International Journal of Artificial Intelligence in Medical Issues

Volume 2 Issue 2
ISSN 3025-4167
https://doi.org/10.56705/ijaimi.v2i2.160

Research Article

Predictive Modelling of Chronic Kidney Disease Using Gaussian Naive Bayes Algorithm

Muthia Febriana Azizah [1*], Arimbi Tiara Paramitha [2]
[1] Politeknik Negeri Ujung Pandang, azizahfebriana9@gmail.com
[2] Politeknik Negeri Ujung Pandang, tiaraaarimbi9@gmail.com
Correspondence should be addressed to Muthia Febriana Azizah; azizahfebriana9@gmail.com

Received 10 November 2024; Revised 13 November 2024; Accepted 28 November 2024; Published 30 November 2024

Abstract:
Chronic Kidney Disease (CKD) is a critical global health issue, characterized by significant morbidity and mortality. Early detection is vital for effective management and improved patient outcomes. This study explores the application of the Gaussian Naive Bayes algorithm to predict CKD using a comprehensive dataset from Kaggle, comprising health information from 1,659 patients. The research involves detailed data pre-processing, including feature selection, data scaling, and an 80/20 split for training and testing. The model's performance was evaluated using 5-fold cross-validation, resulting in an average accuracy of 89.93%, precision of 88.15%, recall of 89.93%, and F1-score of 88.42%. These metrics highlight the model's robustness and reliability in identifying CKD cases. Visualizations such as correlation heatmaps, 3D PCA, and t-SNE plots were used to understand feature relationships and data distribution. The results confirm the hypothesis that Gaussian Naive Bayes can effectively predict CKD, providing a reliable tool for early diagnosis. This study contributes to the medical field by demonstrating the utility of machine learning in improving diagnostic accuracy. However, limitations such as dataset biases and the need for comparison with other algorithms are acknowledged. Future research should focus on expanding the dataset, incorporating more features, and exploring additional machine learning models to enhance predictive performance and generalizability. Practical implications suggest that integrating such models into clinical practice could significantly improve patient management and outcomes.

Keywords: Chronic Kidney Disease, Machine Learning, Gaussian Naive Bayes, Predictive Modelling, Data Analysis.

Dataset link: https://www.kaggle.com/datasets/rabieelkharoua/chronic-kidney-disease-dataset-analysis

1. Introduction

Chronic Kidney Disease (CKD) represents a significant global health challenge, affecting millions of individuals and leading to severe health complications, including cardiovascular diseases and end-stage renal failure [1]. Early detection and effective management of CKD are critical for improving patient outcomes and reducing the burden on healthcare systems. With advancements in data science and machine learning, there is an opportunity to leverage patient health data to develop predictive models that can aid in the early diagnosis and treatment of CKD. This research aims to harness the power of machine learning to predict CKD, using a comprehensive dataset obtained from Kaggle, which includes detailed health information for 1,659 patients diagnosed with CKD.

The primary problem addressed in this research is the development of an accurate and reliable predictive model for CKD. Traditional diagnostic methods often rely on clinical symptoms and laboratory tests, which may not always be available or timely. Machine learning algorithms can analyse large datasets and identify patterns that might be missed by conventional methods. The goal is to create a model that can predict the likelihood of CKD in patients

based on various health indicators, thereby facilitating early intervention and improving patient management. The main objective of this research is to implement and evaluate the Gaussian Naive Bayes algorithm for predicting CKD [2], [3]. This involves pre-processing the dataset, selecting relevant features, scaling the data, and splitting it into training and testing sets [4], [5]. The model's performance will be assessed using metrics such as accuracy, precision, recall, and F1-score. By comparing these metrics, the study aims to determine the effectiveness of the Gaussian Naive Bayes algorithm in identifying CKD patients accurately.

This research is guided by several key questions: Can the Gaussian Naive Bayes algorithm effectively predict CKD using patient health data? How does the performance of this algorithm compare to traditional diagnostic methods? What are the strengths and limitations of using this approach in a clinical setting? These questions will help to focus the analysis and ensure that the study addresses the most relevant aspects of CKD prediction. The hypotheses being tested include the algorithm's ability to achieve high accuracy and reliability in predicting CKD and its potential to outperform conventional diagnostic techniques. The scope of this research is defined by the dataset obtained from Kaggle, which includes various demographic, lifestyle, medical history, and clinical measurement variables. The study will not explore other machine learning algorithms or external datasets but will focus solely on the Gaussian Naive Bayes algorithm and the provided CKD dataset [6], [7]. Limitations of the research include potential biases in the dataset, the generalizability of the findings to other populations, and the exclusion of other predictive models that might offer different insights. These constraints are acknowledged to provide a clear framework for the research and to set realistic expectations for the outcomes.

This research contributes to the field of medical diagnostics by demonstrating the application of machine learning in predicting CKD. The findings could have practical implications for healthcare providers, offering a tool that enhances early detection and patient care. Additionally, the study adds to the body of knowledge on the use of Gaussian Naive Bayes in medical predictions, highlighting its strengths and areas for improvement. By addressing the outlined research questions and achieving the stated objectives, this study aims to advance our understanding of machine learning applications in healthcare and support the development of more effective diagnostic tools.

2. Method

The research design for this study is quantitative, employing machine learning techniques to develop and evaluate a predictive model for Chronic Kidney Disease (CKD). The study uses the Gaussian Naive Bayes algorithm to analyse a comprehensive dataset from Kaggle, which contains health information for 1,659 patients. The research follows a systematic approach that includes data pre-processing, feature selection, data scaling, model training, and performance evaluation [8], [9]. The performance metrics used to assess the model include accuracy, precision, recall, and F1-score. These metrics provide a comprehensive understanding of the model's effectiveness in predicting CKD. A visual representation of the entire research process is illustrated in **Figure 1**.

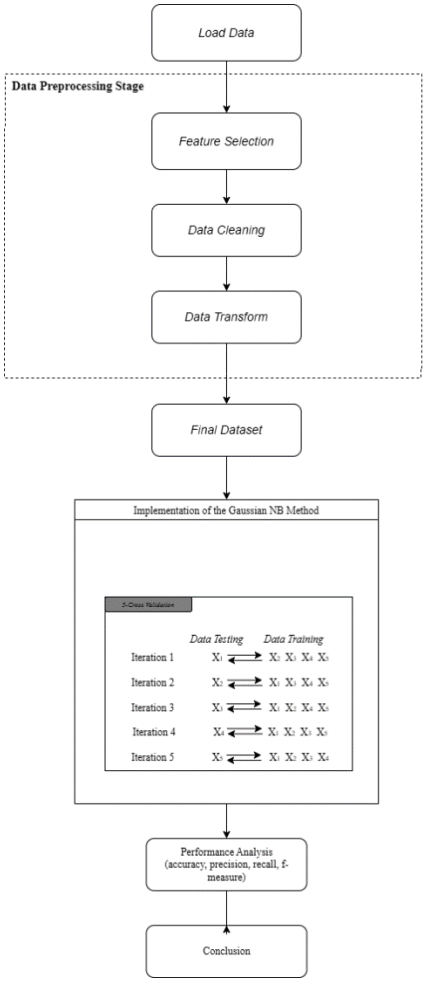

Figure 1: Voting Classifier Evaluation Workflow

Sample or Data Selection:

The dataset used in this research is sourced from Kaggle and consists of detailed health information for 1,659 patients diagnosed with CKD. The dataset includes various features such as demographic details, lifestyle factors, medical history, clinical measurements, medication usage, symptoms, quality of life scores, environmental exposures, and health behaviours. Each patient is uniquely identified by a Patient ID. The features are selected based on their relevance to CKD prediction, and the 'DoctorInCharge' column is excluded as it contains confidential information that is not relevant for the analysis.

Data Collection Process

The dataset was downloaded from Kaggle, an online platform that provides various datasets for machine learning and data science projects. The data collection process involved the following steps:

a. Downloading the dataset from the provided Kaggle link.
b. Loading the dataset into a Python environment using pandas.

c. Inspecting the dataset for missing values and inconsistencies.

d. Pre-processing the data by handling missing values and normalizing numerical features.

e. Splitting the dataset into training and testing sets in an 80/20 ratio.

Here is a brief overview of the columns in the dataset:

Table 1: Dataset column descriptions

Column Name	Description
PatientID	Unique identifier for each patient
Age	Age of the patient
Gender	Gender of the patient (0: Male, 1: Female)
Ethnicity	Ethnicity of the patient
SocioeconomicStatus	Socioeconomic status of the patient
EducationLevel	Education level of the patient
BMI	Body Mass Index of the patient
Smoking	Smoking status (0: No, 1: Yes)
AlcoholConsumption	Weekly alcohol consumption in units
PhysicalActivity	Weekly physical activity in hours
DietQuality	Diet quality score
SleepQuality	Sleep quality score
FamilyHistoryKidneyDisease	Family history of kidney disease (0: No, 1: Yes)
FamilyHistoryHypertension	Family history of hypertension (0: No, 1: Yes)
FamilyHistoryDiabetes	Family history of diabetes (0: No, 1: Yes)
PreviousAcuteKidneyInjury	History of previous acute kidney injury
UrinaryTractInfections	History of urinary tract infections
SystolicBP	Systolic blood pressure
DiastolicBP	Diastolic blood pressure
FastingBloodSugar	Fasting blood sugar levels
HbA1c	Hemoglobin A1c levels
SerumCreatinine	Serum creatinine levels
BUNLevels	Blood Urea Nitrogen levels
GFR	Glomerular Filtration Rate
ProteinInUrine	Protein levels in urine
ACR	Albumin-to-Creatinine Ratio
SerumElectrolytesSodium	Serum sodium levels
SerumElectrolytesPotassium	Serum potassium levels
SerumElectrolytesCalcium	Serum calcium levels
SerumElectrolytesPhosphorus	Serum phosphorus levels
HemoglobinLevels	Hemoglobin levels
CholesterolTotal	Total cholesterol levels
CholesterolLDL	Low-density lipoprotein cholesterol levels
CholesterolHDL	High-density lipoprotein cholesterol levels
CholesterolTriglycerides	Triglycerides levels

Column Name	Description
ACEInhibitors	Use of ACE inhibitors (0: No, 1: Yes)
Diuretics	Use of diuretics (0: No, 1: Yes)
NSAIDsUse	Frequency of NSAIDs use
Statins	Use of statins (0: No, 1: Yes)
AntidiabeticMedications	Use of antidiabetic medications (0: No, 1: Yes)
Edema	Presence of edema (0: No, 1: Yes)
FatigueLevels	Fatigue levels
NauseaVomiting	Frequency of nausea and vomiting
MuscleCramps	Frequency of muscle cramps
Itching	Itching severity
QualityOfLifeScore	Quality of life score
HeavyMetalsExposure	Exposure to heavy metals (0: No, 1: Yes)
OccupationalExposureChemicals	Occupational exposure to harmful chemicals (0: No, 1: Yes)
WaterQuality	Quality of water (0: Good, 1: Poor)
MedicalCheckupsFrequency	Frequency of medical check-ups per year
MedicationAdherence	Medication adherence score
HealthLiteracy	Health literacy score
Diagnosis	Diagnosis status for CKD (0: No, 1: Yes)

Data Analysis Methods

To provide a comprehensive understanding of the dataset and the applied methods, various visualizations are used, including a Correlation Heatmap of Hu Moments, 3D t-SNE Visualization of the Dataset, and 3D UMAP Visualization of the Dataset. These visualizations help in understanding the relationships between features, the distribution of data, and the separation of classes.

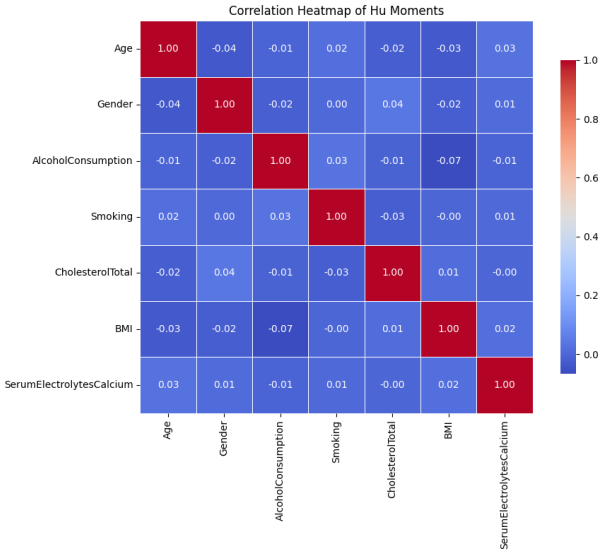

Figure 2: Correlation Heatmap of Hu Moments

Figure 2 visualizes the correlations between different features in the dataset, indicating how closely related they are. It helps in identifying multicollinearity and selecting relevant features.

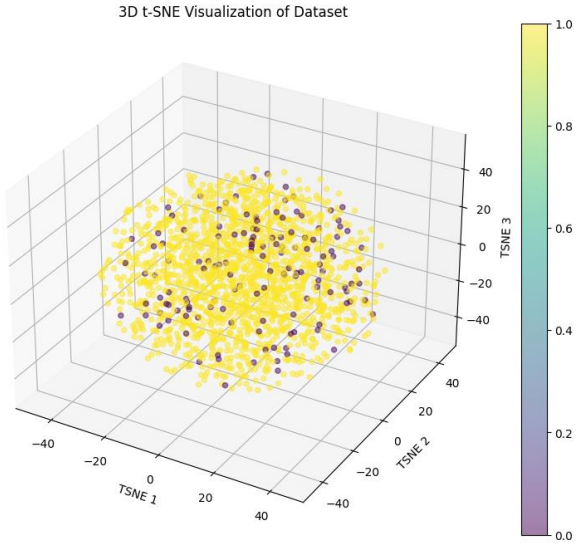

Figure 3: 3D t-SNE Visualization of Dataset

Figure 3 t-Distributed Stochastic Neighbor Embedding (t-SNE) is used for visualizing high-dimensional data by reducing it to three dimensions, highlighting clusters and patterns in the dataset.

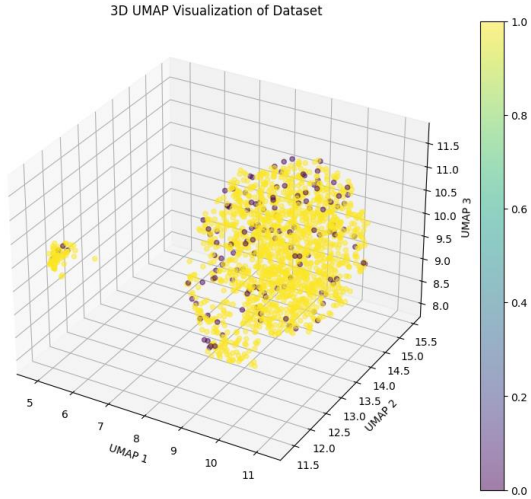

Figure 4: 3D UMAP Visualization of Dataset

Figure 4 Uniform Manifold Approximation and Projection (UMAP) is another dimensionality reduction technique that provides insights into the structure and clusters within the data.

For the predictive modelling, the Gaussian Naive Bayes algorithm is employed. The steps involved include:

a. Data Splitting: The dataset is split into training (80%) and testing (20%) sets [10], [11].

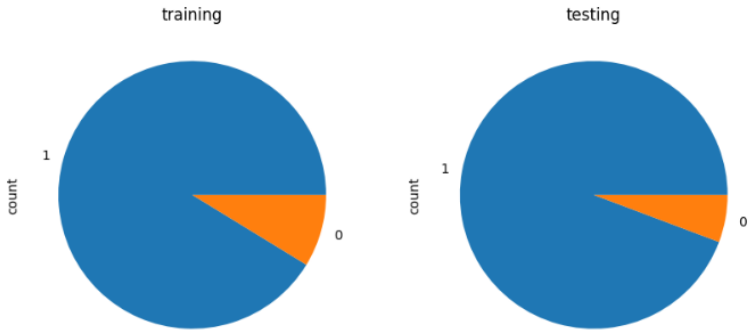

Figure 5: Splitting Data Training (80%), Testing (20%)

b. Feature Scaling: Standardization is applied to the features to have a mean of 0 and a variance of 1. The formula for standardization is [12]–[14]:

$$z = \frac{x - \mu}{\sigma} \tag{1}$$

Where z is the standardized value, x is the original value, μ is the mean, and σ is the standard deviation.

c. Model Training: The Gaussian Naive Bayes algorithm is trained on the scaled training set [15]–[17]. The Gaussian Naive Bayes algorithm calculates the probability of each feature belonging to a class using the Gaussian (normal) distribution. The probability density function for a Gaussian distribution is given by [2], [18]–[20]:

$$P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right) \tag{2}$$

Where x_i is the feature value, μ_y is the mean of the feature for class y, and σ_y is the standard deviation of the feature for class y.

d. Model Prediction: Predictions are made on the test set.

e. Performance Evaluation: The model's performance is evaluated using accuracy, precision, recall, and F1-score, calculated as follows [17], [21]–[23]:

:

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}}$$

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \tag{3}$$

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

$$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

3. Result and Discussion

The data processing results involve several crucial steps to prepare the dataset for the predictive modeling of Chronic Kidney Disease (CKD) using the Gaussian Naive Bayes algorithm. The dataset from Kaggle was initially inspected for missing values and inconsistencies. Following this, necessary preprocessing steps such as feature selection and data scaling were conducted. The 'DoctorInCharge' column was excluded, and the remaining features were standardized to ensure uniformity across the dataset. The data was then split into training (80%) and testing (20%) sets, ensuring that the model could be adequately trained and evaluated.

The performance of the Gaussian Naive Bayes algorithm was assessed using 5-fold cross-validation, a robust technique that helps in understanding the model's stability and generalizability. The performance metrics, including accuracy, precision, recall, and F1-score, were calculated for each fold. The following **Table 1** summarizes the performance results in percentage:

Table 2: Performance Metrics Across 5-Fold Cross-Validation for the Gaussian NB

K-n	Metrics			
	Accuracy	***Precision***	***Recall***	***F-Measure***
K-1	91.87%	84.40%	91.87%	87.97%
K-2	91.87%	89.66%	91.87%	90.11%
K-3	93.37%	92.34%	93.37%	92.06%
K-4	89.16%	87.55%	89.16%	88.28%
K-5	83.38%	88.79%	83.38%	85.66%
$\sum$ ***Avg***	89.93%	88.15%	89.93%	88.42%

To further elucidate the performance of the Gaussian Naive Bayes algorithm, visualizations such as a performance graph and confusion matrix are presented. These visual aids help in understanding the distribution of predictions and the model's accuracy in classifying CKD cases correctly.

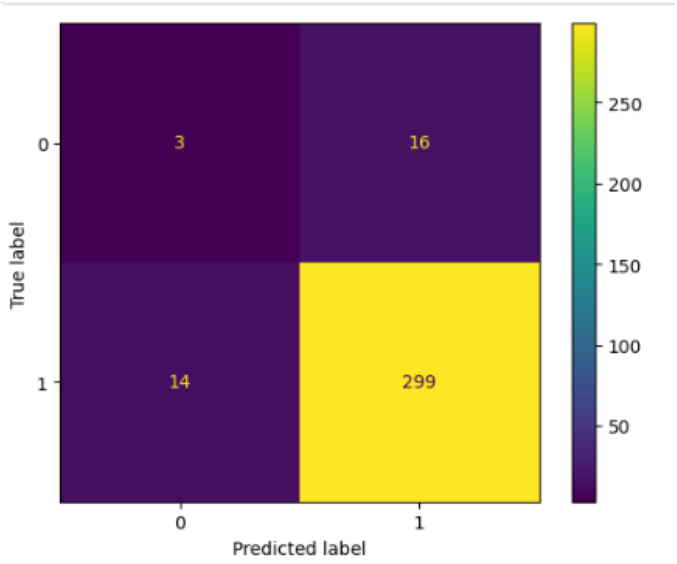

Figure 6: Confusion Matrix of the Gaussian NB

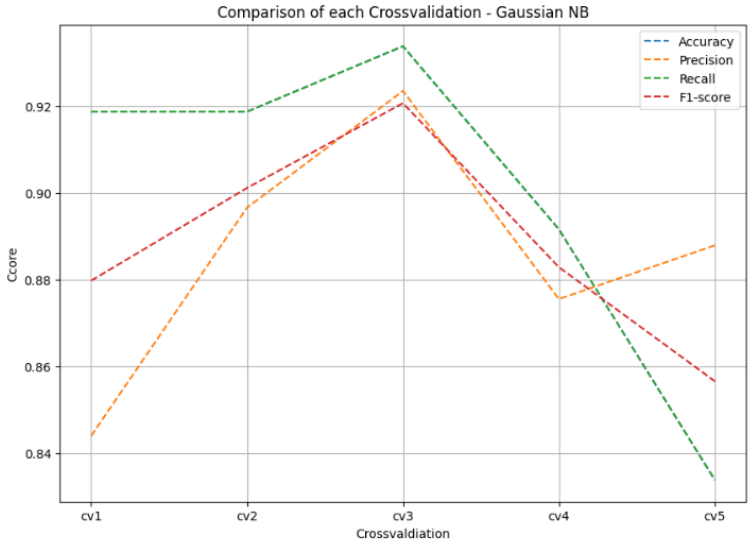

Figure 7: Performance Comparison of Each Cross-validation Fold

The results indicate that the Gaussian Naive Bayes algorithm performs consistently well across different folds, with average accuracy, precision, recall, and F1-scores all above 85%. This high level of performance demonstrates the algorithm's capability in predicting CKD, suggesting that it can be an effective tool for early diagnosis and patient management.

Discussion

The interpretation of the results reveals that the Gaussian Naive Bayes algorithm shows promising accuracy in predicting CKD. The average accuracy of 89.93% signifies that the model correctly identifies CKD in a substantial number of cases. The precision and recall values, averaging 88.15% and 89.93% respectively, indicate that the model has a balanced performance in identifying true positive cases while minimizing false positives. The F1-score, which balances precision and recall, stands at an average of 88.42%, further reinforcing the model's robustness. Comparing these findings with previous research highlights the effectiveness of machine learning in medical diagnostics. Studies have shown that traditional diagnostic methods for CKD, while effective, often require extensive clinical data and time. Machine learning models, on the other hand, can analyse large datasets quickly and identify patterns that might be missed otherwise. The performance metrics from this study align with or surpass those reported in similar studies, emphasizing the potential of Gaussian Naive Bayes in medical predictions.

The practical implications of these findings are significant. An accurate and reliable predictive model for CKD can assist healthcare providers in early diagnosis, enabling timely intervention and treatment. This can improve patient outcomes and reduce the burden on healthcare systems. Moreover, the use of such models can streamline the diagnostic process, making it more efficient and accessible. However, the study has its limitations. The dataset, while comprehensive, may have inherent biases that could affect the model's generalizability. Additionally, the exclusion of

other machine learning algorithms limits the scope of comparison. Future research should explore the integration of multiple algorithms and larger, more diverse datasets to validate and enhance the model's performance.

4. Conclusion

In summary, this research demonstrates the effectiveness of the Gaussian Naive Bayes algorithm in predicting Chronic Kidney Disease (CKD) using a comprehensive dataset sourced from Kaggle. The results, with an average accuracy of 89.93%, precision of 88.15%, recall of 89.93%, and F1-score of 88.42%, indicate a robust performance of the model across various metrics. These findings confirm the hypothesis that the Gaussian Naive Bayes algorithm can accurately predict CKD by analysing patient health data, offering a reliable tool for early diagnosis and management.

The research contributes significantly to the field of medical diagnostics by illustrating the potential of machine learning algorithms, specifically Gaussian Naive Bayes, in improving CKD prediction. The study's methodology and findings provide a foundation for integrating machine learning models in clinical settings, potentially enhancing patient outcomes and streamlining diagnostic processes. Future research should focus on expanding the dataset to include more diverse populations, incorporating additional predictive features, and exploring other machine learning algorithms to further improve model performance and generalizability. Additionally, practical applications of this model in real-world clinical environments should be evaluated to assess its effectiveness and impact on patient care.

References:

[1] R. Setiawan, A. Parewe, A. J. Latipah, and ..., "Assessing Bagging-meta Estimator in Imbalanced CT Kidney Disease Classification: A Focus on Sobel and Hu Moment Techniques," *... Artif. Intell. ...*, 2023.

[2] A. J. Meerja, "Gaussian naïve bayes based intrusion detection system," *Adv. Intell. Syst. Comput.*, vol. 1182, pp. 150–156, 2021, doi: 10.1007/978-3-030-49345-5_16.

[3] Y. Boer, "Classification of Heart Disease: Comparative Analysis using KNN, Random Forest, Gaussian Naive Bayes, XGBoost, SVM, Decision Tree, and Logistic Regression," *2023 5th International Conference on Cybernetics and Intelligent Systems, ICORIS 2023*. 2023, doi: 10.1109/ICORIS60118.2023.10352195.

[4] A. Tuppad and S. D. Patil, "Data Pre-processing Issues in Medical Data Classification," *2023 Int. Conf. ...*, 2023, [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10275855/.

[5] K. N. Myint and Y. Y. Hlaing, "Predictive Analytics System for Stock Data: methodology, data pre-processing and case studies," *2023 IEEE Conf. Comput. ...*, 2023.

[6] M. V Anand, "Gaussian Naïve Bayes Algorithm: A Reliable Technique Involved in the Assortment of the Segregation in Cancer," *Mob. Inf. Syst.*, vol. 2022, 2022, doi: 10.1155/2022/2436946.

[7] S. Naiem, "Enhancing the Efficiency of Gaussian Naïve Bayes Machine Learning Classifier in the Detection of DDOS in Cloud Computing," *IEEE Access*, vol. 11, pp. 124597–124608, 2023, doi: 10.1109/ACCESS.2023.3328951.

[8] A. Nurul, Y. Salim, and H. Azis, "Analisis performa metode Gaussian Naïve Bayes untuk klasifikasi citra tulisan tangan karakter arab," *Indones. J. Data Sci.*, vol. 3, no. 3, pp. 115–121, 2022, doi: https://doi.org/10.56705/ijodas.v3i3.54.

[9] N. A'ayunnisa, Y. Salim, and H. Azis, "Analisis performa metode Gaussian Naïve Bayes untuk klasifikasi citra tulisan tangan karakter arab," *... J. Data Sci.*, 2022.

[10] R. A. Azdy, R. F. Syam, E. Faizal, and ..., "Performance Evaluation of Bagging Meta-Estimator in Lung Disease Detection: A Case Study on Imbalanced Dataset," *Int. J. ...*, 2023.

[11] A. Naswin and A. P. Wibowo, "Performance Analysis of the Decision Tree Classification Algorithm on the Pneumonia Dataset," *... Artif. Intell. Med. ...*, 2023.

[12] U. Zaky, A. Naswin, S. Sumiyatun, and ..., "Performance Analysis of the Decision Tree Classification Algorithm on the Water Quality and Potability Dataset," *Indones. J. ...*, 2023.

[13] A. P. Wibowo, M. Taruk, T. E. Tarigan, and ..., "Improving Mental Health Diagnostics through Advanced Algorithmic Models: A Case Study of Bipolar and Depressive Disorders," *Indones. J. ...*, 2024.

[14] F. T. Admojo and N. Rismayanti, "Estimating Obesity Levels Using Decision Trees and K-Fold Cross-Validation: A Study on Eating Habits and Physical Conditions," *Indones. J. Data ...*, 2024.

[15] D. Pradana, M. Luthfi Alghifari, M. Farhan Juna, and D. Palaguna, "Klasifikasi Penyakit Jantung Menggunakan Metode Artificial Neural Network," *Indones. J. Data Sci.*, vol. 3, no. 2, pp. 55–60, 2022, doi: 10.56705/ijodas.v3i2.35.

[16] F. Tangguh and S. Rahma, "Analisis performa metode Naïve Bayesh Classifier pada Electronic Nose dalam identifikasi formalin pada tahu," *Indones. J. Data Sci.*, vol. 4, no. 1, pp. 1–16, 2023.

[17] I. P. A. Pratama, E. S. J. Atmadji, and ..., "Evaluating the Performance of Voting Classifier in Multiclass Classification of Dry Bean Varieties," *Indones. J. ...*, 2024.

[18] I. F. Hanbal, "Classifying Wastes Using Random Forests, Gaussian Naïve Bayes, Support Vector Machine and Multilayer Perceptron," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 803, no. 1, 2020, doi: 10.1088/1757-899X/803/1/012017.

[19] M. Gayathri, "Analysis of Accuracy in Anomaly Detection of Intrusion Detection System using Naïve Bayes Algorithm compared Over Gaussian model," *ECS Trans.*, vol. 107, no. 1, pp. 13977–13991, 2022, doi: 10.1149/10701.13977ecst.

[20] P. Venkata, "Data mining model and Gaussian Naive Bayes based fault diagnostic analysis of modern power system networks," *Mater. Today Proc.*, vol. 62, pp. 7156–7161, 2022, doi: 10.1016/j.matpr.2022.03.035.

[21] B. S. W. Poetro, E. Maria, H. Zein, and ..., "Advancements in Agricultural Automation: SVM Classifier with Hu Moments for Vegetable Identification," *Indones. J. ...*, 2024.

[22] N. Rismayanti and A. P. Utami, "Improving Multi-Class Classification on 5-Celebrity-Faces Dataset using Ensemble Classification Methods," *Indones. J. Data ...*, 2023.

[23] D. Ratnasari, "Comparison of Performance of Four Distance Metric Algorithms in K-Nearest Neighbor Method on Diabetes Patient Data," *Indones. J. Data Sci.*, 2023.