Journal of Artificial Intelligence and Engineering Applications

Website: https://ioinformatic.org/

15th February 2025. Vol. 4. No. 2; e-ISSN: 2808-4519

Optimizing Naïve Bayes Algorithm Through Principal Component Analysis To Improve Dengue Fever Patient Classification Model

Santi Nurjulaiha[1]*, Rudi Kurniawan[2], Arif Rinaldi Dikananda[3], Tati Suprapti[4]

[1,3]Informatics engineering, STMIK IKMI Cirebon
[2,4] Software engineering, STMIK IKMI Cirebon
santicrb1403@gmail.com [1]*

Abstract

Dengue fever is an infectious disease that has a significant impact on public health in tropical regions, including Indonesia. Early detection and proper classification of DHF patients is essential to reduce severity and mortality. For this reason, a method that can improve the accuracy in diagnosing this disease is needed. Principal Component Analysis (PCA) and Naïve Bayes (NB) are two commonly used techniques in medical data analysis. PCA is used to reduce the dimensionality of data to reduce complexity, while Naïve Bayes is used for classification of data based on probability. This study aims to optimize the use of PCA and Naïve Bayes in improving the accuracy of the dengue patient classification model. The method used in this study involves processing a medical dataset of dengue patients containing various clinically relevant attributes. The dataset was then processed using PCA to reduce dimensionality and identify key features that affect classification. Next, Naïve Bayes was applied to classify the data based on the selected features. This study compares the performance of classification models that use a combination of PCA and Naïve Bayes with models that only use Naïve Bayes without dimensionality reduction. The results show that the use of PCA in data processing significantly improves the accuracy of the classification model compared to the model that only uses Naïve Bayes. The combination of PCA and Naïve Bayes produces a more efficient model and has a higher accuracy rate in identifying patients with DHF risk. Thus, the application of PCA and Naïve Bayes in the classification of DHF patients can be an effective tool in assisting the medical diagnosis process, which in turn can reduce misdiagnosis and improve patient recovery rates. This research contributes to the development of artificial intelligence technology in the medical field, especially to improve the accuracy of dengue disease diagnosis, and serves as a basis for further research in the use of machine learning techniques in healthcare. This study analyzes the performance of the Naïve Bayes algorithm in classifying dengue fever patient data, by comparing models that use Principal Component Analysis (PCA) as a dimension reduction method and models that do not use it. The results show that the Naïve Bayes model without PCA has an accuracy of 49.96%, which is close to the random guess rate. This finding indicates that the model is less effective in recognizing patterns in the data. In contrast, the application of PCA successfully increased the model's accuracy to 50.03%.

Keywords: *Naïve Bayes, Principal Component Analysis, Classification, Dengue Fever*

1. Introduction

In the era of development of information technology and data science, classification model has become one of the fundamental techniques in data mining and machine learning to predict the class or category of a data sample based on a set of input features [1]-[2]-[3]-[4]-[5]. The use of machine learning in healthcare, especially in the classification of infectious diseases such as dengue fever, is highly relevant given the diagnostic challenges faced in detecting such diseases. Dengue fever, caused by the dengue virus and transmitted through the bite of the Aedes aegypti mosquito, is a growing global health problem. Although various preventive measures have been taken, the incidence of dengue remains high, especially in tropical and subtropical countries [6]. The use of machine learning techniques such as Principal Component Analysis (PCA) and Naïve Bayes shows great potential in improving accuracy and efficiency in patient classification based on clinical and laboratory data. PCA, which is useful for reducing the dimensionality of data without losing important information, has been used in various studies to improve the performance of prediction models [7]. Naïve Bayes, as a simple yet effective probabilistic model, has also been widely used in disease classification with limited data [8]. Therefore, this study combines the two methods to improve the classification accuracy of dengue fever patients, by utilizing dimension reduction techniques through PCA and classification using Naïve Bayes.

Although there are various approaches in dengue patient classification, both based on classical statistical methods and machine learning, there are still many challenges that cannot be overcome, especially related to the complexity of large and dynamic data. Traditional methods such as Poisson regression and ARIMA, while useful, cannot handle large data with many variables that affect the incidence of dengue epidemics, such as weather factors and disease vectors. This indicates that machine learning-based approaches, which can handle big data more effectively [9]. The use of PCA to reduce the dimensionality of data has shown that this technique can improve model accuracy by identifying key features that influence disease incidence, while reducing data complexity [7]. Naive Bayes is a probability classification technique and method. In many complex real-world situations, Naive Bayes also performs much better. Due to its simplicity, Naïve Bayes is a common machine learning model that allows all attributes to contribute equally to the final decision [10]. Naïve Bayes, although effective in handling categorical and numerical data, has limitations in terms of accuracy when faced with more complex data [8]. Therefore, combining PCA and Naïve Bayes is expected to overcome this problem and produce a more accurate and efficient classification model, although techniques such as backpropagation in artificial neural networks can be used for classification, there have been no studies combining such techniques with PCA in the context of classifying dengue patients [11]. This approach suggests that while each technique has its strengths, combining the two can bring greater benefits in improving the efficiency and accuracy of the model.

This study aims to improve the classification model using the naïve bayes algorithm based on principal component analysis on dengue fever patient data. Dengue as one of the infectious diseases that is highly endemic in tropical areas requires a classification model that is not only accurate but also efficient in processing large and diverse clinical and laboratory data. Therefore, this study was conducted to develop and improve PCA and Naïve Bayes-based classification models that are able to process clinical and epidemiological data more accurately and quickly. The importance of accurately classifying dengue patients lies in its potential impact on outbreak management and public health policy. One of the major challenges in the management of this disease is the wide variability of clinical symptoms, which often overlap with other diseases. Therefore, having a model that can classify patients appropriately is essential to ensure that patients who need immediate care get the right attention. PCA will be used to reduce the dimensionality of the often very large and varied data, while still retaining the relevant information required for the classification process [7]. In this study, Naïve Bayes will be applied to the data that has been reduced in dimension through PCA to ensure efficiency and speed in the classification process. The expected result is a model that is not only accurate but also fast in classifying patients, which is important in disease outbreak management. In addition to improving accuracy, this research also aims to develop a model that can be practically applied in clinical and epidemiological settings. In the medical world, speed in decision-making is crucial, especially in dealing with rapidly evolving diseases such as dengue fever. Therefore, a fast and efficient classification model can support medical personnel in making more informed decisions, by providing accurate predictions in a short time. Through combining PCA and Naïve Bayes, this research aims to overcome the challenges of processing large data that often requires long computational time, as well as improve classification accuracy to enable faster and more precise treatment [12]. This research also has the potential to have a major impact on public health policy and dengue outbreak control. By generating a more accurate classification model, this research can help in mapping the severity of the disease, predicting the likelihood of spread, as well as identifying patients most at risk. This will allow health authorities to better plan interventions, allocate resources more efficiently, and take more effective preventive measures. For example, by catching patients who may be suffering from severe dengue early, more intensive medical measures can be taken to prevent the disease from progressing further. In this context, the success of this research could contribute to improving the capacity of health systems to respond to outbreaks faster and more appropriately, thereby reducing the impact of future dengue epidemics. This research also has the potential to enrich the literature in the field of medical data processing and machine learning. By developing PCA and Naïve Bayes-based classification models, this research will add to the understanding of how dimension reduction techniques can be used in the context of infectious disease classification, as well as contribute to the development of models that can be adapted for other diseases with similar epidemic patterns. In addition, this research is expected to introduce methods that can be used to handle large and complex data which is often a challenge in the medical field.

This research will adopt an experimental approach to develop a classification model that combines Principal Component Analysis (PCA) and Naïve Bayes techniques. The research process begins with the collection of clinical data and laboratory test results from patients diagnosed with dengue fever. This data will include various variables, such as patient demographics, clinical symptoms, physical examination results, as well as relevant laboratory data, including platelet count, hematocrit, and serology test results. The collected data will be analyzed and processed through several stages, starting with data pre-processing to ensure data quality and consistency [11]. Data pre-processing is an important first step in ensuring that the data used in the classification model does not contain errors or gaps that could affect the results of the analysis. This stage includes data cleaning, filling in missing values, and normalizing or standardizing numerical data [13]. In addition, categorical data coding is also performed, if needed, to prepare the data for use by machine learning algorithms. Once the data is processed, the next step is to apply the PCA technique to reduce the dimensionality of the data. PCA will be used to identify the principal components that explain the greatest variation in the data and reduce data complexity by eliminating less relevant or redundant features. The goal of dimensionality reduction is to improve the efficiency of model processing without compromising classification accuracy. PCA also helps overcome multicollinearity issues that can arise in data that has many related features [7]. After the data has been reduced in dimension using PCA, the next stage is the application of the Naïve Bayes algorithm to classify dengue fever patients. Naïve Bayes is well known for its simplicity and ability to handle data with different categories, be it numerical or categorical. Although Naïve Bayes is often considered a less complex model compared to other methods such as SVM or neural networks, it can provide adequate results in disease classification with limited data. Once the model is built, an evaluation phase will be conducted to measure the performance of the model in classifying dengue patients [8]. The evaluation method used will include cross-validation, where the dataset is divided into subsets to ensure that the model is thoroughly tested. The measurements used to evaluate the performance of the model will include accuracy, precision, recall, and F1-score. Accuracy measures how many correct predictions are made by the model, while precision and recall provide deeper insight into how the model handles imbalanced data, where one class (e.g., severe dengue) may be rarer than another [6].

This research has some significant implications in dengue disease management by producing a more efficient and accurate classification model. If the objectives of this research are achieved, the resulting model will have several important implications, both in the clinical and public health policy contexts. By improving the accuracy in patient classification based on clinical and laboratory data, the model can assist doctors and medical personnel in diagnosing dengue more quickly. In emergency situations or outbreaks, speed in determining whether a patient has dengue or another similar illness is critical to providing timely treatment. With a faster classification model, it is expected to reduce delays in diagnosis, which in turn can reduce mortality from dengue [11]. From a health policy perspective, a more accurate model could help health authorities to plan more appropriate interventions, such as allocating medical resources to areas of greatest need or prioritizing patients most at risk. This research could also provide a basis for the development of a more responsive and data-driven dengue epidemic monitoring system, allowing decision-makers to track the spread of the disease more efficiently and plan more appropriate control strategies [12]. In terms of scientific contribution, this research can enrich the literature in the field of epidemiology and medical data processing. By combining PCA and Naïve Bayes, this study proposes a new approach in the classification of infectious diseases based on clinical and laboratory data. This research can also serve as a basis for the development of similar classification models for other infectious diseases that have similar epidemic patterns to dengue fever, such as malaria or chikungunya [7]. The model generated from this research also has the potential to be implemented in developing countries, where medical resources are often limited, yet the incidence of infectious diseases such as dengue is still high. By using efficient machine learning techniques, such as PCA and Naïve Bayes, the model can be operated in environments with limited computing capacity, while providing a solution to improve the health system's ability to deal with disease outbreaks [9]. In addition to its practical applications, this research also has the potential to improve our understanding of the factors that influence the incidence of dengue fever. By using extensive data and analyzing it with sophisticated techniques, this model can help researchers and policy makers to better understand the epidemic patterns of dengue fever, which in turn can lead to more effective disease control strategies in the future [8].

2. Related Work

Research conducted by Altayeb and Arabiat (2024) entitled “Enhancing Stroke Prediction Using the Waikato Environment for Knowledge Analysis” focuses on using machine learning and PCA to improve stroke prediction in medical data. In this study, Naïve Bayes algorithm was used to classify the data, and PCA was applied to reduce the dimension of the huge data, making it easier to manage and analyze. While the main focus of this study was to predict stroke, the application of PCA to reduce data complexity is highly relevant to your research which aims to improve the classification of dengue patient data using Naïve Bayes. The techniques applied in this study provide valuable insights for processing more complex medical data [14]. Shenify (2024) in his research entitled “Sentiment Analysis of Saudi E-commerce Using Naïve Bayes and SVM” used Naïve Bayes and Support Vector Machine (SVM) to perform sentiment analysis on e-commerce data in Saudi Arabia. This research shows how classification algorithms such as Naïve Bayes and SVM can be used to classify customer opinions in e-commerce platforms. Although the context focuses on sentiment analysis in the commercial sector, the techniques used in this study can be adapted for medical applications, such as your study that used Naïve Bayes to classify dengue fever patient data [15]. Research conducted by Alghushairy et al. (2024) entitled “An Efficient Support Vector Machine Algorithm-Based Network Outlier Detection System” uses Support Vector Machine (SVM) to detect anomalies or outliers in network systems. This research applies classification techniques to detect unusual patterns in network data. Although the main focus is on network data, this research provides insight into how classification techniques such as SVM can be used to detect patterns in large and complex datasets, which has similarities with your research which also aims to classify medical data of dengue fever patients [16]. Research conducted by Hassan et al. (2024) entitled “Predicting Student Dropout Rates Using Supervised Machine Learning: Insights from the 2022 National Education Survey in Somaliland” used machine learning to predict student dropout rates in Somaliland. It uses various supervised learning algorithms, including Naïve Bayes, to predict factors that influence dropout rates and attempts to classify student data based on these predictions. While the main focus is on educational data, this study provides useful insights into the use of classification algorithms to predict outcomes in large datasets, similar to the classification of medical patient data in your study [17].

Monteverde-Suárez et al. (2024) in a study entitled “Predicting Students' Academic Progress and Related Attributes in First-Year Medical Students: An Analysis with Artificial Neural Networks and Naïve Bayes” used a combination of artificial neural networks (ANN) and Naïve Bayes to predict the academic progress of first-year medical students. This research uses a combination of ANN and Naïve Bayes models to analyze academic data and other related attributes. Although the focus was on medical students, the use of Naïve Bayes to classify complex and diverse data is highly relevant to your research which also used the Naïve Bayes algorithm on medical data, although the type of data and the purpose of classification were different [18]. Stonier et al. (2024) in a study entitled “Cardiac Disease Risk Prediction Using Machine Learning Algorithms” examined the application of machine learning algorithms, including Naïve Bayes, to predict the risk of heart disease. This study used patient medical datasets to predict the potential risk of heart disease using classification techniques. While this study focused on heart disease, the techniques used in medical classification are highly relevant to your research, which focuses on classifying dengue patient data. These two studies used similar techniques, although the diseases predicted were different [19]. Mahendra et al. (2024) in a study entitled “Performance Enhancement of Naïve Bayes Method Using AdaBoost for Classification of Diabetes Mellitus Dataset Type II” combined Naïve Bayes with the AdaBoost method to improve classification accuracy on diabetes mellitus datasets. This research shows that by combining ensemble methods, classification performance can be improved. A similar approach can be applied to your research to improve disease classification by combining Naïve Bayes and other techniques, such as PCA, to maximize accuracy. The difference lies in the type of disease analyzed, namely diabetes mellitus in this study and dengue fever in your study [20]. Research conducted by Saheed et al. (2024) entitled “Feature Selection in Intrusion Detection Systems: A New Hybrid Fusion of Bat Algorithm and Residue Number System” focuses on feature selection for intrusion detection systems using the Bat algorithm and Residue Number System. The feature selection technique used in this research focuses on selecting the best attributes from large and complex datasets, which is highly relevant to the PCA approach used in your research to select important features in medical data. Although this research focuses on intrusion detection systems, the feature selection concepts used have similarities in the context of medical data processing [21]. Singh et al. (2024) in a study entitled “Spectral-Spatial Classification with Naïve Bayes and Adaptive FFT for Improved Classification Accuracy of Hyperspectral Images” applied Naïve Bayes along with adaptive FFT techniques to improve classification accuracy on hyperspectral image data. This research shows how a combination of classification techniques can improve accuracy on image data, which can also be adapted for medical data in your research. Although focused on hyperspectral images, the approach used to improve classification accuracy is relevant to disease classification in your study using dengue patient data [22].

Morakis and Adamopoulos (2024) in a study titled “Hybrid Machine Learning Algorithms to Evaluate Prostate Cancer” used hybrid machine learning algorithms to evaluate prostate cancer. This research shows that by combining several algorithms, such as Naïve Bayes, the accuracy in prostate cancer classification can be improved. Although this study focused on prostate cancer, the use of hybrid machine learning techniques to improve medical classification can provide useful insights for your research that focuses on classifying dengue fever patient data [23]. Research conducted by Sulak and Koklu (2024) entitled “Analysis of Depression, Anxiety, Stress Scale (DASS-42) with Methods of Data Mining” focuses on the application of data mining methods, including Naïve Bayes, to analyze data related to the depression, anxiety, and stress scale (DASS-42). This research shows how Naïve Bayes can be used to classify mental health data effectively. Although the focus of this research is on mental health, the use of Naïve Bayes for medical data classification is highly relevant to your research which also focuses on medical disease classification [24]. Hameed et al. (2024) in a study entitled “Leukemia Diagnosis

Using Machine Learning Classifiers Based on MRMR Feature Selection" used machine learning algorithms and MRMR feature selection to diagnose leukemia. This research shows how classification algorithms such as Naïve Bayes, combined with feature selection techniques, can improve diagnosis accuracy. Although this study focused on leukemia, the feature selection approach used has similarities to the PCA technique in your study, which aimed to improve the classification of dengue medical data [25].

Shams et al. (2024) in a study entitled "Enhancing Crop Recommendation Systems with Explainable Artificial Intelligence: A Study on Agricultural Decision-Making" focuses on improving agricultural recommendation systems using explainable artificial intelligence (XAI). Although the context of this study is agriculture, the use of XAI to explain decisions generated by AI models provides insight into how to explain classification decisions in medical systems, which is highly relevant to your research on improving dengue disease prediction with Naïve Bayes [26]. Khasim et al. (2024) in their research entitled "Using Deep Learning and Machine Learning: Real-Time Discernment and Diagnostics of Rice-Leaf Diseases in Bangladesh" used deep learning and machine learning to diagnose rice leaf diseases in Bangladesh. Although focused on plant diseases, the classification techniques used in this study provide valuable guidance that can be applied to human disease classification, as done in your study of dengue fever patient data [27]. Albataineh et al. (2024) in a study titled "COVID-19 CT-Images Diagnosis and Severity Assessment Using Machine Learning Algorithm" applied machine learning algorithms to diagnose COVID-19 based on CT images and assess disease severity. The use of Naïve Bayes in this study shows how the algorithm can be used in medical image processing for disease diagnosis, although it is different from the focus of your study which classifies dengue fever patient data [28]. Yaman et al. (2024) in a study entitled "A Neural Network Approach for Classification of Fault-Slip Data in Geoscience" applied neural networks for the classification of geoscience data related to fault-slip data. Although this study focuses on geoscience, the use of classification methods such as neural networks can provide useful ideas in using similar classification techniques on medical data, such as dengue patient data in this study [29].

Nguyen et al. (2024) in a study titled "An Ensemble Model of Logistic Regression, Naïve Bayes, and Adaboost for Assessing Landslide Spatial Probability" combined Naïve Bayes with other algorithms in an ensemble model to predict the likelihood of landslides. This research shows how the combination of various algorithms can improve prediction accuracy, which can be applied in your context to improve medical data classification results using Naïve Bayes and PCA [30]. El Mahjouby et al. (2024) in a study titled "Machine Learning Algorithms for Forecasting and Categorizing Euro-to-Dollar Exchange Rates" focused on using machine learning algorithms to predict and classify the Euro to Dollar exchange rate. Although this research focuses on economic prediction, the use of machine learning techniques for classification in an economic context can be applied in the medical field, especially in classifying patient data for disease prediction [31]. Abdlkader and Ghanim (2024) in their research entitled "Design and Analysis of Face Recognition System Based on VGG-Face-16 with Various Classifiers" focused on face recognition using algorithm-based classification systems such as Naïve Bayes and VGG-Face-16. Although this focuses more on face recognition technology, the use of classification methods to identify patterns in data is highly relevant to your research, which also uses classification algorithms for medical data [32]. Qahar et al. (2024) in a study titled "Factor Analysis Influencing Mobile JKN User Experience Using Sentiment Analysis" used Naïve Bayes for sentiment analysis in the JKN application. This study shows how classification algorithms are used to assess user experience in health-based applications. A similar approach can be applied to your research, which focuses on disease classification in medical data, although the context is different in terms of user data analysis [33].

Li et al. (2024) in a study titled "Mapping Urban Floods via Spectral Indices and Machine Learning Algorithms" used machine learning to map urban floods through satellite images. This research uses classification algorithms to process image data and produce flood maps. Although the focus is not on medical data, the image processing and classification techniques used can provide guidance in your research that focuses on medical data classification using Naïve Bayes [34]. Abdullah et al. (2024) in their research entitled "Development of a Machine Learning Algorithm for Fake News Detection" focused on fake news detection using Naïve Bayes algorithm and ensemble method. Although the topic analyzed was fake news, the use of classification algorithms such as Naïve Bayes to detect patterns in data can be adapted for medical data classification, as was done in your study of dengue fever patients [35].

Zada et al. (2024) in a study titled "Fine-Tuning Cyber Security Defenses: Evaluating Supervised Machine Learning Classifiers for Windows Malware Detection" tested various supervised learning algorithms, including Naïve Bayes, to detect malware on Windows systems. Although focused on cybersecurity, the classification techniques used in this study can be applied to improve classification in medical data [36]. Anamisa et al. (2024) in a study entitled "Performance Test of Naïve Bayes and SVM Methods on Classification of Malnutrition Status in Children" tested the use of Naïve Bayes and SVM to classify malnutrition status in children. This research shows how both algorithms can be applied in the classification of malnutrition-related medical data. Although the types of diseases classified are different, the techniques used in this study are very relevant for the classification of patient data in your study [37].

Barracloug et al. (2024) in a study entitled "Artificial Intelligence System for Malaria Diagnosis" focused on the use of artificial intelligence systems to diagnose malaria. The use of machine learning algorithms, including Naïve Bayes, to diagnose this disease has similarities to the approach you use to diagnose and classify dengue patient data, although the type of disease is different [38]. Jamil et al. (2024) in a study titled "Sentiment Analysis: Classifying Public Comments on YouTube in Disaster Management Simulation in Indonesia Using Naïve Bayes and Support Vector Machine" used Naïve Bayes and SVM to classify public comments related to disaster management. Although focused on analyzing public comments, the classification approach used can be applied in the classification of medical data in your research [39]. CheSuh et al. (2024) in their research entitled "Improve Quality of Service for the Internet of Things Using Blockchain & Machine Learning Algorithms" used machine learning to improve the quality of IoT services with the help of blockchain technology. Although the focus is on managing IoT services, the techniques used to classify the data can provide insight into how to apply classification to large datasets, such as the dengue patient data [40]. Arsalane et al. (2024) in their study entitled "Performance Evaluation of Machine Learning Algorithms for Meat Freshness Assessment" focused on the use of machine learning to assess meat freshness. Although not focused on diseases, the application of classification algorithms such as Naïve Bayes to assess the quality of food products can provide useful ideas for the classification of medical data in your research [41].

3. Method

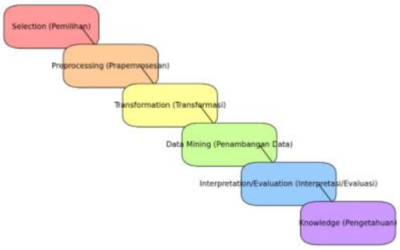

Selection (Pemilihan)
Preprocessing (Prapemrosesan)
Transformation (Transformasi)
Data Mining (Penambangan Data)
Interpretation/Evaluation (Interpretasi/Evaluasi)
Knowledge (Pengetahuan)

Fig. 1: Stages of Knowledge Discovery in Databases (KDD)

Fig.1 explaining the stages of Knowledge Discovery in Databases (KDD). This research is conducted through an experimental method approach that will be applied to classification in machine learning with the Naïve Bayes algorithm to explore and verify the proposed hypothesis. This approach was chosen to provide a comprehensive and in-depth understanding of the proposed research topic. Experimental methods in machine learning are applied to evaluate and compare the performance of various classification algorithms and data processing techniques. Experimental methods in classification are an important approach in scientific research to identify patterns and relationships between variables. In the context of classification, there are various methods that can be used, such as K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Naïve Bayes (NB), Convolutional Long Short-Term Memory (LSTM), and so on [42]-[43]-[44]. The stages carried out in this experimental method use the stages of Knowledge Discovery in Databases (KDD).

3.1. Data Selection

In this data selection stage, determining and selecting relevant data for analysis, often involves collecting data from various sources and ensuring that the data matches the desired analysis objectives [45].

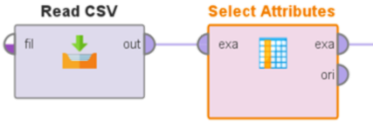


Fig. 2: Step of Data Selection

Based on Fig. 2, there are two initial stages performed in the data processing process, namely reading data from external sources and selecting relevant attributes. This stage is an important part of the Selection step in the Knowledge Discovery in Databases (KDD) framework, where the available raw data is transformed into more structured and relevant data for further analysis. Read CSV, aims to read data from files in CSV (Comma Separated Values) format. The CSV format is often used to store tabular data, where rows represent entries or observations, while columns contain associated attributes or variables. In this process, data is imported from external sources into the analysis system, so that it can be used for subsequent steps. The output of this block is a raw dataset containing all the attributes and records from the source file. This stage is fundamental, as it ensures that the initial data is available in a suitable format for further processing.

Once the data has been successfully imported, the next stage is the attribute selection process, represented by the Select Attributes block. This process is designed to select attributes that are relevant to the purpose of the analysis, based on certain criteria such as statistical relevance, data quality, or research focus. Not all attributes in the original dataset may have high informative value or relevance to the analysis objective. Therefore, this selection process aims to filter out insignificant attributes, so that only relevant and high-quality attributes are carried forward to the next stage. The end result of this process is a more focused dataset, with fewer attributes that still represent important information from the original dataset.

3.2. Data Preprocessing

Data Processing the selected data to improve its quality, including data cleaning, handling missing values, normalization, and data transformation to make the data ready for further analysis [46].

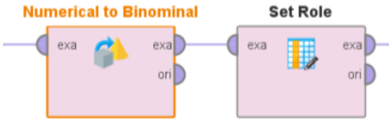


Fig. 3: Stages of Data Preprocessing

Fig. 3 explains that the parameters used in the “Numerical to Binominal” operator to convert numerical data into binary data. The first parameter, Attribute filter type, indicates that the attribute filtering or selection process is performed on a subset of the available attributes in the dataset. This means that not all available attributes will be processed, but only a certain subset that is relevant to the purpose of the conversion. Furthermore, the Attributes parameter refers to the selection of attributes that will be selected for conversion. Only those attributes that have been selected based on certain criteria, such as those with numeric values or those that meet a specified threshold, will be converted into binary form. This process generates binary data where the selected attributes will be categorized into two values, e.g. “1” for categories that are higher than the threshold and “0” for categories that are lower. Using these parameters, the Numerical to Binominal operator can efficiently convert relevant attributes for analysis or machine learning models, thereby improving the accuracy and efficiency of data processing. After doing numericel to binominal, the next step is Set Role, which serves to define the role of each attribute in the dataset, such as determining the target attribute (dependent variable) and predictor attribute (independent variable). Defining the role of these attributes is very important to ensure a clear and organized data structure according to the needs of the analysis or machine learning model to be used. In research, the target attribute is usually the main focus to be analyzed or predicted, while the predictor attribute is used to model the relationship to the target. This process provides clarity on the function of the attributes, thus supporting the success of the analysis effectively.

These two steps have a complementary relationship. Numerical to Binominal is performed first to ensure the data format is appropriate, especially if numeric attributes require transformation. Next, Set Role provides a clear structure to the dataset by explicitly defining the role of each attribute. Overall, these two stages ensure that the resulting dataset is ready for further analysis, with a data structure that is relevant, organized, and in line with the needs of the analysis methodology in the research framework.

3.3. Data Transformation

Mengubah atau merubah bentuk data yang telah diproses, seperti pengurangan dimensi atau pembentukan fitur baru, untuk memudahkan penerapan algoritma data mining [47].

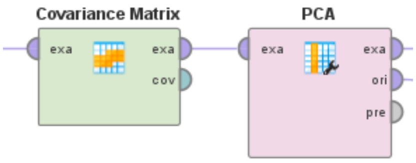


Fig. 4:Stage of Data Transformation

Fig.4 shows two important stages in data transformation, namely Covariance Matrix and Principal Component Analysis (PCA). The first stage is the calculation of the covariance matrix using the Covariance Matrix components, which aims to understand the relationship between variables in the data. This covariance matrix provides information on how two variables change together, with positive values indicating a direct relationship and negative values indicating an inverse relationship. This process accepts input data in the form of example data and produces two main outputs: unaltered example data to be passed on to the next process and a covariance matrix that is used for further analysis.

3.4. Data Mining

Using data mining techniques to find patterns or information from data. This involves applying various algorithms, such as classification, clustering, or association, to extract knowledge [48]-[49].

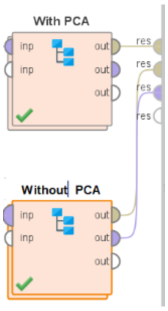


Fig. 5: Stages of Data Mining

Fig. 5 shows the data analysis comparison process between two paths: with PCA (Principal Component Analysis) and without PCA. This study aims to evaluate the effect of using PCA on the performance of the analysis process. In the path with PCA, the data first undergoes dimensional reduction using PCA before further processing. This dimensional reduction aims to simplify the data by reducing the number of variables while still maintaining the main information. PCA eliminates insignificant attributes and only leaves components that have a large contribution to data variance. The data resulting from this dimensional reduction is then used in the analysis process which is indicated by the output of this pathway.

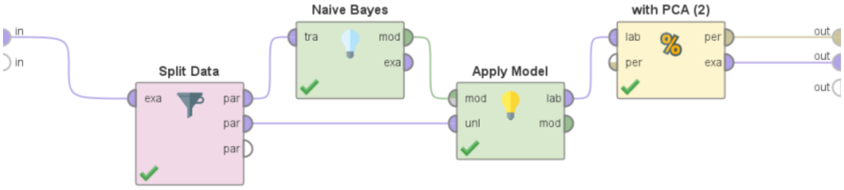


Fig. 6: Data Mining Process using PCA

Fig. 6 shows the process of data analysis based on Machine Learning algorithms. The process starts with the Split Data module, which is used to split the input data into two subsets, namely training data (training set) and testing data (testing set). Next, the training data is used to train a classification model using the Naïve Bayes algorithm, which is a probability-based method to predict the target class based on input features. Once the model is trained, the Apply Model module is applied to test the performance of the model on the testing data. The figure also shows the application of the Principal Component Analysis (PCA) method in the data analysis process. PCA is used to reduce the dimensionality of the data before the Naïve Bayes algorithm is applied. In this process flow, PCA is not explicitly seen as a separate module, but its integration is visible from the process name, which is "with PCA". This indicates that the previous process has included a data processing step using PCA to reduce the number of features used in the analysis.

3.5. Evaluation and Interpretation

The evaluation and interpretation stages in the Knowledge Discovery in Databases (KDD) process are important steps to ensure the results of data mining analysis are relevant and provide added value according to the initial objectives. In the evaluation stage, data mining results are assessed in terms of quality, accuracy, and relevance to the analysis objectives using evaluation metrics such as accuracy, precision, and F1-score. This process also involves validating the results to ensure reliability, identification of significant patterns, and avoidance of errors such as overfitting or data bias. After that, the interpretation stage aims to understand and communicate the results to stakeholders. The results that have been evaluated are compiled and visualized in the form of graphs, diagrams, or reports so that they can be clearly understood. Ultimately, these two stages complement each other to ensure that the findings are not only technically valid, but also able to provide relevant insights and support decision-making [50].

3.6. Knowledge

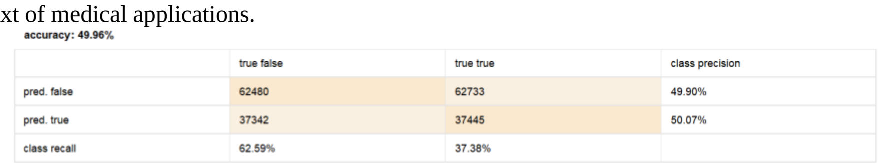
The knowledge stage in the Knowledge Discovery in Databases (KDD) process is the final step that aims to transform the results of data mining analysis into knowledge that can be understood and utilized by stakeholders for more effective decision-making. This stage is not just about presenting the results, but also integrating the insights into a specific context that is relevant to the needs of the organization or users. Presentation is done in a clear format, such as detailed reports, data visualizations, or interactive presentations, to ensure that the information produced is easily accessible and makes a real impact. This process involves several important elements. First, the use of data visualizations, such as bar charts, heat maps, line graphs, or network graphs, which help simplify the interpretation of complex patterns. These visualizations allow non-technical users to understand trends and relationships discovered during analysis. Second, the preparation of the report, which usually includes the methodology, key findings, evaluation of model performance, and strategic recommendations. The report should be designed with clear language, without sacrificing important technical details, so that it can reach audiences with various backgrounds.

Furthermore, contextualizing the results is important to ensure that the insights gained are relevant to the original problem or objective. For example, in a business context, analysis results can be linked to marketing strategies, risk management, or optimization of operational processes. Finally, interactive presentation, such as through web- or app-based dashboards, further facilitates independent exploration of the data by stakeholders. This stage not only completes the KDD cycle, but also serves as a bridge between the insights generated and their real-world application. By ensuring that the results presented are clear, accurate and relevant, the knowledge stage supports the transformation of data into strategic decisions that underpin innovation and sustainable growth [51].

4. Result And Discussion

4.1. Result

The results of applying the Naïve Bayes algorithm in the classification of dengue fever patient data, by comparing the performance of models that use Principal Component Analysis (PCA) and models that do not use PCA. The main objective of this study is to evaluate the extent to which the use of PCA, as a dimension reduction technique, can improve the effectiveness of classification models in terms of accuracy and precision. The application of PCA is expected to reduce data complexity by eliminating irrelevant or redundant variables, which in turn can improve computational efficiency as well as reduce the possibility of overfitting. In contrast, a model that does not use PCA will perform Naïve Bayes on the original data without any dimensionality reduction. The comparison between the two approaches is expected to provide deeper insights into the role of PCA in improving classification accuracy, as well as its implications in accelerating and improving the accuracy of dengue diagnosis. The results obtained will be critically analyzed to assess the practical contribution of applying PCA in the context of medical applications.



accuracy: 49.96%

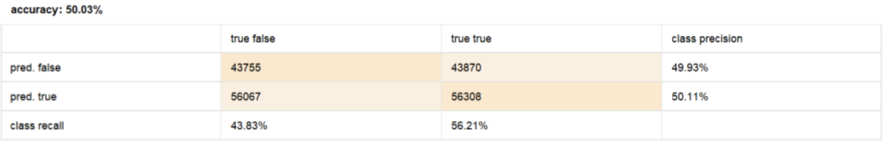
	true false	true true	class precision
pred. false	62480	62733	49.90%
pred. true	37342	37445	50.07%
class recall	62.59%	37.38%	

Fig. 7: Classification Evaluation Result of Naïve Bayes Algorithm without PCA

Based on the evaluation results in Figure 4.9, the naïve bayes classification model without PCA tested shows unsatisfactory performance with an accuracy of 49.96%, which is almost equivalent to the random guess rate. This indicates that the model has not been able to identify patterns in the data to produce reliable predictions. From the confusion matrix, it was found that the model successfully predicted 62,480 samples as “false” correctly (true negatives), but made the mistake of predicting 37,342 samples as “true” even though they were actually “false” (false positives). For the “true” class, the model correctly predicted 37,445 samples (true positives), but incorrectly predicted 62,733 samples as “false” even though they were actually “true” (false negatives). The larger number of false negatives compared to true positives indicates that the model had difficulty in recognizing samples that were actually “true.”

The model precision for the “false” class was 49.90%, which means that almost half of the “false” predictions generated by the model were correct. The precision for the “true” class was slightly higher at 50.07%, indicating that about half of the “true” predictions were correct. This precision value of only around 50% indicates that the model was not able to consistently distinguish between the “true” and “false” classes. Recall or sensitivity also shows a significant discrepancy. The recall for the “false” class reaches 62.59%, which means that the model is able to recognize more than half of the samples that are truly “false” well. However, the recall for the “true” class was only 37.38%, indicating that the model often failed to identify samples that should belong to the “true” class. The disparity between the recall values of these two classes indicates that the model tends to be biased towards the “false” class, thus predicting samples as “false” more often than “true.”

Overall, the model's performance showed fundamental weaknesses in accuracy, precision, and recall. The near-random accuracy indicates that the model did not successfully capture significant relationships in the data. In addition, the low precision and recall, especially for the “true” class, indicate that the model is unable to reliably classify samples with the “true” label. The model's bias towards the “false” class is also a concern, which may impact its use in real situations where recognition of the “true” class may be more important.



accuracy: 50.03%

	true false	true true	class precision
pred. false	43755	43870	49.93%
pred. true	56067	56308	50.11%
class recall	43.83%	56.21%	

Fig. 8: Classification Evaluation Result of Naïve Bayes Algorithm With PCA

Fig.6 describes the results of Classification Evaluation Result of Naïve Bayes Algorithm With Principal Component Analysis (PCA) as a dimension reduction technique, the performance of the classification model shows a slight improvement compared to before. The model accuracy was recorded at 50.03%, which is still at a level close to random guessing. PCA, which serves to simplify the data by reducing feature redundancy, allows the model to work more efficiently. However, the evaluation results show that this dimensionality reduction has not had a significant impact on the model's ability to distinguish between the “true” and “false” classes. From the confusion matrix, the model was able to correctly predict 43,755 samples as “false” (true negatives) and 56,308 samples as “true” (true positives). However, the model still incorrectly predicted 56,067 samples as “true” even though they were actually “false” (false positives), and 43,870 samples as “false” even though they were actually “true” (false negatives).

The precision and recall values provide a more detailed picture of the model's performance after PCA. The precision for the “false” class was recorded at 49.93%, which means that only about half of the “false” predictions matched the original label. For the “true” class, the precision was slightly better at 50.11%, showing similar results. The recall for the “false” class is 43.83%, which indicates that the model only recognizes a fraction of the samples with the label “false.” In contrast, the recall for the “true” class reached 56.21%, indicating that the model was better able to recognize samples with the label “true.” This discrepancy indicates the model's bias towards the “true” class, which may be influenced by the data distribution or certain feature characteristics.

Table 1: PCA Experiment

NO.	Dimensionality reduction	value	Accurasy Without PCA	Accurasy With PCA
1.	Fixed number	1	49.96%	49.91%
2.	Fixed number	2	49.96%	49.97%
3.	Fixed number	3	49.96%	50.03%
4.	Fixed number	4	49.96%	49.97%
5.	Fixed number	5	49.96%	49.97%
6.	Fixed number	6	49.96%	49.97%
7.	Fixed number	7	49.96%	49.97%
8.	Fixed number	8	49.96%	49.97%
9.	Fixed number	9	49.96%	49.97%
10.	Fixed number	10	49.96%	49.97%
11.	keep variance	0.1	49.96%	49.91%
12.	keep variance	0.2	49.96%	49.91%
13.	keep variance	0.3	49.96%	49.97%
14.	keep variance	0.4	49.96%	49.97%
15.	keep variance	0.5	49.96%	49.97%
16.	keep variance	0.6	49.96%	50.03%
17.	keep variance	0.7	49.96%	50.03%
18.	keep variance	0.8	49.96%	50.03%
19.	keep variance	0.9	49.96%	50.03%
20.	keep variance	1.0	49.96%	50.03%
21.	none			49.9%

Based on table 1, the PCA Application Process is carried out first in determining the value of dimension reduction fix number and dimension reduction keep variance. Experiments on dimension reduction fix number and keep variance are carried out 10 times each with the value of the fix number value from 1 to 10, for the value of the keep variance value starting from 0.1 to 1.0. After the experiment, the value value is obtained which has the highest value in the application of PCA in dimension reduction fix number with a value of 3 with an accuracy value of 50.03%. In the experiment in dimension reduction keep variance obtained five values that produce high accuracy value in the application of PCA including value 0.6, 0.7, 0.8, 0.9, and 1.0 with the acquisition of accuracy value of 50.03%. Based on this acquisition, the highest accuracy value has the same value of 50.03%, so in this study only uses a dimension reduction fix number with a value of 3 in the application of PCA.

Overall, although PCA succeeded in simplifying the data by reducing the number of features used by the model, the application of this technique was not effective enough to significantly improve the model performance. Despite small improvements in accuracy, precision and recall, the model still struggled to reduce prediction errors, especially for the minority ("false") class. This suggests that PCA may not be suitable for data with more complex non-linear relationships, or that deeper issues such as class imbalance have not been resolved. Therefore, to achieve more meaningful performance improvements, additional measures such as data balancing, use of more sophisticated algorithms, or hyperparameter tuning should be considered.

4.2. Discussion

In the study of the application of the Naïve Bayes algorithm in dengue disease classification, the results were disappointing, both in the model that used Principal Component Analysis (PCA) and the one that did not use the technique. In the model without PCA, the accuracy of only 49.96% is almost equivalent to random guessing, indicating the inability of the model to effectively distinguish between the "true" and "false" classes. The apparent imbalance in precision and recall values is more indicative of a bias towards the "false" class, with higher recall for this class compared to the "true" class. This phenomenon is highly relevant to the existing discussion in the literature on how Naïve Bayes, although often praised for its simplicity, can have difficulties on datasets that have significant class imbalance [25]-[52]. When PCA was applied to the model, although there was a slight improvement in accuracy (50.03%), the overall performance was still inadequate, and the model still struggled to identify samples from the "true" class. PCA aims to simplify the data by reducing its dimensionality, but the evaluation results showed that the application of this technique was not effective enough in addressing the issues present in this dataset. A number of studies have also reported that while PCA can reduce data complexity, its effectiveness is highly dependent on the nature of the relationships between features in the data. PCA may not provide optimal results on data that has complex non-linear relationships [14]-[53]. Thus, despite the slight improvement, PCA does not seem to be sufficient to address the main issues in the dataset, especially class imbalance which may affect the overall classification performance.

This research is in line with previously published studies on the application of PCA and Naïve Bayes to medical datasets. For example, a study by Gunawardana et al. in 2024 revealed that the use of simple algorithms such as Naïve Bayes can give limited results on datasets with complex or imbalanced features, such as in the case of disease classification. They suggested the use of more sophisticated algorithms or more robust pre-processing techniques to deal with such problems [52]. In addition, research by Hameed et al. in 2024 entitled Leukemia Diagnosis using Machine Learning Classifiers based on MRMR Feature Selection also emphasized the importance of using appropriate feature selection techniques to improve algorithm performance on medical datasets [25]. In this context, the application of PCA provides only a slight improvement, as the technique is unable to address the deeper class imbalance issues present in dengue datasets. In this regard, further literature provides an understanding that although PCA successfully reduces data dimensionality and redundancy, the application of PCA on more complex datasets with non-linear feature relationships or datasets with class imbalance may result in suboptimal performance [17]. Research conducted by Mahendra et al. in 2024 also applied Naïve Bayes in diabetes mellitus classification, they found that dimensionality reduction does not always provide better results if the class imbalance problem has not been addressed. They put more emphasis on applying data balancing techniques or using more complex algorithms, such as AdaBoost, to improve model accuracy [20]. Meanwhile, Altayeb and Arabiat's research showed that PCA can improve accuracy in some cases, but only when the relationship between features is fairly linear and there is no obvious class imbalance problem [14]. This suggests that this study does not fully match the results obtained by other researchers in the literature, who often point out that PCA is not the sole solution for improving model performance on complex datasets.

Overall, these findings have significant implications in both scientific and practical contexts. From a scientific point of view, the results of this study reveal that although PCA can simplify data by reducing dimensionality, its application in dengue disease classification using Naïve Bayes is not sufficient to address the main problem present in the dataset, which is significant class imbalance. This suggests that PCA provides little benefit to datasets with complex or non-linear relationships. As described by Hameed et al. in 2024, PCA is more effective when the dataset has a linear structure and is not affected by extreme class imbalance issues [25]. From a practical perspective, these results are highly relevant for medical professionals looking to develop machine learning-based diagnostic tools for diseases such as dengue fever. The use of Naïve Bayes without data balancing techniques or more sophisticated algorithms may result in inaccurate predictions, especially for minority classes such as "true" in this dataset. This finding confirms the importance of adopting a more holistic approach in model development, as suggested by Gunawardana et al. in 2024, who suggested using a combination of pre-processing techniques and more sophisticated algorithms in cases like this [52]. In addition, the use of class balancing techniques or even ensemble algorithms such as AdaBoost used by Mahendra et al. in 2024 may be more beneficial in correcting the class imbalance that occurs [20].

Another practical implication is the importance of considering a diversity of techniques in medical data modeling, especially in diseases with complex and imbalanced data. For example, research by Demilie in 2024 entitled Plant disease detection and classification techniques: a comparative study of the performances showed that a combination of more effective feature selection techniques and the application of class balancing methods can further improve classification accuracy [53]. Therefore, although PCA can simplify the model, a more thorough and integrated approach, including the use of optimization techniques and selection of appropriate algorithms, is required to produce a model that can be used in real medical applications. This research also opens up opportunities for further exploration of other techniques that can be used to improve accuracy, such as hyperparameter optimization or the use of hybrid methods that combine various machine learning algorithms to mitigate the weaknesses present in a single approach [23]. Overall, these findings provide important insights into how to address the issues involved in applying machine learning algorithms for disease diagnosis, while also highlighting the challenges to be faced in future similar research.

5. Conclusion

This study evaluates the performance of the Naïve Bayes algorithm in classifying dengue fever patient data, by comparing models that use Principal Component Analysis (PCA) as a dimension reduction technique and models without PCA. Based on the evaluation results, the Naïve Bayes model without PCA shows an accuracy of 49.96%, which is almost equivalent to a random guess. This indicates that the model is unable to effectively identify patterns in the data. On the other hand, applying PCA increased the accuracy of the model to 50.03%, but this increase was not significant in improving the classification performance.

Further analysis showed that both models with and without PCA were biased towards the dominant ("false") class. Precision for both classes was around 50%, while recall showed significant discrepancies. In the model without PCA, the recall of the "false" class reached 62.59%, but the recall of the "true" class was only 37.38%. After the application of PCA, the bias towards the "false" class was reduced, but the performance gap between classes was still visible, with the recall of the "true" class reaching 56.21% and the recall of the "false" class dropping to 43.83%. These results indicate that the application of PCA, while reducing the dimensionality of the data, is not sufficient to address the main issues in the dataset, especially related to the unbalanced label distribution.

This Study makes an important contribution in understanding the limitations of the Naïve Bayes algorithm and dimensionality reduction techniques such as PCA in medical classification applications. It also emphasizes the importance of using diverse evaluation metrics to thoroughly understand model performance. From a practical perspective, the findings provide insights for the development of more advanced and adaptive machine learning-based diagnostic tools. Finally, this study opens up opportunities for exploration of techniques such as data balancing and ensemble algorithms to improve model performance in the future..

Acknowledgement

All praise and gratitude go to Allah SWT for all His grace, gifts and guidance, so that the author can complete this research well.
During the process of writing the results of this research, many parties have been involved in helping and providing support and guidance to the author. On this occasion, I would like to say thank you to:

1. Mr. Assoc. Prof. Dr. Dadang Sudrajat, S.Si., M.Kom, as Chairman of STMIK IKMI Cirebon.
2. Mr. Dian Ade Kurnia, M.Kom, as Vice Chairman I for Academic Affairs, Cooperation, Research and Innovation.
3. Mrs. Dra. Nining R, M.Si., as Vice Chairman II of Finance.
4. Mrs. Fatihanursari Dikananda, S.Tr.I.Kom., M.Kom as Vice Chair III for Student Affairs and Alumni.
2. Mr. H. Eka Jayawangsa, BBA, as Vice Chair IV for Facilities and Infrastructure.
3. Mrs. Gifthera Dwilestari, S.I.Kom., M.Kom, as the Head of the Informatics Engineering Study Program.
4. Mr. Rudi Kurniawan, MT as the Main Supervisor.
5. Mr. Yudhistira Arie Wijaya, M.Kom as Second Supervisor lecturer.
6. Friends and all those who have helped.
7. My mother, my father and family who have provided support and prayers during my college journey.here.

References

[1] L. Chaves, "Data mining techniques for early diagnosis of diabetes: a comparative study," *Appl. Sci.*, vol. 11, no. 5, p. 2218, 2021, doi: 10.3390/app11052218.
[2] C. Fan, M. Chen, X. Wang, J. Wang, and B. Huang, "A Review on Data Preprocessing Techniques Toward Efficient and Reliable Knowledge Discovery From Building Operational Data," *Front. Energy Res.*, vol. 9, 2021, doi: 10.3389/fenrg.2021.652801.
[3] A. Mohammed, "Decision tree, naïve bayes and support vector machine applying on social media usage in nyc / comparative analysis," *Tikrit J. Pure Sci.*, vol. 22, no. 9, pp. 94–99, 2023, doi: 10.25130/tjps.v22i9.881.
[4] M. Ragab, "Optimized Classification Model for Biomedical Data Analysis," *Ann. Adv. Biomed. Sci.*, vol. 6, no. 1, 2023, doi: 10.23880/aabsc-16000204.
[5] S. P. Simelane, C. Hansen, and C. Munghemezulu, "The Use of Remote Sensing and GIS for Land Use and Land Cover Mapping in Eswatini: A Review," *South African J. Geomatics*, vol. 10, no. 2, pp. 181–206, 2022, doi: 10.4314/sajg.v10i2.13.
[6] V. Vijay Anuradha, N. Anbalagan, "Clinical presentation and platelet profile of dengue fever: a retrospective study," *Cureus*, 2022, doi: 10.7759/cureus.28626.
[7] C. Ouattara, "Spatio-temporal determinants of dengue epidemics in the central region of burkina faso," *Trop. Med. Infect. Dis.*, vol. 8, no. 11, p. 482, 2023, doi: 10.3390/tropicalmed8110482.
[8] N. Hamdani Hatta, H., Puspitasari, A. Septiarini, and H. Henderi, H., "Dengue classification method using support vector machines and cross-validation techniques," *Iaes Int. J. Artif. Intell.*, vol. 11, no. 3, p. 1119, 2022, doi: 10.11591/ijai.v11.i3.pp1119-1129.
[9] Y. Salim Wah, C. Reeves, M. Smith, W. Yaacob, R. Mudin, and N. Haque, U., "Prediction of dengue outbreak in selangor malaysia using machine learning techniques," *Sci. Rep.*, vol. 11, no. 1, 2021, doi: 10.1038/s41598-020-79193-2.
[10] Y. A. Wijaya, N. Suarna, Iin, R. Hamonangan, and R. Nining, "Comparison of machine learning algorithm for Santander dataset," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 1088, no. 1, p. 012032, 2021, doi: 10.1088/1757-899x/1088/1/012032.
[11] R. Tarigan, "Artificial neural network for classification of dengue fever using backpropagation algorithm," *J. Artif. Intell. Eng. Appl.*, vol. 3, no. 1, pp. 468–478, 2023, doi: 10.59934/jaiea.v3i1.357.
[12] A. and P. Rahman S., "Performance analysis of the hybrid voting method on the classification of the number of cases of dengue fever," *Int. J. Inf. Commun. Technol.*, vol. 8, no. 1, pp. 10–19, 2022, doi: 10.21108/ijoict.v8i1.614.
[13] N. A. Salim *et al.*, "Prediction of Dengue Outbreak in Selangor Malaysia Using Machine Learning Techniques," *Sci. Rep.*, vol. 11, no. 1, 2021, doi: 10.1038/s41598-020-79193-2.
[14] M. Altayeb and A. Arabiat, "Enhancing stroke prediction using the waikato environment for knowledge analysis," *IAES Int. J. Artif. Intell.*, vol. 13, no. 3, pp. 3010–3017, 2024, doi: 10.11591/ijai.v13.i3.pp3010-3017.
[15] M. Shenify, "Sentiment analysis of Saudi e-commerce using naïve bayes algorithm and support vector machine," *Int. J. Data Netw. Sci.*, vol. 8, no. 3, pp. 1607–1612, 2024, doi: 10.5267/j.ijdns.2024.3.006.
[16] O. Alghushairy *et al.*, "An Efficient Support Vector Machine Algorithm Based Network Outlier Detection System," *IEEE Access*, vol. 12, pp. 24428–24441, 2024, doi: 10.1109/ACCESS.2024.3364400.
[17] M. A. Hassan, A. H. Muse, and S. Nadarajah, "Predicting Student Dropout Rates Using Supervised Machine Learning: Insights from the 2022 National Education Accessibility Survey in Somaliland," *Appl. Sci.*, vol. 14, no. 17, 2024, doi: 10.3390/app14177593.
[18] D. Monteverde-Suárez *et al.*, "Predicting students' academic progress and related attributes in first-year medical students: an analysis with artificial neural networks and Naïve Bayes," *BMC Med. Educ.*, vol. 24, no. 1, 2024, doi: 10.1186/s12909-023-04918-6.
[19] A. A. Stonier, R. K. Gorantla, and K. Manoj, "Cardiac disease risk prediction using machine learning algorithms," *Healthc. Technol. Lett.*, vol. 11, no. 4, pp. 213–217, 2024, doi: 10.1049/htl2.12053.
[20] I. G. A. P. Mahendra, I. M. A. Wirawan, and I. G. A. Gunadi, "Enhancement performance of the Naïve Bayes method using AdaBoost for classification

of diabetes mellitus dataset type II," *Int. J. Adv. Appl. Sci.*, vol. 13, no. 3, pp. 733–742, 2024, doi: 10.11591/ijaas.v13.i3.pp733-742.

[21] Y. K. Saheed, T. O. Kehinde, M. Ayobami Raji, and U. A. Baba, "Feature selection in intrusion detection systems: a new hybrid fusion of Bat algorithm and Residue Number System," *J. Inf. Telecommun.*, vol. 8, no. 2, pp. 189–207, 2024, doi: 10.1080/24751839.2023.2272484.

[22] A. K. Singh, R. Sunkara, G. R. Kadambi, and V. Palade, "Spectral-Spatial Classification With Naive Bayes and Adaptive FFT for Improved Classification Accuracy of Hyperspectral Images," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 17, pp. 1100–1113, 2024, doi: 10.1109/JSTARS.2023.3327346.

[23] D. Morakis and A. Adamopoulos, "Hybrid Machine Learning Algorithms to Evaluate Prostate Cancer," *Algorithms*, vol. 17, no. 6, 2024, doi: 10.3390/a17060236.

[24] S. Sulak and N. Koklu, "Analysis of Depression, Anxiety, Stress Scale (DASS-42) With Methods of Data Mining," *Eur. J. Educ.*, vol. 59, no. 4, 2024, doi: 10.1111/ejed.12778.

[25] S. M. Hameed, W. A. Ahmed, and M. A. Othman, "Leukemia Diagnosis using Machine Learning Classifiers based on MRMR Feature Selection," *Eng. Technol. Appl. Sci. Res.*, vol. 14, no. 4, pp. 15614–15619, 2024, doi: 10.48084/etasr.7720.

[26] M. Y. Shams, S. A. Gamel, and F. M. Talaat, "Enhancing crop recommendation systems with explainable artificial intelligence: a study on agricultural decision-making," *Neural Comput. Appl.*, vol. 36, no. 11, pp. 5695–5714, 2024, doi: 10.1007/s00521-023-09391-2.

[27] S. Khasim, I. S. Rahat, H. Ghosh, K. Shaik, and S. K. Panda, "Using Deep Learning and Machine Learning: Real-Time Discernment and Diagnostics of Rice-Leaf Diseases in Bangladesh," *EAI Endorsed Trans. Internet Things*, vol. 10, 2024, doi: 10.4108/eetiot.4579.

[28] Z. Albataineh, F. Aldrweesh, and M. A. Alzubaidi, "COVID-19 CT-images diagnosis and severity assessment using machine learning algorithm," *Cluster Comput.*, vol. 27, no. 1, pp. 547–562, 2024, doi: 10.1007/s10586-023-03972-5.

[29] S. Yaman, B. Karakaya, and M. Köküm, "A neural network approach for classification of fault-slip data in geoscience," *Ain Shams Eng. J.*, vol. 15, no. 1, 2024, doi: 10.1016/j.asej.2023.102325.

[30] B.-Q.-V. Nguyen, L.-H.-P. Ho, and Y.-T. Kim, "An Ensemble Model of Logistic Regression, Naïve Bayes, and Adaboost for Assessing the Landslide Spatial Probability-Study Case: Phuoc Son, Quang Nam, Vietnam and Umyeon, Seoul, Korea," *Civ. Eng. Archit.*, vol. 12, no. 3, pp. 2010–2028, 2024, doi: 10.13189/cea.2024.121307.

[31] M. El Mahjouby, M. Taj Bennani, M. Lamrini, B. Bossoufi, T. A. H. Alghamdi, and M. El Far, "Machine Learning Algorithms for Forecasting and Categorizing Euro-to-Dollar Exchange Rates," *IEEE Access*, vol. 12, pp. 74211–74217, 2024, doi: 10.1109/ACCESS.2024.3404824.

[32] D. F. Abdlkader and M. F. Ghanim, "Design and analysis of face recognition system based on VGG-Face-16 with various classifiers," *IAES Int. J. Artif. Intell.*, vol. 13, no. 2, pp. 1499–1510, 2024, doi: 10.11591/ijai.v13.i2.pp1499-1510.

[33] M. Y. A. Qahar, Y. Ruldeviyani, U. N. Mukharomah, M. A. Fidyawan, and R. Putra, "Factor analysis influencing Mobile JKN user experience using sentiment analysis," *IAES Int. J. Artif. Intell.*, vol. 13, no. 2, pp. 1782–1793, 2024, doi: 10.11591/ijai.v13.i2.pp1782-1793.

[34] L. Li, A. Woodley, and T. Chappell, "Mapping Urban Floods via Spectral Indices and Machine Learning Algorithms," *Sustain.* , vol. 16, no. 6, 2024, doi: 10.3390/su16062493.

[35] N. A. S. Abdullah, N. I. A. Rusli, and N. S. Yuslee, "Development of a machine learning algorithm for fake news detection," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 35, no. 3, pp. 1732–1743, 2024, doi: 10.11591/ijeecs.v35.i3.pp1732-1743.

[36] I. Zada *et al.*, "Fine-Tuning Cyber Security Defenses: Evaluating Supervised Machine Learning Classifiers for Windows Malware Detection," *Comput. Mater. Contin.*, vol. 80, no. 2, pp. 2917–2939, 2024, doi: 10.32604/cmc.2024.052835.

[37] D. R. Anamisa, A. Jauhari, and F. A. Mufarroha, "PERFORMANCE TEST OF NAIVE BAYES AND SVM METHODS ON CLASSIFICATION OF MALNUTRITION STATUS IN CHILDREN," *Commun. Math. Biol. Neurosci.*, vol. 2024, 2024, doi: 10.28919/cmbn/8429.

[38] P. A. Barracloug *et al.*, "Artificial Intelligence System for Malaria Diagnosis," *Int. J. Adv. Comput. Sci. Appl.*, vol. 15, no. 3, pp. 920–932, 2024, doi: 10.14569/IJACSA.2024.0150392.

[39] M. Jamil, H. Hadiyanto, and R. Sanjaya, "Sentiment Analysis: Classifying Public Comments on YouTube in Disaster Management Simulation in Indonesia Using Naïve Bayes and Support Vector Machine," *Ing. des Syst. d'Information*, vol. 29, no. 2, pp. 437–446, 2024, doi: 10.18280/isi.290205.

[40] L. N. CheSuh, R. Á. Fernández-Diaz, J. M. Alija-Perez, C. Benavides-Cuellar, and H. Alaiz-Moreton, "Improve quality of service for the Internet of Things using Blockchain & machine learning algorithms," *Internet of Things (Netherlands)*, vol. 26, 2024, doi: 10.1016/j.iot.2024.101123.

[41] A. Arsalane, A. Klilou, and N. El Barbri, "Performance evaluation of machine learning algorithms for meat freshness assessment," *Int. J. Electr. Comput. Eng.*, vol. 14, no. 5, pp. 5858–5865, 2024, doi: 10.11591/ijece.v14i5.pp5858-5865.

[42] W. Darmawan, "Komparasi Metode Klasifikasi Untuk Analisis Sentimen Pengguna Twitter Terhadap Penerapan Kurikulum Merdeka," *Ic-Tech*, vol. 18, no. 1, pp. 9–15, 2023, doi: 10.47775/ictech.v18i1.262.

[43] J. Maulani and M. Sari, "Komparasi Metode K-Nearest Neighbor (Knn) Dengan Support Vector Machine (Svm) Terhadap Tingkat Akurasi Klasifikasi Kualitas Air," *Smart Comp Jurnalnya Orang Pint. Komput.*, vol. 12, no. 2, 2023, doi: 10.30591/smartcomp.v12i2.4205.

[44] Y. Widhiyasana, T. Semiawan, I. G. A. Mudzakir, and M. R. Noor, "Penerapan Convolutional Long Short-Term Memory Untuk Klasifikasi Teks Berita Bahasa Indonesia," *J. Nas. Tek. Elektro Dan Teknol. Inf.*, vol. 10, no. 4, pp. 354–361, 2021, doi: 10.22146/jnteti.v10i4.2438.

[45] B. Molina-Coronado, U. Mori, A. Mendiburu, and J. Miguel-Alonso, "Survey of Network Intrusion Detection Methods From the Perspective of the Knowledge Discovery in Databases Process," *Ieee Trans. Netw. Serv. Manag.*, vol. 17, no. 4, pp. 2451–2479, 2020, doi: 10.1109/tnsm.2020.3016246.

[46] M. Defriani and I. Jaelani, "Recognition of Regional Traditional House in Indonesia Using Convolutional Neural Network (CNN) Method," *J. Comput. Networks Archit. High Perform. Comput.*, vol. 4, no. 2, pp. 104–115, 2022, doi: 10.47709/cnahpc.v4i2.1562.

[47] Z. Amri, "Prediksi Tingkat Kelulusan Mahasiswa Menggunakan Algoritma Naïve Bayes, Decision Tree, ANN, KNN, Dan SVM," *Edumatic J. Pendidik. Inform.*, vol. 7, no. 2, pp. 187–196, 2023, doi: 10.29408/edumatic.v7i2.18620.

[48] Y. Elda, S. Defit, Y. Yunus, and R. Syaljumairi, "Klasterisasi Penempatan Siswa Yang Optimal Untuk Meningkatkan Nilai Rata-Rata Kelas Menggunakan K-Means," *J. Inf. Dan Teknol.*, pp. 103–108, 2021, doi: 10.37034/jidt.v3i3.130.

[49] - Rezki, S. Defit, and S. Sumijan, "Metode K-Means Clustering Untuk Mengukur Tingkat Kedisiplinan Pegawai (Studi Kasus Di Pemerintah Kabupaten Padang Pariaman)," *J. Coscitech (Computer Sci. Inf. Technol.*, vol. 4, no. 1, pp. 116–125, 2023, doi: 10.37859/coscitech.v4i1.4728.

[50] M. D. Hendriyanto, A. A. Ridha, and U. Enri, "Analisis Sentimen Ulasan Aplikasi Mola Pada Google Play Store Menggunakan Algoritma Support Vector Machine," *Intecoms J. Inf. Technol. Comput. Sci.*, vol. 5, no. 1, pp. 1–7, 2022, doi: 10.31539/intecoms.v5i1.3708.

[51] E. Tohidi, "Analisa Sentimen Komentar Video Youtube Di Channel Tvonenews Tentang Calon Presiden Prabowo Subianto Menggunakan Support Vector Machine," *Jati (Jurnal Mhs. Tek. Inform.*, vol. 8, no. 1, pp. 660–667, 2024, doi: 10.36040/jati.v8i1.8560.

[52] J. R. N. A. Gunawardana, S. D. Viswakula, R. P. Rannan-Eliya, and N. Wijemunige, "Machine learning approaches for asthma disease prediction among adults in Sri Lanka," *Health Informatics J.*, vol. 30, no. 3, 2024, doi: 10.1177/14604582241283968.

[53] W. B. Demilie, "Plant disease detection and classification techniques: a comparative study of the performances," *J. Big Data*, vol. 11, no. 1, 2024, doi: 10.1186/s40537-023-00863-9.