JURNAL ILMIAH FIFO
Vol. 17 No. 2 (2025) 102-118
DOI: 10.22441/fifo.2025.v17i2.001

P-ISSN 2085-4315
E-ISSN 2502-8332

A Comparative Study of Machine Learning with Statistical Feature Selection for Risk Detection of Diabetic

Isnen Hadi Al Ghozali[1*], M. Askar Fathin[2], Andy Rio Handoko[3]

[1,3]Fakultas Teknologi Informasi, Universitas Budi Luhur
[2]Sekolah Vokasi, Universitas Gadjah Mada
E-mail: [1]2111601163@student.budiluhur.ac.id, [2]muhammadaskar@mail.ugm.ac.id,
[3]andy.handoko@budiluhur.ac.id
Corresponding Author[*]

(***received*****:** 24-06-24, ***revised*****:** 08-09-24, ***accepted*****:** 20-08-25)

Abstract

Elevated glucose levels in the circulation are indicative of diabetes, a chronic medical condition. Prolonged unregulated blood glucose levels pose a significant risk of severe consequences, including renal failure, myocardial infarction, and lower limb amputation. The objective of this study is to conduct a comparative analysis of SVM, Naive Bayes, XGBoost, Random Forest, and ANN models in order to forecast the occurrence of diabetes. The research methodology comprises seven primary stages: (1) literature review, (2) data collection, (3) exploratory data analysis (EDA), (4) data preprocessing, (5) feature selection, (6) model development, and (7) model evaluation and comparison. The XGBoost model is the most suitable option, as indicated by the model evaluation results. The XGBoost model achieved a precision of 0.88, a recall of 0.87, and an accuracy of 0.8690. The XGBoost model has a RMSE of 0.3620 and a MSE of 0.1310.

Keyword*: SVM, XGBoost, Random Forest, Naive Bayes, ANN*

Abstrak

Peningkatan kadar glukosa dalam sirkulasi merupakan indikasi diabetes, kondisi medis kronis. Tidak terkontrolnya kadar gula darah dalam jangka waktu lama berisiko menimbulkan komplikasi berbahaya, seperti gagal ginjal, serangan jantung, dan amputasi kaki. Tujuan penelitian ini adalah mengkomparasikan model SVM, Naive Bayes, XGBoost, Random Forest, dan ANN untuk memprediksi diabetes. Metode penelitian terdiri dari tujuh tahap: (1) melakukan tinjauan literatur, (2) mengumpulkan data, (3) melakukan analisis data eksplorasi, (4) pra-pemprosesan data, (5) pemilihan fitur, (6) mengembangkan model, dan (7) evaluasi dan komparasi model. Berdasarkan hasil evaluasi model, model XGBoost merupakan pilihan paling optimal. Model XGBoost mencapai akurasi 0,8690, presisi 0,88, dan recall 0,87. Model XGBoost menunjukkan RMSE sebesar 0,3620 dan MSE sebesar 0,1310.

Kata Kunci: *SVM, XGBoost, Random Forest, Naive Bayes, ANN*

1. INTRODUCTION

High blood glucose, often known as blood sugar, is a defining feature of diabetes, a chronic metabolic disease that can seriously harm the heart, blood vessels, kidneys, eyes, nerves, and heart. [1] Individuals with diabetes are more susceptible to experiencing symptoms such as cardiovascular disease, visual impairment, lower extremity amputation, and renal illness as a result of persistently elevated levels of glucose in their body. Although there is no definitive remedy for diabetes, many interventions such as shedding excess weight, adopting a nutritious diet, engaging in regular physical exercise, and receiving medical treatment can significantly mitigate the adverse consequences of the disease for numerous people. As to the World Health Organization (WHO), the global diabetes population rose to 422 million in 2014 and is projected to surge to 700 million by 2045. [2] The Behavioral Risk Factor Surveillance System (BRFSS) is an annual health-related telephone survey conducted by the Center for Disease Control (CDC). Annually, the study gathers information from more than 400,000 individuals in the United States about habits that pose health risks, persistent medical conditions, and the utilization of preventative therapies. [3]

The prediction process is significantly influenced by machine learning (ML) and deep learning (DL) algorithms, including reinforcement learning, unsupervised learning, and supervised learning. Predictive algorithms are advanced mathematical models that use past data to forecast future results. [4] These algorithms generate precise predictions by evaluating data patterns, correlations, and trends. Classification algorithms are

components of supervised learning. Machine learning is a widely used tool for making predictions with classification algorithms. Some machine learning algorithms are pure, including Support Vector Machine (SVM) [5], [6], Naive Bayes [7], and Artificial Neural Network (ANN) [8], [9]. In its development, there is ML, which involves an improvement of basic algorithms such as Random Forest, the result of Decision Trees [10], [11], and XGBoost, the outcome of Gradient Tree Boosting. [12], [13]

ML, or machine learning, is widely utilized in the medical domain. Utilizing Support Vector Machines (SVM) in health research yields a significant level of precision. The study aimed to forecast the occurrence of diabetes [1], with an accuracy rate of just 81.3%. For the prediction of prostate cancer, the support vector machine (SVM) algorithm obtained the maximum accuracy of 94.55%. [14] Concurrently, the research [15] yielded a mean accuracy of 97.36%. The highest level of accuracy achieved in clinical diagnosis studies was 97.7%. The research [6] yielded the most optimal outcomes, attaining a precision rate of 98.3%. ANN and XGBoost are commonly utilized in the medical area as alternative classification methods. Research conducted on the categorization of sleep disorders using Artificial Neural Networks (ANN) resulted in an accuracy rate of 92.92%. [16] A study [17] utilized an optimized Artificial Neural Network (ANN) technique to create a model capable of accurately detecting COVID-19. The model achieved an accuracy range of 89.83–92.37 percent. However, the ANN system only obtained an accuracy of 83.6% in detecting diabetes. [1] The XGBoost algorithm has significant variability in its accuracy. In the study [13] used XGBoost was employed to optimize the CNN algorithm, resulting in an accuracy of 87.07% in the classification of colorectal polyp sizes. For clinical diagnosis, XGBoost achieves an accuracy range of 82.25–97.8%. [7] XGBoost demonstrated a predictive accuracy range of 83.86–97.01 percent in forecasting depression levels. [18]Random Forest and Naive Bayes might be regarded as alternative classification algorithms. Random Forest demonstrates an accuracy range of 83.28–96% for predictions in the medical domain. [19] An accuracy of 99.37 ± 0.05% was achieved in the classification of valvular heart disease using the optimized Random Forest method. [20] The accuracy of diagnosing and detecting Alzheimer's disease using Random Forest is 86.93%. [12] A Random Forest model may achieve a detection accuracy of 95.41% for hepatitis C virus. [21] Random Forest can accurately predict diabetes even when dealing with class imbalances and missing values, with an accuracy rate of 88.03%. [22] Naive Bayes is less prevalent in the medical domain compared to the previously mentioned algorithms. The Naive Bayes algorithm demonstrated an accuracy of 92.39% in healthcare detection. [4] Nevertheless, when it comes to detecting diabetes, Naive Bayes algorithm only attained an accuracy rate of 82.4%. On imbalanced data, Naive Bayes achieved accuracy ranging from 71.88% to 96.44%. [23].

In the study we analyzed, the emphasis on feature selection was still lacking. The goal of feature selection is to identify the highest possible amount of relevant characteristics while reducing the costs associated with data processing. [24] Multiple studies, specifically studies [25], [26], [27], and [28], have employed feature selection in their process of constructing models. Typically, these studies provide prediction models that are applicable to a wide range of situations. However, research specifically targeting the prediction of diabetes risk has not yet taken into account the process of selecting relevant features. Furthermore, the problem of data imbalance has not been thoroughly investigated. Only two articles, specifically studies [22] and [28], addressed the issue of dealing with imbalanced data throughout the process of constructing a model. Addressing the problem of data imbalance is essential to avoid overfitting in the model being developed. Hence, this study will employ feature selection techniques to construct models and address imbalanced data, hence mitigating the risk of overfitting in the final model.

2. METHODOLOGY

A. Data Sources and Research Variables

The indicators of the Behavioral Risk Factor Surveillance System (BRFSS) for 2015 were employed in this investigation. A pristine dataset of 253,680 survey responses collected for the CDC's BRFSS 2015 was employed in this study. The target variable "Diabetes_binary" is composed of two distinct classes. The absence of diabetes is denoted by the value 0, while the presence of prediabetes or diabetes is denoted by the value 1. The dataset is imbalanced and contains 21 feature variables. Twenty-one feature variables are included in the dataset.

B. Research Framework

This study use a comparison methodology to determine the optimal model. Comparative studies seek to analyze and evaluate two or more variables in order to ascertain the presence of a comparison between the items being studied. Figure 3 depicts the seven processes that were carried out in the study to arrive at its findings. The research methodology comprises seven primary stages: (1) literature review, (2) data collection, (3) exploratory

P-ISSN 2085-4315
E-ISSN 2502-8332

data analysis (EDA), (4) data preprocessing, (5) feature selection, (6) model development, and (7) model evaluation and comparison. Table 1 presents a comprehensive analysis of the existing literature in a methodical manner. The findings of a comprehensive analysis of existing literature reveal that there are five alternative approaches for constructing diabetes risk prediction models, including SVM, Naive Bayes, XGBoost, Random Forest, and ANN.

Table 1. Systematic Literature Review

No.	Authors	Dataset	Methodology	Evaluation Metrics	Research Conclusions
1	H. Zhang, Y. Zhang [5]	Coal and gangue datasets	Enhanced Sparrow Search Algorithm for optimizing SVM parameters	Accuracy, Precision, Recall	The improved Sparrow Algorithm provides better SVM parameter optimization results compared to traditional methods. Accuracy for ISSA-SVM is 94.27%. Base model SVM reach 80.10% of accuracy.
2	S. Liang, A. Q. M. Sabri, F. Alnajjar, C. K. Loo [6]	Self-collected video datasets of autistic behavior	Deep Learning for temporal feature extraction followed by SVM classification	Accuracy, Interpretability	Combining temporal features with SVM yields accurate and explainable classification of autistic behaviors. TCDN-SVM achieves accuracy of 98.3%
3	D. N, N. P. K S [7]	Public health records datasets	Combining predictive techniques and analytics for improved clinical diagnosis	Accuracy, Sensitivity, Specificity	Predictive analytics overall enhances the accuracy of clinical diagnoses. Proposed model reach accuracy range of 99.26 - 100%. NAive Bayes model achieved accuracy range of 78.92 - 95.17%
4	C. Liu, Z. Gu, J. Wang [10]	KDD Cup 1999, NSL-KDD	Hybrid intrusion detection system combining K-Means, Random Forest, and Deep Learning	Accuracy, F1-Score, Processing Time	The hybrid system is more effective in detecting intrusions compared to conventional methods. NSL-KDD has a detection rate of 98.57%, while the CIC-IDS2017 dataset has a detection rate of 99.67%.
5	Z. Huang, D. Chen [11]	Wisconsin Breast Cancer Dataset	VIM feature selection and hierarchical Random Forest algorithm for breast cancer diagnosis	Accuracy, Precision, Recall	This method significantly improves breast cancer diagnostic performance. The most accurate method is the HCRF algorithm with VIM as the feature selection method, which gets 97.05% and 97.76%.
6	G. P. Shukla, S. Kumar, S. K. Pandey, R. Agarwal, N. Varshney, A. Kumar	Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset	Learning algorithms for the detection and diagnosis of Alzheimer's disease based on imaging and clinical data	Accuracy, F1-Score, Sensitivity	Learning algorithms are effective in detecting and diagnosing Alzheimer's disease. Proposed model (RF, XGBoost, and CNN) achieves accuracy of 97.52%.
7	P. Haldar et al.	Publicly available colorectal polyp dataset	Binary CNN combined with XGBoost for	Accuracy, Precision, Recall	The combination of XGBoost with Binary CNN improves the classification performance for

P-ISSN 2085-4315
E-ISSN 2502-8332

			colorectal polyp size classification		colorectal polyp size. Proposed model (CNN with XGBoost) reaches an accuracy of 87.07%.
8	M. Varan, J. Azimjonov, B. Maçal	Prostate cancer imaging dataset	Utilizing key radiomics features and a fine-tuned Linear SVM for prostate cancer classification	Accuracy, Precision, Recall	Enhanced radiomics features with Linear SVM yield more accurate prostate cancer classification results. Proposed model (Linear SVM with IC feature) achieves accuracy of 90.91%.
9	H. M. Alshamlan	UCI Machine Learning Repository	Filter method to improve the performance of FF-SVM (Fuzzy Support Vector Machine)	Accuracy, Precision, Recall	The filter method effectively enhances the performance of the FF-SVM algorithm in data classification. FF-SVM achieved accuracy range of 92.7 - 100%
10	T. S. Alshammari	Sleep Disorder dataset	Various ML algorithms for classifying sleep disorders such as SVM, Random Forest, and k-Nearest Neighbors (k-NN)	Accuracy, F1-Score	ML algorithms effectively classify various sleep disorders based on the available data. ANN reaches an accuracy of 92.92%.
11	S. Punitha, T. Stephan, R. Kannan, M. Mahmud, M. S. Kaiser, S. B. Belhaouari	COVID-19 CT scan dataset	Using ANN optimized with Swarm Intelligence for detecting COVID-19 from lung CT images	Accuracy, Precision, Recall	Swarm Intelligence optimization improves ANN accuracy in detecting COVID-19 from CT images. Proposed model (RF, XGBoost, and CNN) achieves accuracy of 92.37%.
12	J.-G. Choi, I. Ko, S. Han	Self-collected actigraphy data	Various ML algorithms for classifying depression levels based on actigraphy data	Accuracy, F1-Score	ML algorithms are effective in classifying depression levels from actigraphy data. Every machine learning model exhibits varying degrees of accuracy. The accuracy achieved by XGBoost was 86.47%, MLP achieved 71.25% accuracy, logistic regression achieved 66.18% accuracy, and the lowest accuracy was achieved by SVC at 65.45%.
13	C. Zhang, X. Wang, S. Chen, H. Li, X. Wu, X. Zhang	UCI Machine Learning Repository	Random Forest modified using Kappa measure and Binary Artificial Bee Colony algorithm for feature selection	Accuracy, Precision, Recall	Modifications improve Random Forest performance in classification and feature selection. Proposed model (modified Random Forest) achieved accuracy range of 69.63 - 98.21%.
14	T. Sinha Roy, J. K. Roy, N. Mandal	IoT sensor data for heart diseases	Deep Learning model based on CNN combined	Accuracy, F1-Score,	Combining CNN with Random Forest improves the accuracy of valvular heart disease

P-ISSN 2085-4315
E-ISSN 2502-8332

			with Random Forest for classifying and analyzing valvular heart diseases	Precision, Recall	classification. The Conv-RF approach achieved an accuracy of 99.37 ± 0.05%.
15	T.-H. S. Li, H.-J. Chiu, P.-H. Kuo	Hepatitis C Virus dataset	Detection model combining Random Forest, Logistic Regression, and ABC (Artificial Bee Colony) algorithm for Hepatitis C virus detection	Accuracy, Precision, Recall	The combined model provides more accurate Hepatitis C virus detection. Proposed model reaches an accuracy of 98.84%.
16	L. Jia, Z. Wang, S. Lv, Z. Xu	Pima Indian Diabetes Dataset	Efficient probabilistic ensemble classification algorithm (PE_DIM) for handling class imbalance and missing values in diabetes prediction	Accuracy, Precision, Recall	PE_DIM improves accuracy and reliability of diabetes predictions by addressing imbalance and missing data issues. PE_DIM achieves accuracy of 94.53%.
17	Z. Ahmed, B. Issac, S. Das	UCI Machine Learning Repository	Combination of OPTICS (Ordering Points to Identify the Clustering Structure) and k-Naive Bayes for class imbalance classification with overlapping issues	Accuracy, Precision, Recall	This combination effectively addresses imbalance and overlapping classification problems. ok-NB achieved accuracy range of 71.88 - 96.44%.
18	H. C. S. C. Lima, F. E. B. Otero, L. H. C. Merschmann, M. J. F. Souza	UCI Machine Learning Repository	Hybrid feature selection algorithm for hierarchical classification	Accuracy, Precision, Recall	The hybrid algorithm improves hierarchical classification performance through better feature selection.
19	G. J. Ansari, J. H. Shah, M. C. Q. Farias, M. Sharif, N. Qadeer, H. U. Khan	ICDAR 2003 Scene Text dataset	Feature selection technique optimized using Genetic Algorithm for text classification in diverse natural scenes	Accuracy, Precision, Recall	The Genetic Algorithm enhances feature selection and classification accuracy for natural scene text. Proposed model achieved an accuracy range of 88.1 - 91.8%.

P-ISSN 2085-4315
E-ISSN 2502-8332

20	A. K. Mandal, Md. Nadim, H. Saha, T. Sultana, Md. D. Hossain, E.-N. Huh	High-dimensional, low sampling size datasets	Feature subset selection using ensemble methods with wrapper-based search for high-dimensional, low sample size data classification	Accuracy, Precision, Recall	This method is effective in selecting optimal feature subsets for high-dimensional, low sample size data. Proposed approach achieved an accuracy of 99.83 ± 0.73%.
21	L. Al-Shalabi	UCI Machine Learning Repository	A novel algorithm for feature selection that is predicated on the stability and correlation of features	Accuracy, Precision, Recall	This algorithm improves classification performance by selecting stable and highly correlated features. Proposed model achieved an accuracy range of 69.31 - 95.59%.
22	F. Feng, K.-C. Li, J. Shen, Q. Zhou, X. Yang	UCI Machine Learning Repository	Cost-sensitive learning and feature selection algorithms for imbalanced classification	Accuracy, Precision, Recall	This approach improves classification performance under class imbalance conditions. Proposed model achieves accuracy of 96%.

Support Vector Machine, or SVM, is a widely used supervised machine learning algorithm that is employed in a variety of applications for both classification and regression tasks. The primary objective of the SVM technique is to determine the hyperplane that can accurately separate the data belonging to different classes. [29] A hyperplane functions as a discriminant for different classes. In the context of Support Vector Machines (SVM), a support vector is defined as the data object that is located closest to the hyperplane. Support vectors, due to their positioning that closely overlaps with other classes, pose the greatest challenge in classification. [30] The XGBoost algorithm is a very efficient tree boosting algorithm that is known for its scalability, sparsity-awareness, ability to handle data compression and sharding, and cache-aware access. Figure 1 depicts the XGboost algorithm model.

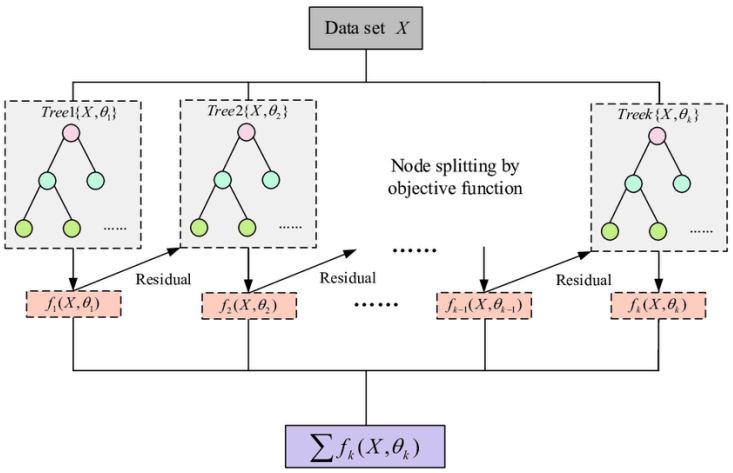


Figure 1 XGboost Algorithm Model [31]

The Random Forest technique, introduced by Breiman in 2001. This supervised learning algorithm runs based on a straightforward yet efficient divide-and-conquer principle: Initially, the data undergoes a process of sampling, and thereafter, a random tree predictor is generated for each fragment. Subsequently, forecasts can be derived from the average values produced by these predictors. [21] The final prediction in a binary task is determined by aggregating the ballots of all the trees. [11] The Naive Bayes (NB) algorithm operates under the

P-ISSN 2085-4315
E-ISSN 2502-8332

assumption that characteristics are mutually independent. This classifier is seen as an easy and practical way to use guided machine learning to sort things into groups. The idea behind it is that the info will be distributed normally. [32] The Bayes theorem is predicated on the assumption that each selected feature contributes equally as well is independent, in which case the likelihood and prior probabilities are multiplied and divided by the evidence. Naïve Bayes is a technique that employs the Bayes theorem, which is as follows:

$$\text{Option : P (A | B)} = \frac{P\,(B \mid A)\,P\,(A)}{P\,(B)} \quad (1)$$

with:
P(A | B) : The posterior probability refers to the conditional likelihood of event A given that event B has occurred.
P(B | A) : The conditional probability of occurrence B given that event A has already occurred.
P(A) : The likelihood of event A happening
P(B) : The likelihood of event B happening

McCulloch and Pitts established the fundamental principles of Artificial Neural Networks (ANN) in 1943. The neuron processed both the input data and the output data, which were of a boolean type [8]. Artificial neurons are specifically engineered to replicate the anatomical and operational characteristics of natural neurons, using mathematical descriptions of the nucleus, dendrites, synapses, and axons. Figure 32 depicts the scheme of the Artificial Neural Network (ANN) model.

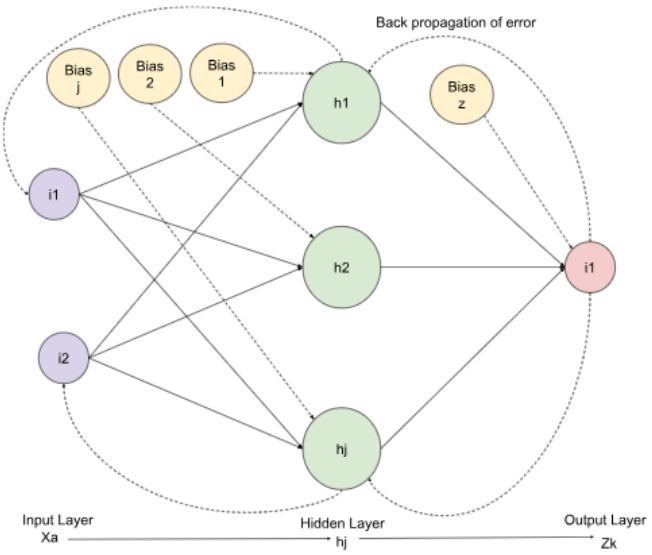


Figure 2 ANN Algorithm Model

During the subsequent phase, we gather a dataset consisting of lifestyle variables that have the potential to contribute to the development of diabetes. Specifically, we focus on the BRFSS indicators for the year 2015. We will analyze the gathered data using exploratory data analysis (EDA). The purpose of doing the exploratory data analysis (EDA) method is to discover genuine mistakes, get a deeper understanding of data patterns, identify outliers or anomalous occurrences, and uncover meaningful links between variables. Once we have comprehended the data, we proceed with preprocessing the necessary data in order to acquire sanitized data. Prior to constructing a model, we employ the variance inflation factor (VIF) and chi-square to carefully choose the features. VIF measures the degree to which the variance of regression estimates is amplified as a result of the existence of correlated independent variables. A higher correlation between the independent variables will result in an even higher VIF rating. Multicollinearity can be detected by seeing a Variance Inflation Factor (VIF) value that exceeds 10. [33] The formulation of VIF is as follows:

$$VIF = \frac{1}{1 - R^2} \quad (2)$$

JURNAL ILMIAH FIFO
Vol. 17 No. 2 (2025) 102-118
DOI: 10.22441/fifo.2025.v17i2.001
P-ISSN 2085-4315
E-ISSN 2502-8332

The R^2 number is used to assess the degree to which the independent variables effectively explain each other. The chi-square test is used as a statistical method to compare the observed results with the expected results. [34] The purpose of this test is to ascertain if a difference between the actual and predicted data is due to random chance or to a correlation between the variables under investigation. The Chi-Square test is determined by the subsequent formula:

$$x_c^2 = \frac{\Sigma(O_i - E_i)^2}{E_i} \quad (3)$$

where:
c : Number of independent variables or parameters that can vary in a statistical model.
O : Measured or recorded value
E : Expected Value

We will build prediction models employing the selected attributes based on the VIF and Chi-Square criteria. Prior to constructing a prediction model, our initial step will involve addressing the issue of imbalanced data. We implemented the NearMiss algorithm in our framework. In order to decrease the quantity, this method chooses samples from the majority class that are either the nearest or the farthest from the minority class. This method decreases the dataset's size, resulting in faster training and processing time. NearMiss can mitigate overfitting by simplifying the dataset. [35] Subsequently, we assess the derived models by means of precision, recall, accuracy, and AUC metrics. The formula used for evaluation is as follows:

$$Precision\ (Prec) = \frac{TP}{TP + FP} \quad (4)$$

$$Recall\ (Rec) = \frac{TP}{TP + FN} \quad (5)$$

$$Accuracy\ (Acc) = \frac{TP + TN}{TP + TN + FP + FP} \quad (6)$$

accompanied by:
True Positive (TP) : the quantity of valid predictions in which the classification precisely identifies the positive class as positive.
True Negative (TN) : the number of precise predictions in which the classifier correctly indicates that the negative class is negative.
False Positive (FP) : those cases where the classifier gets a negative class name wrong.
False Negative (FN) : how often the classifier gets a positive class wrong and calls it a negative one.

The AUC of the Receiver Operating Characteristic (ROC) curve is a statistic used to assess classification model performance across various threshold levels. The ROC curve is a graphical depiction of the probability distribution, while the AUC (Area Under the Curve) measures how well the data points can be separated from one another. The AUC is a square-shaped zone with a constant value between 0 and 1.

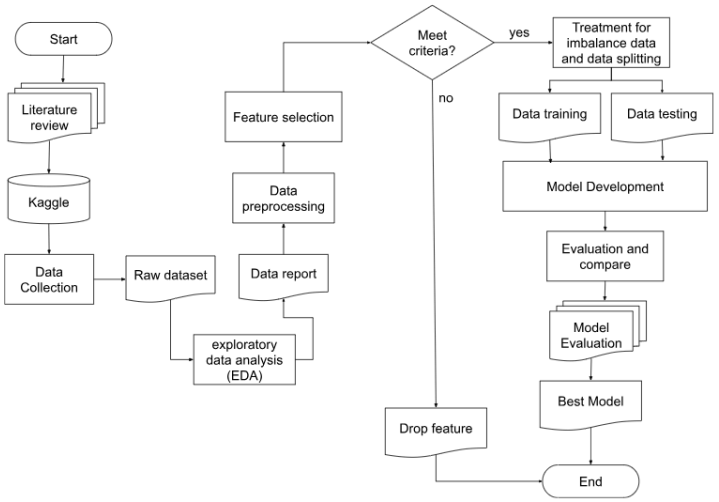


Figure 3 Research Framework

JURNAL ILMIAH FIFO
Vol. 17 No. 2 (2025) 102-118
DOI: 10.22441/fifo.2025.v17i2.001

P-ISSN 2085-4315
E-ISSN 2502-8332

Accuracy, recall, and precision are essential factors for assessing categorization models. The optimal model is determined by comparing the accuracy of each model. Furthermore, we compared these results with the mean squared error (MSE) and root mean square error (RMSE) values to confirm the accuracy of the indicators used in the model. This formula calculates the MSE and RMSE.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}\right)^2 \quad (7)$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}\right)^2} \quad (8)$$

with:
Y_i : mean value of y
$\hat{Y}$: predicted value of y

The MSE and RMSE readings are consistently positive. The optimal values for MSE and RMSE are both 0, whereas the worst value is represented by positive infinity (+∞). The reason we prefer using RMSE is that it provides a measure of the least number of mistakes, and it allows us to manage larger residual errors with more care.

3. RESULTS AND DISCUSSION

3.1. Data Collection

This study utilizes a dataset obtained from Kaggle. This study utilizes the Behavioral Risk Factor Surveillance System (BRFSS) indicators for the year 2015. This study employed an untainted dataset comprising 253,680 survey responses that were gathered for the CDC's BRFSS 2015. The dataset contains a total of 21 distinct characteristics. The dependent variable for this research is Diabetes_binary. The variable is bounded inclusively between 0 and 1. A result of 0 indicates the lack of diabetes, while a value of 1 indicates the presence of prediabetes or diabetes. The study involves 21 predictor variables, encompassing factors such as blood pressure, cholesterol levels, body mass index, smoking habits, history of stroke or heart disease, physical activity, dietary habits, alcohol consumption, access to healthcare, general health, mental health, physical health, mobility, gender, age, education, and income.

3.2 Exploratory Data Analysis (EDA)

The dataset utilized comprises 253,680 records. Every record is of the float data type. No missing values were detected in the dataset. Out of the 22 features present in the dataset, 15 of them have a binary value. Figure 4 illustrates the arrangement of features that have two possible values.

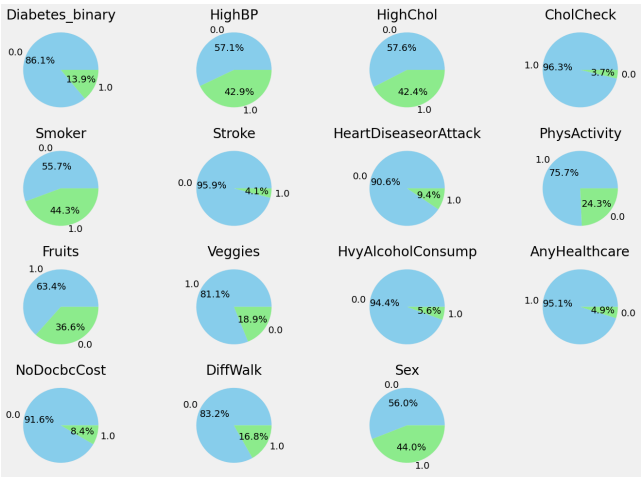
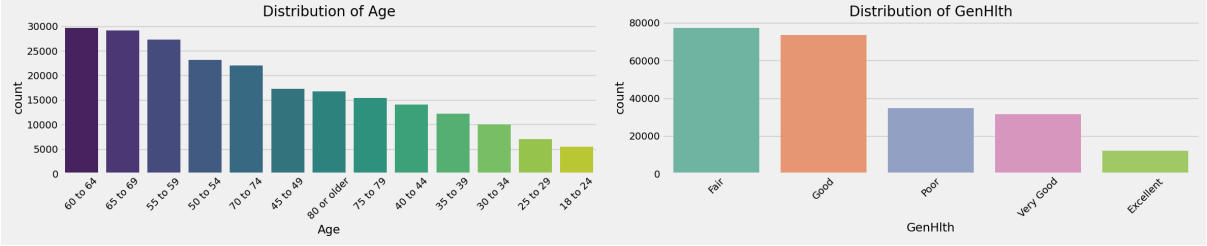


Figure 4 Binary variable distribution

The highest percentage of people who smoke, have high blood pressure, or have high cholesterol is over 40%. Those who have had a stroke or have not checked their cholesterol in the previous five years account for

JURNAL ILMIAH FIFO
Vol. 17 No. 2 (2025) 102-118
DOI: 10.22441/fifo.2025.v17i2.001
P-ISSN 2085-4315
E-ISSN 2502-8332

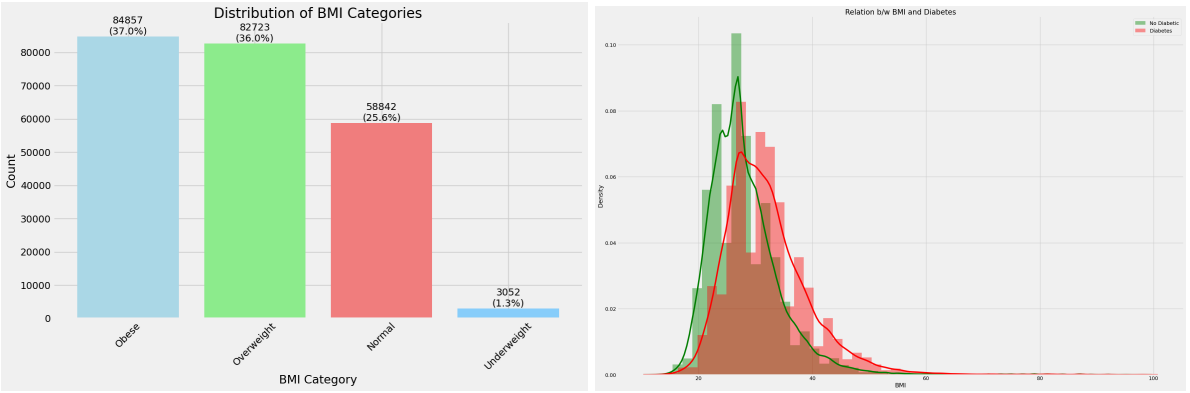
the lowest number, less than 5%. Figure 5 (a) shows the age distribution of respondents. The age range of 55–64 years holds the majority of individuals, while the age range of 18–44 years has the lowest percentage. Figure 5 (b) shows the general health distribution of respondents. "College students" have the highest frequency of education, followed by "some college or technical school," while "never attended school" has the lowest frequency.



(a) (b)

Figure 5 (a) Age distribution; (b) General health distribution

Figure 6 (a) displays the distribution of BMI. A significant proportion of individuals in this sample exhibit overweight or obesity, while only 25.6% have a BMI within the normal range. Figure 6 (b) displays the overlap between the Body Mass Index (BMI) of persons with prediabetes or diabetes and individuals without prediabetes or diabetes. Therefore, it may be inferred that BMI cannot be relied upon as a definitive indicator for identifying the prediabetes or diabetes stage.



(a) (b)

Figure 6 (a) BMI distribution; (b) BMI distribution for non diabetes and diabetes

3.3 Data Preprocessing

Data preparation is the process of transforming raw data into high-quality data that is appropriate for further processing in the subsequent step. Additionally, we do data cleansing throughout this phase. A total of 24,206 duplicate entries were identified among the 253,680 items in the dataset. We will remove redundant data from the collection, resulting in 229,474 remaining entries. During this stage, we convert the data values from floating-point numbers to integers in order to make their processing easier.

3.4 Feature Selection

Feature selection involves determining the most pertinent subset of features from a given dataset. The objective is to enhance the efficiency of machine learning models by decreasing data dimensionality, mitigating overfitting, accelerating model training time, and enhancing the interpretability of outcomes. Initially, we proceeded by determining the correlation between the predictor factors and the dependent variable. According to Figure 7, the GenHlth feature exhibits the strongest association, whilst the NoDocbcCost feature has the weakest correlation. The variables GenHlth, HighBP, DiffWalk, and BMI exhibit a high correlation. The variables HighChol, Age, HeartDiseaseorAttack, PhysHlth, Stroke, CholCheck, and MentHlth exhibit a moderate positive connection. Meanwhile, a moderate correlation with a negative association was discovered in the variables of

P-ISSN 2085-4315
E-ISSN 2502-8332

Income, Education, Physical Activity, and Heavy Alcohol Consumption. The features with poor connection are Smoker, Sex, AnyHealthcare, NoDocbcCost, Fruits, and Veggies.

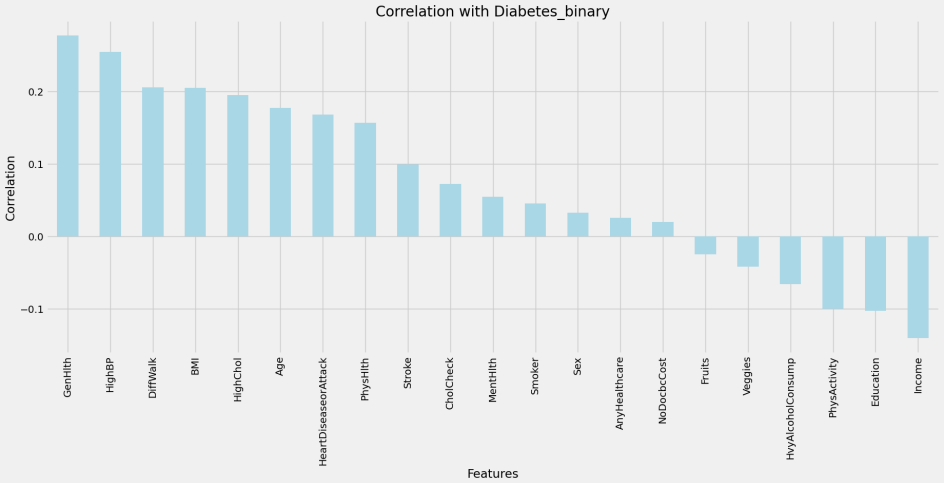


Figure 7 Correlation values with Diabetes_binary

Once correlations have been identified, the next stage involves doing VIF and chi-square tests. Table 2 displays the outcomes of VIF and chi-square computations. The VIF test results indicate that GenHlth had the greatest value of 1.7415, while HvyAlcoholConsump had the lowest value of 1.0278. Since all characteristics have values below 10, there is no presence of multicollinearity. The maximum chi-square value for the PhysHlth feature is 97,988.7617, while the lowest value is 7.9497 for the AnyHealthcare feature. According to these findings, five characteristics were excluded, specifically Fruits, Veggies, NoDocbcCost, CholCheck, and AnyHealthcare.

Table 2 VIF and Chi-square Score

Feature	VIF	Chi-square
PhysHlth	1.5943	97,988.7617
BMI	1.1418	15,507.7362
MentHlth	1.2218	11,419.5848
Age	1.3590	8,539.9063
HighBP	1.3152	8,098.5482
DiffWalk	1.5139	7,875.4962
GenHlth	1.7415	7,671.7328
HeartDiseaseorAttack	1.1704	5,822.1457
HighChol	1.1664	4,869.3127
Income	1.4318	3,377.0993
Stroke	1.0779	2,156.6784
HvyAlcoholConsump	1.0278	937.4011
PhysActivity	1.1306	617.5639
Education	1.2721	479.1129
Smoker	1.0761	253.8261
Sex	1.0767	137.8371
NoDocbcCost	1.1357	83.6628
Veggies	1.0981	82.0988
Fruits	1.0980	54.6889
CholCheck	1.0360	48.9041
AnyHealthcare	1.1099	7.9497

P-ISSN 2085-4315
E-ISSN 2502-8332

3.5 Model Development

Addressing imbalanced data is a crucial factor to take into account during model development. The Diabetes_binary variable consists of 194,377 instances with a value of 0 and 35,097 instances with a value of 1. This circumstance signifies that the dataset is undergoing undersampling. In order to tackle this situation, we employ the NearMiss technique to choose samples from the majority class, guaranteeing their equilibrium with the minority class. We will choose a subset of 35,097 from the total of 194,377 records. Therefore, the dataset was constructed using a total of 70,194 records. After balancing the dataset, we will partition it into a training dataset and a testing dataset with a ratio of 70:30. The model's construction utilized the parameters shown in Table 3. This research introduces two enhanced models, namely XGBoost and Random Forest, which are improvements over the original model. Meanwhile, the remaining three models are fundamental machine learning models.

Table 3 Model Development

No.	Model	Base model	Parameter
1	Support vector machines (SVM)	-	kernel='rbf', C=1.0
2	XGBoost	Gradient boosted trees	eval_metric= 'error', learning_rate= 0.1
3	Random Forest	Decision Tree	max_depth=12 ,n_estimators =10, random_state=42
4	Naive Bayes (NB)	Bayes theorem	priors=None, var_smoothing=1e-9
5	Artificial Neural Networks (ANN)	-	optimizer='adam', loss='binary_crossentropy', metrics=['accuracy']

3.6 Model Evaluating and Comparing

Five models were developed specifically for predicting the risk of diabetes. Figure 8 displays the precision and recall value attained by each generated model. Overall, all models exhibit high precision and recall scores, surpassing 0.8. The XGBoost and ANN models had the highest precision, however only the XGBoost model achieved the maximum recall. According to Figure 8, the XGBoost model attains the best accuracy of 0.87. The Random Forest and Artificial Neural Network (ANN) models both attained an accuracy of 0.86. The Naive Bayes model exhibited the lowest accuracy, attaining a mere 0.82. The Naive Bayes model is noteworthy because of its high precision, however it falls short in achieving high accuracy. The presence of a tradeoff between precision and recall results in ramifications for the overall accuracy level.

Figure 8 Precision, Recall, and Accuracy

Table 4 shows the results of model evaluation using training and testing data. Using training data, the XGBoost model achieved the highest level of accuracy with 0.8725. The lowest level of accuracy was in the Naive Bayes model, with 0.8203. Using testing data, the XGBoost model still achieved the highest accuracy value of 0.8690. The Naive Bayes model continues to be the model with the lowest accuracy. According to a

P-ISSN 2085-4315
E-ISSN 2502-8332

comparison of accuracy using training and testing datasets, the five models developed did not experience overfitting. Most models experienced a decrease in accuracy when using the testing dataset, but an anomaly occurred in the SVM model. The SVM model shows better accuracy results using the testing dataset. The built model has low MSE and RMSE values. The lowest MSE value is 0.1310 for the XGBoost model. The highest MSE value is 0.1798 for the Naive Bayes model. The lowest RMSE value in the XGBoost model is 0.3620. Meanwhile, the highest value achieved by the Naive Bayes model was 0.4240.

Table 4 Model Evaluation

Model	Training set accuracy	Test set accuracy	MSE	RMSE
XGBoost	0.8725	0.8690	0.1310	0.3620
Random Forest	0.8689	0.8593	0.1407	0.3752
ANN	0.8579	0.8583	0.1417	0.3764
SVM	0.8493	0.8508	0.1492	0.3862
Naive Bayes	0.8203	0.8202	0.1798	0.4240

In binary classification problems, the Area Under the Curve (AUC) statistic provides an all-around evaluation of a model's performance. The AUC value is a numerical measure that varies between 0 and 1. A value of 0.5 suggests a model that has no ability to distinguish between classes (random guessing), while a value of 1.0 shows a model that is excellent in its discrimination. The Support Vector Machine (SVM) model has robust discriminatory ability, accurately differentiating between the positive and negative categories with an approximate success rate of 85%. XGBoost demonstrates the highest Area Under the Curve (AUC) score compared to the other models, indicating its superior capability to distinguish between the classes. An AUC of 0.8682 indicates that the XGBoost model accurately prioritizes positive instances above negative ones almost 87% of the time. The Random Forest model has robust performance, achieving an AUC of 0.8584. The Artificial Neural Network (ANN) model demonstrates a notable AUC score of 0.8645, indicating its robust capability to differentiate between the classes with an accuracy of roughly 86% in prioritizing true positives over false positives. Among the models listed, Naive Bayes has the lowest AUC score of 0.8188, indicating a relatively good performance.

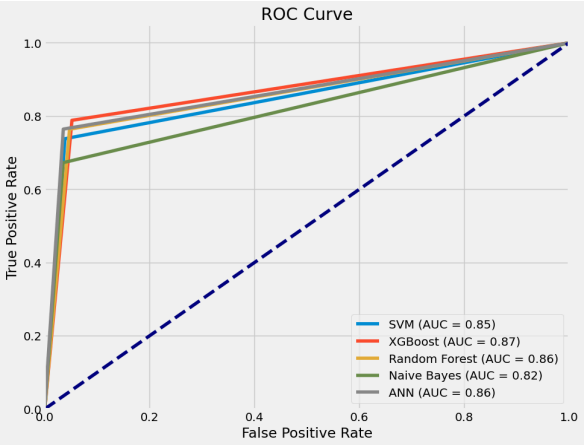


Figure 9 ROC Curve

3.7 Discussion

The efficacy of machine learning algorithms in classification tasks that involve numerous classes is assessed using a confusion matrix, which is a metric. The perplexity matrix is a type of tabular representation that illustrates four unique combinations of actual and anticipated values. The confusion matrix for the

JURNAL ILMIAH FIFO
Vol. 17 No. 2 (2025) 102-118
DOI: 10.22441/fifo.2025.v17i2.001

P-ISSN 2085-4315
E-ISSN 2502-8332

constructed model is seen in Figure 10. This graph facilitates the assessment of the efficacy of the built prediction model. The confusion matrix provides insights into the critical issue of model performance. Overall, it would be far more detrimental if the model erroneously predicted that a patient does not have diabetes, despite the fact that they really have.

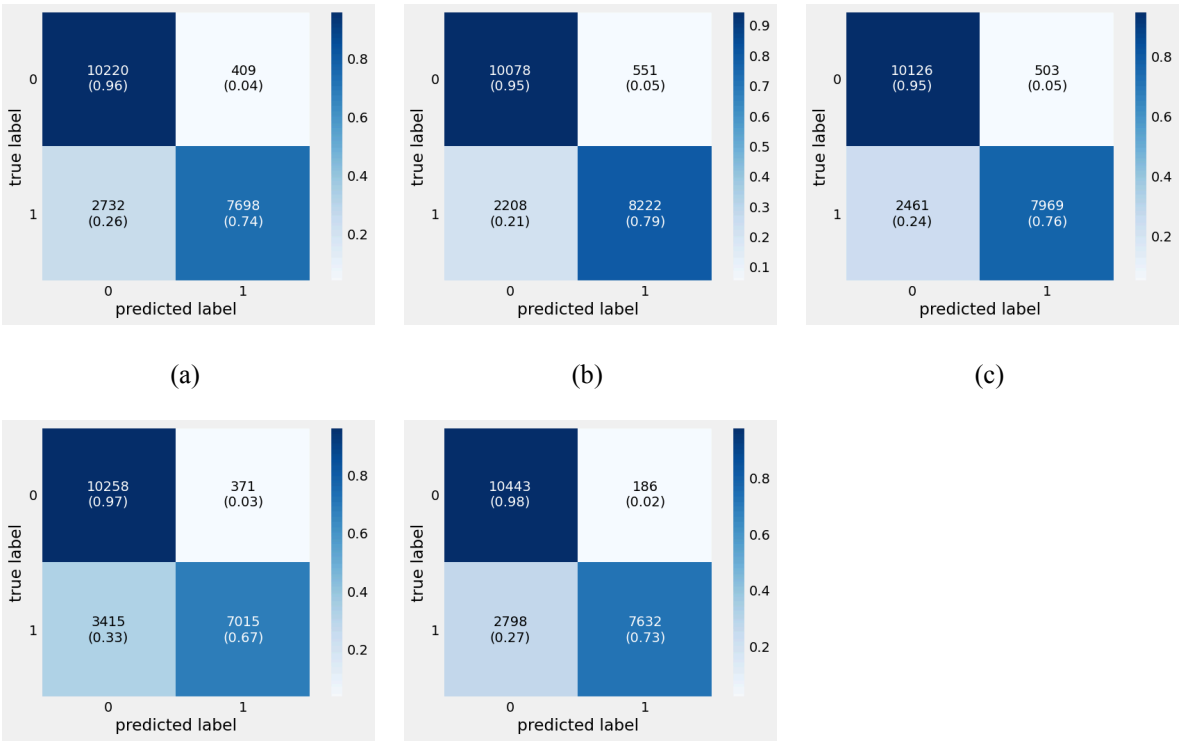


Figure 10 Confusion Matrix for: (a) SVM; (b) XGBoost; (c) Random Forest; (d) Naive Bayes; (e) ANN

The results of this study show that the XGBoost model is the best model for predicting diabetes. The XGBoost model achieved an accuracy of 0.8690, a precision of 0.88, and a recall of 0.87. The results are close to those of the study [13] and are still within the range achieved in studies [7] and [18], although they do not reach optimum results. The SVM model in this study achieved better accuracy than the study [1] in predicting diabetes. Despite the lower accuracy compared to the study [14]. The ANN model in this study achieved better accuracy than the study [1] for diabetes prediction. However, this study's ANN model achieved lower accuracy than studies [16] and [17]. This research's Random Forest model only achieved 0.8593, lower than the study [22] which achieved 0.8803. The accuracy results for the Naive Bayes model are similar to those in the study [1]. Even though the Naive Bayes model achieves the lowest accuracy compared to other models, this accuracy achievement is still within the accuracy range, according to studies [23].

The model developed by this research produces an MSE value in the range of 0.1310 to 0.1798. We can categorize this value as very good because it is close to 0. This research's model produces an RMSE value within the range of 0.3620 - 0.4240. We categorize this value as good because it remains below 0.5. These two indicators indicate that the features used to build the model are effective. This result also confirms the correct handling of imbalanced data and feature selection. The absence of an overfitting model serves as proof of this. All the selected features act as good predictors.

All of the models exhibit strong performance, as evidenced by their AUC ratings consistently exceeding 0.8. This indicates that all of them are highly effective classifiers. Nevertheless, XGBoost stands out as the top performance, achieving the greatest AUC score of 0.8682, with ANN and Random Forest closely trailing after. Although Naive Bayes is slightly less advanced than the other methods, it still works adequately.

4. CONCLUSIONS

For this study, we have gathered a total of 253,680 data entries from the BRFSS indicators dataset. The

dataset utilized consists of 21 distinct features. The dataset exhibits a data imbalance in the dependent variable. A total of 194,377 entries were found to have a value of 0, whereas only 35,097 records had a value of 1. Using the data profile, we performed feature selection and identified 15 out of the 21 available features as the most optimal. In addition, we utilize the NearMiss approach to address undersampling, hence achieving a balanced dependent variable.

According to the findings of the model evaluation, the XGBoost model is the most optimal choice. The XGBoost model achieved a precision of 0.88, a recall of 0.87, and an accuracy of 0.8690. The XGBoost model has a RMSE of 0.3620 and a MSE of 0.1310. An AUC value of 0.8682 indicates that the XGBoost model accurately prioritizes positive instances above negative examples almost 87% of the time. This model excels at effectively managing intricate patterns and interactions within the data.

Further research should include the development of feature selection techniques for future investigations. Feature selection strategies can potentially enhance model performance. In addition, the Random Forest and Naive Bayes models possess significant untapped potential for further advancement. The study [20] documented that the Random Forest achieved a performance level of up to 0.99, whereas the study [4] reported that the Naive Bayes model achieved a performance level of up to 0.9239. Further investigation could also concentrate on amalgamating diverse machine learning models to attain enhanced performance.

REFERENCES

[1] H. E. Massari, Z. Sabouri, S. Mhammedi, and N. Gherabi, "Diabetes Prediction Using Machine Learning Algorithms and Ontology," *J. ICT Stand.*, May 2022, doi: 10.13052/jicts2245-800X.10212.

[2] M. M. Farag, M. Fouad, and A. T. Abdel-Hamid, "Automatic Severity Classification of Diabetic Retinopathy Based on DenseNet and Convolutional Block Attention Module," *IEEE Access*, vol. 10, pp. 38299–38308, 2022, doi: 10.1109/ACCESS.2022.3165193.

[3] Z. Xie, O. Nikolayeva, J. Luo, and D. Li, "Building Risk Prediction Models for Type 2 Diabetes Using Machine Learning Techniques," *Prev. Chronic. Dis.*, vol. 16, p. 190109, Sep. 2019, doi: 10.5888/pcd16.190109.

[4] S. Gupta, B. Kishan, and P. Gulia, "Comparative Analysis of Predictive Algorithms for Performance Measurement," *IEEE Access*, vol. 12, pp. 33949–33958, 2024, doi: 10.1109/ACCESS.2024.3372082.

[5] H. Zhang and Y. Zhang, "An Improved Sparrow Search Algorithm for Optimizing Support Vector Machines," *IEEE Access*, vol. 11, pp. 8199–8206, 2023, doi: 10.1109/ACCESS.2023.3234579.

[6] S. Liang, A. Q. M. Sabri, F. Alnajjar, and C. K. Loo, "Autism Spectrum Self-Stimulatory Behaviors Classification Using Explainable Temporal Coherency Deep Features and SVM Classifier," *IEEE Access*, vol. 9, pp. 34264–34275, 2021, doi: 10.1109/ACCESS.2021.3061455.

[7] D. N and N. P. K S, "Improved Clinical Diagnosis Using Predictive Analytics," *IEEE Access*, vol. 10, pp. 75158–75175, 2022, doi: 10.1109/ACCESS.2022.3190416.

[8] N. Assani, P. Matić, N. Kaštelan, and I. R. Čavka, "A Review of Artificial Neural Networks Applications in Maritime Industry," *IEEE Access*, vol. 11, pp. 139823–139848, 2023, doi: 10.1109/ACCESS.2023.3341690.

[9] B. A. S. Emambocus, M. B. Jasser, and A. Amphawan, "A Survey on the Optimization of Artificial Neural Networks Using Swarm Intelligence Algorithms," *IEEE Access*, vol. 11, pp. 1280–1294, 2023, doi: 10.1109/ACCESS.2022.3233596.

[10] C. Liu, Z. Gu, and J. Wang, "A Hybrid Intrusion Detection System Based on Scalable K-Means+ Random Forest and Deep Learning," *IEEE Access*, vol. 9, pp. 75729–75740, 2021, doi: 10.1109/ACCESS.2021.3082147.

[11] Z. Huang and D. Chen, "A Breast Cancer Diagnosis Method Based on VIM Feature Selection and Hierarchical Clustering Random Forest Algorithm," *IEEE Access*, vol. 10, pp. 3284–3293, 2022, doi: 10.1109/ACCESS.2021.3139595.

[12] G. P. Shukla, S. Kumar, S. K. Pandey, R. Agarwal, N. Varshney, and A. Kumar, "Diagnosis and Detection of Alzheimer's Disease Using Learning Algorithm," *Big Data Min. Anal.*, vol. 6, no. 4, pp. 504–512, Dec. 2023, doi: 10.26599/BDMA.2022.9020049.

[13] P. Haldar *et al.*, "XGBoosted Binary CNNs for Multi-Class Classification of Colorectal Polyp Size," *IEEE Access*, vol. 11, pp. 128461–128472, 2023, doi: 10.1109/ACCESS.2023.3332826.

[14] M. Varan, J. Azimjonov, and B. Maçal, "Enhancing Prostate Cancer Classification by Leveraging Key Radiomics Features and Using the Fine-Tuned Linear SVM Algorithm," *IEEE Access*, vol. 11, pp.

88025–88039, 2023, doi: 10.1109/ACCESS.2023.3306515.

[15] H. M. Alshamlan, "An Effective Filter Method Towards the Performance Improvement of FF-SVM Algorithm," *IEEE Access*, vol. 9, pp. 140835–140840, 2021, doi: 10.1109/ACCESS.2021.3119233.

[16] T. S. Alshammari, "Applying Machine Learning Algorithms for the Classification of Sleep Disorders," *IEEE Access*, vol. 12, pp. 36110–36121, 2024, doi: 10.1109/ACCESS.2024.3374408.

[17] S. Punitha, T. Stephan, R. Kannan, M. Mahmud, M. S. Kaiser, and S. B. Belhaouari, "Detecting COVID-19 From Lung Computed Tomography Images: A Swarm Optimized Artificial Neural Network Approach," *IEEE Access*, vol. 11, pp. 12378–12393, 2023, doi: 10.1109/ACCESS.2023.3236812.

[18] J.-G. Choi, I. Ko, and S. Han, "Depression Level Classification Using Machine Learning Classifiers Based on Actigraphy Data," *IEEE Access*, vol. 9, pp. 116622–116646, 2021, doi: 10.1109/ACCESS.2021.3105393.

[19] C. Zhang, X. Wang, S. Chen, H. Li, X. Wu, and X. Zhang, "A Modified Random Forest Based on Kappa Measure and Binary Artificial Bee Colony Algorithm," *IEEE Access*, vol. 9, pp. 117679–117690, 2021, doi: 10.1109/ACCESS.2021.3105796.

[20] T. Sinha Roy, J. K. Roy, and N. Mandal, "Conv-Random Forest-Based IoT: A Deep Learning Model Based on CNN and Random Forest for Classification and Analysis of Valvular Heart Diseases," *IEEE Open J. Instrum. Meas.*, vol. 2, pp. 1–17, 2023, doi: 10.1109/OJIM.2023.3320765.

[21] T.-H. S. Li, H.-J. Chiu, and P.-H. Kuo, "Hepatitis C Virus Detection Model by Using Random Forest, Logistic-Regression and ABC Algorithm," *IEEE Access*, vol. 10, pp. 91045–91058, 2022, doi: 10.1109/ACCESS.2022.3202295.

[22] L. Jia, Z. Wang, S. Lv, and Z. Xu, "PE_DIM: An Efficient Probabilistic Ensemble Classification Algorithm for Diabetes Handling Class Imbalance Missing Values," *IEEE Access*, vol. 10, pp. 107459–107476, 2022, doi: 10.1109/ACCESS.2022.3212067.

[23] Z. Ahmed, B. Issac, and S. Das, "Ok-NB: An Enhanced OPTICS and k-Naive Bayes Classifier for Imbalance Classification With Overlapping," *IEEE Access*, vol. 12, pp. 57458–57477, 2024, doi: 10.1109/ACCESS.2024.3391749.

[24] H. C. S. C. Lima, F. E. B. Otero, L. H. C. Merschmann, and M. J. F. Souza, "A Novel Hybrid Feature Selection Algorithm for Hierarchical Classification," *IEEE Access*, vol. 9, pp. 127278–127292, 2021, doi: 10.1109/ACCESS.2021.3112396.

[25] G. J. Ansari, J. H. Shah, M. C. Q. Farias, M. Sharif, N. Qadeer, and H. U. Khan, "An Optimized Feature Selection Technique in Diversified Natural Scene Text for Classification Using Genetic Algorithm," *IEEE Access*, vol. 9, pp. 54923–54937, 2021, doi: 10.1109/ACCESS.2021.3071169.

[26] A. K. Mandal, Md. Nadim, H. Saha, T. Sultana, Md. D. Hossain, and E.-N. Huh, "Feature Subset Selection for High-Dimensional, Low Sampling Size Data Classification Using Ensemble Feature Selection With a Wrapper-Based Search," *IEEE Access*, vol. 12, pp. 62341–62357, 2024, doi: 10.1109/ACCESS.2024.3390684.

[27] L. Al-Shalabi, "New Feature Selection Algorithm Based on Feature Stability and Correlation," *IEEE Access*, vol. 10, pp. 4699–4713, 2022, doi: 10.1109/ACCESS.2022.3140209.

[28] F. Feng, K.-C. Li, J. Shen, Q. Zhou, and X. Yang, "Using Cost-Sensitive Learning and Feature Selection Algorithms to Improve the Performance of Imbalanced Classification," *IEEE Access*, vol. 8, pp. 69979–69996, 2020, doi: 10.1109/ACCESS.2020.2987364.

[29] S. Rahman and K. Adhikari, "Comparative Analysis of SVM and CNN for Sonar Signal Classification Using Sparse Arrays," *IEEE Access*, vol. 12, pp. 59818–59830, 2024, doi: 10.1109/ACCESS.2024.3393893.

[30] R. Obiedat *et al.*, "Sentiment Analysis of Customers' Reviews Using a Hybrid Evolutionary SVM-Based Approach in an Imbalanced Data Distribution," *IEEE Access*, vol. 10, pp. 22260–22273, 2022, doi: 10.1109/ACCESS.2022.3149482.

[31] R. Guo, Z. Zhao, T. Wang, G. Liu, J. Zhao, and D. Gao, "Degradation State Recognition of Piston Pump Based on ICEEMDAN and XGBoost," *Appl. Sci.*, vol. 10, no. 18, p. 6593, Sep. 2020, doi: 10.3390/app10186593.

[32] S. Naiem, A. E. Khedr, A. M. Idrees, and M. I. Marie, "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.

P-ISSN 2085-4315
E-ISSN 2502-8332

[33] N. Shrestha, "Detecting Multicollinearity in Regression Analysis," *Am. J. Appl. Math. Stat.*, vol. 8, no. 2, pp. 39–42, Jun. 2020, doi: 10.12691/ajams-8-2-1.

[34] F. Al Anshory, S. Siswanto, S. A. Thamrin, and I. Inayah, "Improved Chi Square Automatic Interaction Detection on Students Discontinuation to Secondary School," *J. Varian*, vol. 7, no. 1, pp. 15–26, Oct. 2023, doi: 10.30812/varian.v7i1.2627.

[35] Z. S. Rubaidi, B. B. Ammar, and M. B. Aouicha, "Fraud Detection Using Large-scale Imbalance Dataset," *Int. J. Artif. Intell. Tools*, vol. 31, no. 08, p. 2250037, Dec. 2022, doi: 10.1142/S0218213022500373.