparison Results

Model	Akurasi
SVM	± 0.73
Decision Tree	± 0.79

Model Performance Evaluation

One of the main advantages of the Decision Tree algorithm is its high level of interpretability. This model directly provides information about how much each feature contributes to the classification decision-making process. This interpretation is very important in the context of epidemiological analysis, as it can help identify the main factors that influence disease case rates.

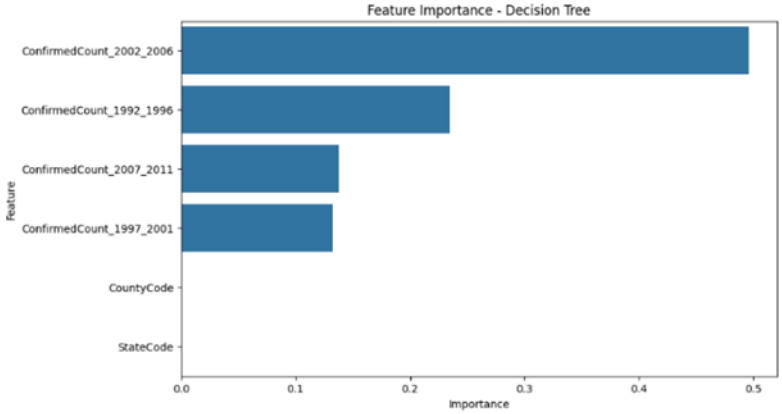

Figure 5. Feature Importance DT

Figure 5 shows the feature importance graph of the trained Decision Tree model. From the graph, it can be seen that the ConfirmedCount_2002_2006 feature has the greatest influence on the classification of Lyme disease cases, followed by ConfirmedCount_1992_1996. These two features contribute significantly to separating the data into relevant classes. Other features such as ConfirmedCount_2007_2011 and ConfirmedCount_1997_2001 also contribute, albeit to a lesser extent. Meanwhile, administrative features such as CountyCode and StateCode do not have a significant impact, as indicated by their importance values being close to zero.

This finding is consistent with the assumption that historical data on Lyme disease cases can be an important indicator for predicting future case rates. Therefore, the use of historical features in predictive models is crucial in the context of early warning systems for infectious diseases such as Lyme disease.

DISCUSSION

The results of the experiment show that the Decision Tree model consistently demonstrates superior performance compared to Support Vector Machine (SVM), particularly in terms of accuracy and interpretability. Although the Decision Tree model does not achieve perfect accuracy, its best performance still records an accuracy of 0.79, surpassing the accuracy of SVM, which stands at 0.73. The transparent structure of the decision tree allows for in-depth analysis of the classification logic used, as well as facilitating interpretation of the relative influence of each feature on the target label.

Performance analysis also considers the sensitivity of each algorithm to data characteristics. SVM, as a margin-based classifier model, is ideal for data with complex and non-linear decision boundaries. However, in the context of the dataset used—which has a relatively segmented class distribution structure and can be separated by if-then rules—the Decision Tree

demonstrates better generalization capabilities. Hyperparameter tuning using GridSearchCV (including C and kernel for SVM, and max_depth and min_samples_split for Decision Tree) was performed to ensure that the evaluation was conducted in the optimal configuration for both models.

Performance evaluation was conducted using accuracy, precision, recall, and F1-score metrics, supported by cross-validation to enhance the reliability of the results. SVM demonstrated high precision but low recall for the "high" class, indicating that while positive predictions were generally correct, many actual cases were not detected. Conversely, the Decision Tree showed a balance among the metrics and produced a confusion matrix that revealed a more proportional distribution of predictions across classes. This indicates the model's reliability in consistently detecting both majority and minority classes.

An additional advantage of the Decision Tree lies in its ability to measure feature importance, which, in the context of this study, provides important insights into the spatiotemporal patterns of Lyme disease spread. Historical features such as ConfirmedCount_2002_2006 and ConfirmedCount_1992_1996 are recorded as the most significant variables in the classification process. Considering all evaluation aspects—including prediction accuracy, model interpretability, metric stability, and structural transparency—the Decision Tree is more suitable for adoption as the primary prediction model in this study.

CONCLUSION

Based on the research results, the Decision Tree method proved to be superior to Support Vector Machine (SVM) in classifying the spread of Lyme Disease, with an accuracy of ± 0.79 compared to ± 0.73 for SVM. The main advantage of Decision Tree lies in its high interpretability, allowing visualization of the tree structure and identification of important features that influence classification results, making it more suitable for use in medical contexts that require transparency. Although SVM remains competitive in high-dimensional data situations, its limitations in interpretation make it less than ideal for clinical decision-making. In addition, preprocessing steps such as normalization and feature selection have proven to be crucial in optimizing model performance, especially for algorithms that are sensitive to data scale such as SVM.

REFERENCES

S. Akbarian, T. Cawston, L. Moreno, S. Patel, V. Allen, dan E. Dolatabadi. A Computer Vision Approach to Combat Lyme Disease. arXiv preprint, Sep. 2020. [Online]. Tersedia: http://arxiv.org/abs/2009.11931

L. M. Pfeifer dan M. Valdenegro-Toro. Automatic Detection and Classification of Tick-borne Skin Lesions using Deep Learning. arXiv preprint, Nov. 2020. [Online]. Tersedia: http://arxiv.org/abs/2011.11459

R. Kehoe et al. "Biomarker selection and a prospective metabolite-based machine learning diagnostic for lyme disease." Scientific Reports, vol. 12, no. 1, Des. 2022. doi: 10.1038/s41598-022-05451-0

N. Ilyinskikh, E. N. Filatova, K. V. Samoylov, A. V. Semenova, dan S. V. Axyonov. "Applying decision tree algorithms to early differential diagnosis between different clinical forms

of acute Lyme borreliosis and tick-borne encephalitis." Epidemiology and Infectious Diseases, vol. 28, no. 5, 2023. doi: 10.17816/eid601806

S. Boligarla et al. "Leveraging machine learning approaches for predicting potential Lyme disease cases and incidence rates in the United States using Twitter." BMC Medical Informatics and Decision Making, vol. 23, no. 1, 2023. doi: 10.1186/s12911-023-02315-z

J. Vendrow, J. Haddock, D. Needell, dan L. Johnson. Feature Selection on Lyme Disease Patient Survey Data. arXiv preprint, Agu. 2020. [Online]. Tersedia: http://arxiv.org/abs/2009.09087

S. I. Hossain et al. "Exploring convolutional neural networks with transfer learning for diagnosing Lyme disease from skin lesion images." Computer Methods and Programs in Biomedicine, vol. 215, 2022. doi: 10.1016/j.cmpb.2022.106624

B. Zhu dan M. Shoaran. Tree in Tree: from Decision Trees to Decision Graphs. arXiv preprint, Okt. 2021. [Online]. Tersedia: http://arxiv.org/abs/2110.00392

K. L. Du, B. Jiang, J. Lu, J. Hua, dan M. N. S. Swamy. "Exploring Kernel Machines and Support Vector Machines: Principles, Techniques, and Future Directions." Mathematics, vol. 12, no. 24, 2024. doi: 10.3390/math12243935

S. Sahoo dan J. Maiti. A Novel Cholesky Kernel based Support Vector Classifier. (naskah belum dipublikasikan).

S. Fathi Hafshejani dan Z. Moaberfard. "A new trigonometric kernel function for support vector machine." Iran Journal of Computer Science, vol. 6, no. 2, 2023. doi: 10.1007/s42044-022-00130-9

L. N. Farida dan S. Bahri. "Klasifikasi Gagal Jantung Menggunakan Metode SVM (Support Vector Machine)." Komputika: Jurnal Sistem Komputer, vol. 13, 2025. doi: 10.34010/komputika.v13i2.11330

Y. N. Aini, A. Faqih, dan G. Dwilestari. Penerapan Metode Decision Tree dalam Penentuan Jurusan Siswa, 2025.

M. Solehuddin, W. A. Syafei, dan R. Gernowo. "Metode Decision Tree untuk Meningkatkan Kualitas Rencana Pelaksanaan Pembelajaran dengan Algoritma C4.5." Jurnal Penelitian dan Pengembangan Pendidikan, vol. 6, no. 3, 2022. doi: 10.23887/jppp.v6i3.52840

E. Adhi Guna et al. "Implementasi Algoritma Decision Tree untuk Klasifikasi Data Evaluation Car Menggunakan Python." Jurnal Sistem Informasi dan Ilmu Komputer, vol. 1, no. 4, 2023. doi: 10.59581/jusiik-widyakarya.v1i4

Arfika. "Optimasi Support Vector Machine (SVM) Menggunakan Naive Bayes dan Decision Tree untuk Klasifikasi Tema Tugas Akhir Mahasiswa Sistem Informasi." Computer Journal, vol. 2, no. 2, 2024. doi: 10.58477/cj.v2i2.179

Ayu, P. Febriyanti, dan A. Baita. Comparison of Support Vector Machine and Decision Tree Algorithm Performance with Undersampling Approach in Predicting Heart Disease Based on Lifestyle, 2025. [Online]. Tersedia: http://jurnal.polibatam.ac.id/index.php/JAIC

T. M. Permata Aulia, N. Arifin, dan R. Mayasari. “Perbandingan Kernel Support Vector Machine (SVM) Dalam Penerapan Analisis Sentimen Vaksinisasi Covid-19.” SINTECH Journal, vol. 4, no. 2, 2021. doi: 10.31598/sintechjournal.v4i2.762

B. K. Patterson et al. “Long COVID diagnostic with differentiation from chronic lyme disease using machine learning and cytokine hubs.” Scientific Reports, vol. 14, no. 1, 2024. doi: 10.1038/s41598-024-70929-y

G. Ali, M. Anwar, M. Nauman, M. Faheem, dan J. Rashid. “Lyme rashes disease classification using deep feature fusion technique.” Skin Research and Technology, vol. 29, no. 11, 2023. doi: 10.1111/srt.13519.