Scientific Journal of Informatics
Vol. 11, No. 3, Aug 2024

p-ISSN 2407-7658 https://journal.unnes.ac.id/journals/sji/index e-ISSN 2460-0040

Principal Component Analysis for Prediabetes Prediction using Extreme Gradient Boosting (XGBoost)

Kartina Diah Kesuma Wardhani[1*], Wenda Novayani[2]
[1,2]Department of Information Technology, Caltex Politechnic of Riau, Indonesia

Abstract.
Purpose: The purpose of this study is to increase the accuracy of the model used for prediabetes prediction. This study integrates Principal Component Analysis (PCA) for reducing the dimension of data with Extreme Gradient Boosting (XGBoost). The study contributes to providing a new alternative for prediabetes prediction in patients by reducing the complexity of the dataset with the aim of increasing the accuracy of the obtained model. PCA and XGBoost identify the best features that have the highest correlation with prediabetes so that they are expected to produce a better predictive model.
Methods: This study utilizes published data sourced from the UCI Machine Learning Repository consisting of 520 records, 16 attributes and 1 label class. The dataset is data collected through direct questionnaires from patients in Sylhet, Bangladesh at the Sylhet Diabetes Hospital. The research method in this study consists of several stages, namely: Data Collection, Data Preprocessing, Dimension Reduction using PCA to reduce the complexity of dimensions in the dataset, Modeling using XGBoost to identify patterns used to predict prediabetes, and Model evaluation used to measure the performance of the resulting model using evaluation metrics such as accuracy, recall, precision and F1-Score.
Result: The current study utilizes XGBoost with Principal Component Analysis for feature selection, resulting in 12 features and a model accuracy of 97.44.
Novelty: The study's originality lies in applying PCA as a preprocessing step to enhance the performance of machine learning models by reducing data dimensionality and focusing on the most critical features. By demonstrating how PCA can improve the efficiency and accuracy of prediabetes prediction models, this research provides valuable insights to inform future studies and contribute to the development of more effective diagnostic tools for early detection and prevention of prediabetes.

Keywords: Prediabetes, Medical data, PCS, XGBoost
Received September 2024 / **Revised** October 2024 / **Accepted** November 2024

INTRODUCTION

Prediabetes is the body's early warning for diabetes. The sooner prediabetes is detected, the sooner preventive measures against diabetes can be taken by patients. Prediabetes that is not well controlled will become diabetes at any time. Prediabetes is condition when blood glucose levels that are above normal but still below the lower limit for type 2 diabetes [1]–[3]. Patients with prediabetes or diabetes can have mild symptoms such as frequent thirst, frequent urination, fatigue, and blurred vision [4], [5].

Generally, prediabetes can be identified using two standard tests, namely the fasting plasma glucose (FPG) test and the oral glucose tolerance test (OGTT) [2], [6]. There are differences between the two tests, FPG functions to measure blood glucose levels after the patient has fasted from the evening. While OGTT functions to measure blood glucose levels after the patient drinks a sweet drink. It is not certain what causes someone to get prediabetes, but there are several factors that can influence the development of prediabetes into diabetes, including genetics, lifestyle, obesity and metabolic syndrome. Someone with obesity who rarely moves in their daily activities, has a habit of consuming unhealthy foods and a family history of diabetes increases the risk of prediabetes at an early age [7].

Over the last decades, multidisciplinary expertise on prediabetes has involved collaborations among researchers, clinicians, epidemiologists, geneticists, and public health experts, leading to improved diagnostic criteria and prevention strategies. Knowing these factors can assist in identifying individuals at

[*] Corresponding author.
Email addresses: diah@pcr.ac.id (Wardhani)
DOI: 10.15294/sji.v11i3.13416

risk, enabling early intervention and preventive measures to reduce the progression to type 2 diabetes and associated complications[8]. On the other hand, preventions and better treatments of diabetes and prediabetes are necessary to increase prosperity and reduce economic costs [9].

Many studies have been undertaken to predict prediabetes based on the symptoms, people's lifestyles, eating habits, and so on [10]. In line with the growth of unlimited data, including medical data, prediabetes predictions can also be made [11]. Predictions are used to develop by expanding methods and technologies to data analysis, such as machine learning[12]–[20]. Machine learning techniques in current years are the choice of researchers to analyze large datasets in order to obtain patterns that can then be used to make predictions [14]. In this case, the current prediabetes prediction also utilizes machine learning to produce patterns that can predict indications of prediabetes in patients.

Machine learning is a derivative of artificial intelligence which is now commonly used to study data patterns to help with decision making, predictions, estimation, forecasting, classification, associations and so on [21]. Prediabetes prediction using machine learning has been done previously by [22] by comparing several machine learning techniques such as XGBoost, Randon Forest, SVM ect with different number of features with this study. Another study by [23] used Gradient Boosting machine learning to predict the risk of transition from prediabetes to Type 2 Diabetes in one and five years. Research by [24], [25] also conducted modeling for prediabetes prediction using XGBoost and Feature Importance XGBoost.

Numerous investigations have focused on predicting prediabetes by harnessing the surge in expansive datasets encompassing medical information. To cope with the extensive data volume, various methodologies within machine learning, like Principal Component Analysis (PCA), can be utilized for reducing dimensions in dataset. PCA can be a valuable tool for reducing dimensionality and selecting features by identifying the most significant variables and reducing the dataset's complexity which its purpose is to enhance subsequent predictive models' accuracy and efficiency [26], [27]. As an unsupervised technique, PCA does not incorporate outcome as a class label from prediabetes during its calculations. PCA reduces a substantial dataset into a compact version while retaining nearly all the original information. This method identifies the meaning of the data and its principal components, making it a widely utilized dimensionality reduction approach. Typically, this technique's purpose is to enhance variance and capture significant feature patterns within a dataset [28].

The current study offers an alternative approach to expand the helpfulness of machine learning algorithms in predicting prediabetes. This is achieved by utilizing Principal Component Analysis (PCA) as a preprocessing step. This research uses PCA to determine the dataset's number of features and records. The outcomes of PCA then serve as the dataset for applying the extreme gradient boosting (XGBoost) algorithm, which operates as a supervised learning technique to discern patterns that could indicate the risk of prediabetes or facilitate predictive modeling.

METHODS

The significant stages of the purpose technique in this study include data acquisition, data preprocessing consisting of exploratory data analysis and transformation, feature selection based on PCA, modeling data utilizing the XGBoost Algorithm, and model evaluation performance. The order stages of the purpose technique are shown in Figure 1.

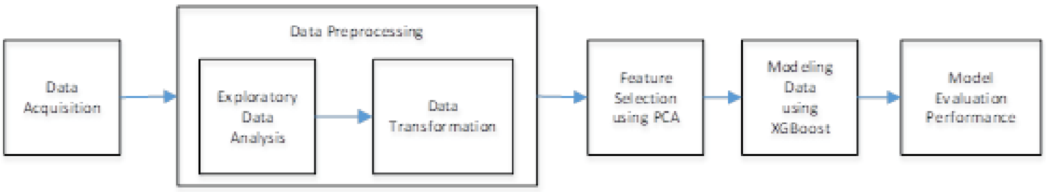
Data Acquisition
Data Preprocessing
Exploratory Data Analysis
Data Transformation
Feature Selection using PCA
Modeling Data using XGBoost
Model Evaluation Performance

Figure 1. Stages of the purpose technique

Dataset acquisition and data preprocessing

Data collection for the diabetes dataset was done by obtaining relevant health data related to diabetes from various web sources. To predict the potential of someone to get diabetes, a dataset that includes information on newly diagnosed or at-risk diabetic patients is required. This study used a dataset obtained via direct questionnaires to patients in Sylhet, Bangladesh at the Sylhet Diabetes Hospital. This dataset, which contains 520 records with 16 attributes and 1 class label, was sourced from the UCI Machine Learning Repository. It is specifically published as part of the prediabetes risk prediction dataset in UCI Machine Learning. Table 1 provides complete details about the dataset.

Table 1. Information of dataset

#	Column	Value	Non-Null Count	Dtype
1	Age	in years ranging from 20 years to 90 years	520 non-null	int64
2	Gender	Male/Female	520 non-null	object
3	Polyuria	Yes/No	520 non-null	object
4	Polydipsia	Yes/No	520 non-null	object
5	Sudden weight loss	Yes/No	520 non-null	object
6	Weakness	Yes/No	520 non-null	object
7	Polyphagia	Yes/No	520 non-null	object
8	Genital Thrush	Yes/No	520 non-null	object
9	Visual Blurring	Yes/No	520 non-null	object
10	Itching	Yes/No	520 non-null	object
11	Irritability	Yes/No	520 non-null	object
12	Delayed Healing	Yes/No	520 non-null	object
13	Partial Paresis	Yes/No	520 non-null	object
14	Muscle Stiffness	Yes/No	520 non-null	object
15	Alopecia	Yes/No	520 non-null	object
16	Obesity	Yes/No	520 non-null	object
17	Class	Positive/Negative	520 non-null	object

Preprocessing Data comprises two stages: Exploratory Data Analysis and Transformation. The initial step involves Exploratory Data Analysis, and Preliminary Data Assessment. Researchers utilize techniques representation of data and statistical techniques to depict characteristics of datasets, encompassing size, quantity, and precision, to comprehend data nature[29]. Exploratory data analysis methods span manual scrutiny and automated tools that visually investigate data variables' relationships, dataset structure, and data value distribution. It helps uncover patterns, enabling data analysts to gain deeper insights from raw data. This study uses an exploratory data analysis in order to identify characteristics of data, show correlations between attributes and labels of data, and descriptive research to adjust variable types. Figure 2 shows the spread of diabetes labels in the dataset, whereas Figure 3 shows the age distribution.

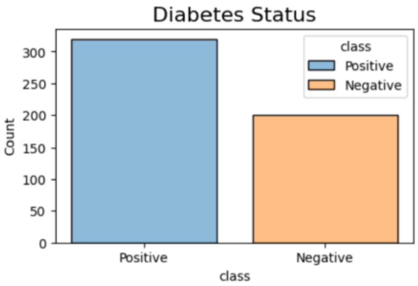

Figure 2. Diabetes label from dataset

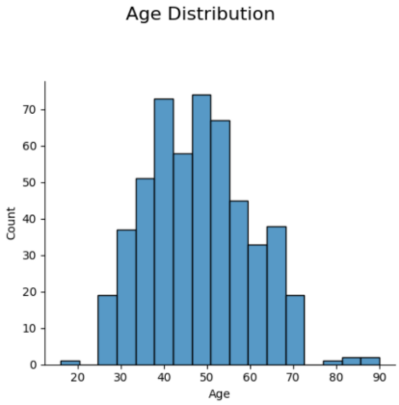

Figure 3. Age distribution in the dataset

After conducting Exploratory Data Analysis, the next step is Data Transformation. This process involves modifying the data format to ensure compatibility with the modeling algorithm used. Various techniques, such as binary transformation, can be applied during data transformation since machine learning algorithms cannot directly process raw text. [30]. Table 2 is the dataset after transformation. The Dtype attribute has changed to an integer value.

Table 2. Dataset after transformation

#	Column	Non-Null Count	Dtype
1	Age	520 non-null	int64
2	Gender	520 non-null	int32
3	Polyuria	520 non-null	int32
4	Polydipsia	520 non-null	int32
5	Sudden weight loss	520 non-null	int32
6	Weakness	520 non-null	int32
7	Polyphagia	520 non-null	int32
8	Genital Thrush	520 non-null	int32
9	Visual Blurring	520 non-null	int32
10	Itching	520 non-null	int32
11	Irritability	520 non-null	int32
12	Delayed Healing	520 non-null	int32
13	Partial Paresis	520 non-null	int32
14	Muscle Stiffness	520 non-null	int32
15	Alopecia	520 non-null	int32
16	Obesity	520 non-null	int32
17	Class	520 non-null	int32

Principal component analysis

PCA is utilized to standardize the dataset by identifying the principal components, which are new variables formed as linear combinations of the original variables. These components are arranged in a way that the first component captures the most variance in the data, the second component captures the next highest amount of variance.

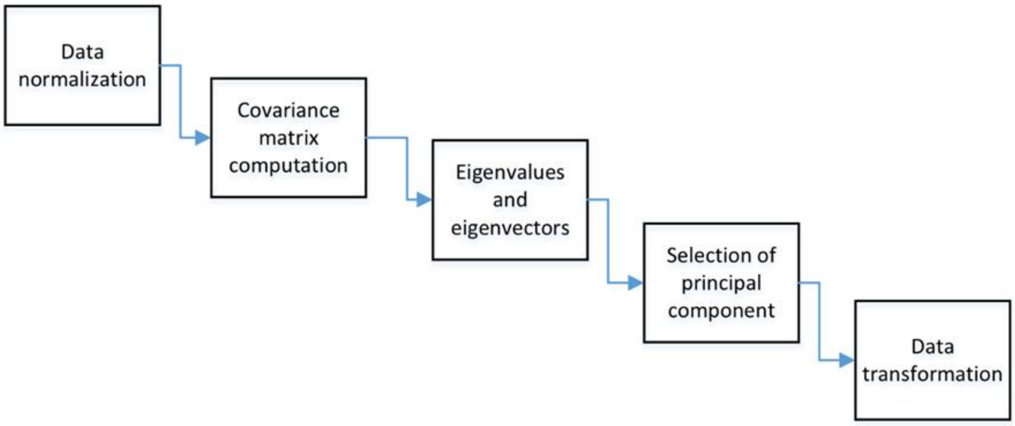
Data normalization
Covariance matrix computation
Eigenvalues and eigenvectors
Selection of principal component
Data transformation

Figure 4. PCA steps

After that, PCA Determines the Number of Components to analyze the variance explained by each principal component and decide how many components are to retain. The next step is Feature Selection. Once the number of components is determined, select the corresponding original features that contribute most to these components. This step helps identify the most important variables for predicting prediabetes.

Data modelling

A step-by-step explanation of prediabetes prediction using XGBoost works by defining an objective function, base learners, boosting, gradient descent optimization, tree construction, and regularization [13], [19].

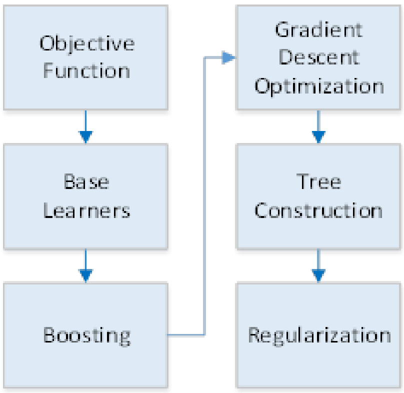

Figure 5. XGBoost step by step

The objective function comprises a loss function quantifying the model's predictive error and a regularization term penalizing intricate models to avert overfitting. The objective function steers the learning process by determining how the model updates its predictions in each iteration. Base Learners: XGBoost uses decision trees as base learners, where each tree predicts the target variable based on the input features. By default, XGBoost uses a shallow tree structure, a "weak learner," to minimize bias and allow for a more significant number of trees in the ensemble. However, depending on the problem's complexity and dataset characteristics, deeper trees can also be used [31].

Boosting: The boosting process involves adding new trees to the ensemble to improve the model's predictions. Each tree is constructed to address the errors or residuals left by the preceding trees. The process begins with a single tree, and the following trees are trained to reduce the discrepancy between the actual target values and the predictions made by the ensemble.

Gradient Descent Optimization: XGBoost employs gradient descent optimization to refine the model's predictions progressively. It computes the gradients of the loss function concerning the model's predictions and adjusts these predictions toward minimizing the loss. The learning rate, often referred to as the shrinkage parameter, governs the magnitude of each update, thereby mitigating overfitting and stabilizing the learning process.

Tree Construction: XGBoost greedily constructs each tree, optimizing the objective function at each split. It evaluates different partitions based on a scoring metric (such as information gain or Gini impurity) and selects the split that maximizes the growth in the objective function. This process repeats until a stopping criterion is met, such as reaching a maximum tree depth or when there is no significant improvement in the objective function.

Regularization: XGBoost uses regularization techniques to regulate model complexity and avoid overfitting. The objective function includes L1 (Lasso) and L2 (Ridge) regularization components, which penalize large coefficients and encourage sparsity, thus creating a simpler model and minimizing overfitting. In general, the formula for predicting target variables with XGBoost is as follows.

The general formula for predicting the target variable using XGBoost can be expressed in formula 1:

$$Y_{hat} = sum\_\{i = 1\}\hat{}\{N\_T\}f\_t(x) \tag{1}$$

where:
Y_hat is the predicted value of the target variable,
N_T is the total number of trees in the ensemble,
$f\_t(x)$ is the prediction made by the t-th tree for the input features x.
Each individual tree in the XGBoost ensemble predicts a value that is a combination of weighted decisions made at each split.

The prediction of a tree can be written in formula 2:

$$f_{t(x)} = w\_\{q(x)\} * h(x, q(x)) \tag{2}$$

where:
$w\_\{q(x)\}$ is the weight assigned to the leaf node q(x) where the input features x fall,
$h(x, q(x))$ is the output value assigned to the leaf node q(x) based on the input features x.

The weights w_{q(x)} are learned during the training process through gradient descent optimization, which minimizes a specific loss function.

Model evaluation performance

Confusion matrix provides a complete overview of model performance, by produce the number of true positives, true negatives, false positives, and false negatives [32]–[35]. This matrix is the basis for calculating various metrics such as accuracy, precision, recall, and F1 score. Accuracy measures the percentage of correct predictions from the total predictions. These metrics are very effective in balanced datasets but may not be ideal for describing the model performance in imbalanced datasets. Precision measures the proportion of correct positive predictions among all positive predictions. The focus of precision is on the accuracy of positive predictions, especially when false positives have a significant impact, thus reflecting the model capabilities for reducing errors in positive predictions. Recall (Sensitivity or True Positive Ratio) measures the percentage of true positive predictions among all true positive cases. Recall places more emphasis on identifying all positive examples and is important when false negatives have serious consequences. Recall evaluates the model's effectiveness in minimizing false negatives. F1-Score: The F1-score is the harmonic means of precision and recall, providing a balance between the two. It is a crucial metric when both false positives and false negatives are important.

RESULTS AND DISCUSSIONS

Testing environment

Prediabetes prediction modeling in this study uses Python 3 with pandas, sklearn, seaborn and matplotlib libraries at https://colab.research.google.com/. The hardware used is AMD Reyzen 7 6800U with Radeon Graphics, and 16GB Memory.

Principal component analysis

Figure 7 shows n_component against the explained variance and describes the cumulative defined variance ratio against the number of components to visually understand how much friction is captured by each additional element. This plot helps determining the point of diminishing returns in terms of explained variance and guides the selection of appropriate components for the analysis. As the number of components (n_components), it captures more variance in the data. The selected component as PCA's result describes the correlation of the selected component to the diabetes label listed in Table 3.

Table 3. Principal component correlation result

Principal component number	Correlation Point
1	0.2893
2	0.1487
3	0.1116
4	0.0773
5	0.0628
6	0.0597
7	0.0541
8	0.0541
9	0.0466
10	0.0451
11	0.2893
12	0.2893

PCA is used for dimensionality reduction, in terms of the number of features and rows, as input for the XGBoost model. The dimensions formed from the PCA results are 12 features and 364 rows.

Data modelling

Data modeling results from the XGBoost algorithm are shown in Figure 8. XGBoost is a tree-based model that employs the boosting process, which iteratively adds new trees to the ensemble to enhance the model's predictions. Each tree in XGBoost is constructed greedily, optimizing the objective function at each split. It evaluates various partitions based on a scoring metric, such as information gain or Gini impurity, and selects the split that maximizes the growth in the objective function. This process continues until a stopping criterion is met, such as reaching the maximum tree depth or achieving a minimum improvement in the objective function.

In this study, the tree model results from XGBoost based on PCA features are initiated with feature F1 as the root. This implies that feature F1, which has the highest correlation with the diabetes label, is selected as the starting point, followed by feature F7. The number of features in the XGBoost tree model combines 12 features resulting from PCA, either all or only a subset. This is due to the random processes applied during the model training.

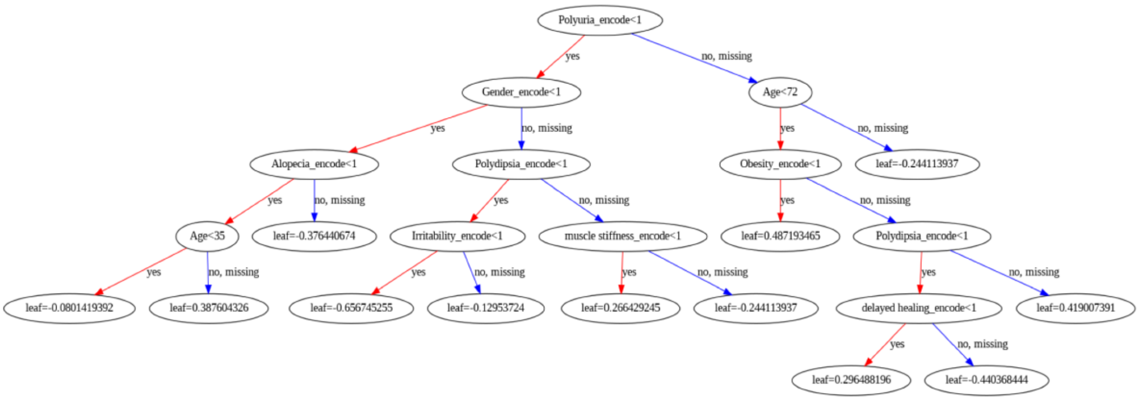

Figure 6. XGBoost model based on PCA

XGBoost also introduces randomness in feature selection. XGBoost considers a random subset of features at each tree split to find the best split. This feature subsampling introduces further randomness into the model-building process. Additionally, XGBoost employs early stopping to prevent overfitting. The training process stops if the performance does not improve for a certain number of iterations. As the iteration at which early stopping occurs may differ between runs, the final ensemble of trees can vary, but the resulting of the accuracy is the same.

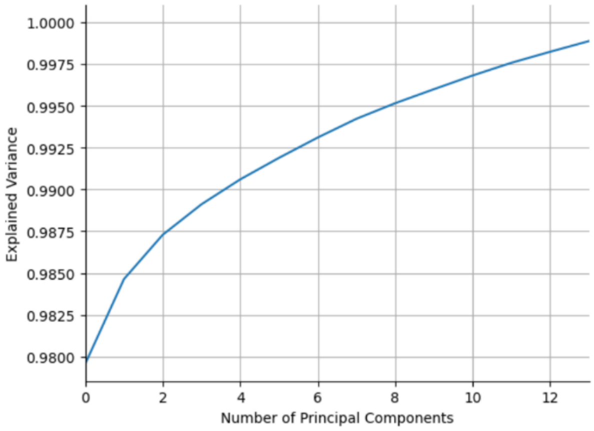

Figure 7. n_component against the explained variance

Model evaluation performance

The model evaluated in this study is divided into three purposes: first, model accuracy based on n_component PCA to determine the number of features that yield the best accuracy. Second, the model evaluation uses a confusion matrix to obtain accuracy, precision, recall, and F1-Score values. Third, a comparison of model evaluation results using the XGBoost algorithm with various feature selection techniques previously applied in the research. The accuracy of the model obtained based on the number of PCA components is shown in Figure 8.

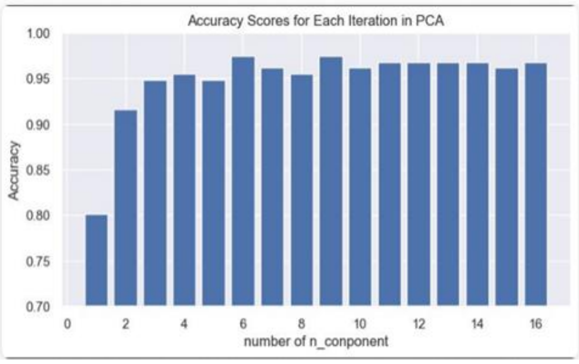

Figure 8. Accuracy score for n-component PCA

The accuracy testing was performed by iterating through the number of PCA components resulting from extracting the diabetes dataset using PCA. Figure 8 shows that the model's accuracy begins to show consistency when the number of PCA components is 12, with an accuracy value of 97.44. Hence, the number of selected features from PCA used in this research is 12.

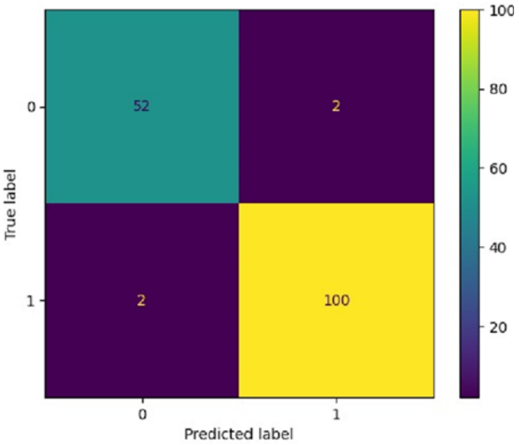

Figure 9. Confusion matrix based on PCA

The model's accuracy is calculated based on the confusion matrix. Figure 9 displays the confusion matrix obtained for evaluating the XGBoost algorithm using a dataset with 12 features selected through PCA feature selection. From the confusion matrix results, values for accuracy, precision, recall, and F1-Score are obtained, as shown in Table 4.

Table 4. Model accuracy using PCA

Accuracy	Precision	Recall	F1-Score
97.44	98.04	98.04	98.04

Discussion

The comparison results of the current study with the previous research are shown in Table 5.

Table 5. Comparation of model accuracy

Method	Number of Feature	Accuracy
XGBoost[24]	16	98.71
XGBoost's Feature Importance[25]	10	98.72
XGBoost[22]	9	80.51
XGBoost[16]	10	70.1
XGboost Using PCA	12	97.44

Previous research on prediabetes prediction using XGBoost machine learning was 80.51 with number of feature 9 and another XGBoost using 10 feature was 70.1, showing lower accuracy compared to current research. In the previous research with the same dataset, used XGBoost[24] with the prediabetes dataset containing 16 features, resulting in a model accuracy of 98.71. The second study utilized XGBoost with feature selection using XGBoost's Feature Importance[25], leading to 10 features and a model accuracy of 98.72. There is an accuracy improvement of 0.01 compared to the first study. The current study utilizes XGBoost with Principal Component Analysis for feature selection, resulting in 12 features and a model accuracy of 97.44. There is a difference in accuracy of 1.28 compared to the previous best accuracy obtained in the earlier research.

The future application of this research is to combine the model into a web or mobile application so that it can be used more widely, for example Personalized Health Monitoring and Mobile Health Apps. Potential improvements from this research are by integrating the dataset used in electronic health records and utilizing other feature selection techniques. Utilization of machine learning techniques such as Long Short-Term Memory (LSTM) which focuses on temporal data such as patient blood sugar levels over time can improve prediction.

CONCLUSION

In this study, XGBoost and PCA are integrated to improve model accuracy by finding features with the best correlation according to PCA. Based on the conducted testing, it can be concluded that PCA can be used for feature selection by reducing dimensionality. PCA preserves most of the data variance in the selected principal components to choose the features that contribute the most to the data variation. Regarding model accuracy, PCA in this research resulted in better accuracy compared to the previous study that used XGBoost and Random Forest. This difference may occur because, in this study, the main features related to prediabetes are considered to have low contributions to the data variation, leading PCA to reduce these features.

REFERENCES

[1] K. Luc, A. Schramm-Luc, T. J. Guzik, and T. P. Mikolajczyk, “Oxidative stress and inflammatory markers in prediabetes and diabetes,” *J. Physiol. Pharmacol.*, vol. 70, no. 6, 2019, doi: 10.26402/jpp.2019.6.01.

[2] I. Owei, N. Umekwe, F. Ceesay, and S. Dagogo-Jack, “Awareness of prediabetes status and subsequent health behavior, body weight, and blood glucose levels,” *J. Am. Board Fam. Med.*, vol. 32, no. 1, pp. 20–27, 2019, doi: 10.3122/jabfm.2019.01.180242.

[3] T. Katangwe, H. Family, J. Sokhi, C. L. Kirkdale, and M. J. Twigg, “The community pharmacy setting for diabetes prevention: A mixed methods study in people with ‘pre-diabetes,’” *Res. Soc. Adm. Pharm.*, vol. 16, no. 8, pp. 1067–1080, 2020, doi: 10.1016/j.sapharm.2019.11.001.

[4] S. A. Amiel, T. Dixon, R. Mann, and K. Jameson, “Hypoglycaemia in Type 2 diabetes,” *Diabet. Med.*, vol. 25, no. 3, pp. 245–254, 2008, doi: 10.1111/j.1464-5491.2007.02341.x.

[5] A. B. Evert *et al.*, “Nutrition therapy for adults with diabetes or prediabetes: A consensus report,” *Diabetes Care*, vol. 42, no. 5, pp. 731–754, 2019, doi: 10.2337/dci19-0014.

[6] R. M. M. Khan, Z. J. Y. Chua, J. C. Tan, Y. Yang, Z. Liao, and Y. Zhao, “From pre-diabetes to diabetes: Diagnosis, treatments and translational research,” *Med.*, vol. 55, no. 9, pp. 1–30, 2019, doi: 10.3390/medicina55090546.

[7] M. K. Ali *et al.*, “Reach and use of diabetes prevention services in the United States, 2016-2017,” *JAMA Netw. Open*, vol. 2, no. 5, pp. 2016–2017, 2019, doi: 10.1001/jamanetworkopen.2019.3160.

[8] J. W. J. Beulens *et al.*, “Risk and management of pre-diabetes,” *Eur. J. Prev. Cardiol.*, vol. 26, no. 2_suppl, pp. 47–54, 2019, doi: 10.1177/2047487319880041.

[9] I. Budiastutik *et al.*, “The effect of Aloe vera on fasting blood glucose levels in pre-diabetes and type 2 diabetes mellitus: A systematic review and meta-analysis,” *J. Pharm. Pharmacogn. Res.*, vol. 10, no. 4, pp. 737–747, 2022, doi: 10.56499/jppres22.1378_10.4.737.

[10] T. Wang *et al.*, “Ideal Cardiovascular Health Metrics and Major Cardiovascular Events in Patients with Prediabetes and Diabetes,” *JAMA Cardiol.*, vol. 4, no. 9, pp. 874–883, 2019, doi: 10.1001/jamacardio.2019.2499.

[11] A. D. Association, “Classification and diagnosis of diabetes,” *Diabetes Care*, vol. 38 Su, 2015, doi: 10.2337/dc15-S005.

[12] U. e. Laila, K. Mahboob, A. W. Khan, F. Khan, and W. Taekeun, “An Ensemble Approach to Predict Early-Stage Diabetes Risk Using Machine Learning: An Empirical Study,” *Sensors*, vol. 22, no. 14, pp. 1–15, 2022, doi: 10.3390/s22145247.

[13] K. D. K. Wardhani and M. Akbar, “Diabetes Risk Prediction Using Extreme Gradient Boosting (XGBoost),” *J. Online Inform. 7(2), 244-250.*, vol. Vol 7.No 2, 2022, doi: 10.15575/join.v7i2.970.

[14] S. Patel, R. Patel, N. Ganatra, and A. Patel, “Predicting a Risk of Diabetes at Early Stage using Machine Learning Approach,” *Turkish J. Comput. Math. Educ.*, vol. 12, no. 10, pp. 5277–5284, 2021.

[15] V. Vaidya and L. K. Vishwamitra Scholar, “Diabetes Detection using Convolutional Neural Network through Feature Sequencing,” *Turkish J. Comput. Math. Educ.*, vol. 12, no. 10, pp. 2783–2789, 2021.

[16] J. J. S. M. Et. al., “Predictive Modeling Framework for Diabetes Classification Using Big Data Tools and Machine Learning,” *Turkish J. Comput. Math. Educ.*, vol. 12, no. 10, pp. 818–823, 2021, doi: 10.17762/turcomat.v12i10.4255.

[17] H. Y. Islam, M. M., Ferdousi, R., Rahman, S., & Bushra, “Likelihood prediction of diabetes at early stage using data mining techniques,” *Comput. Vis. Mach. Intell. Med. image Anal. (pp. 113-125)*, 2020.

[18] S. K. Bhoi *et al.*, “Prediction of Diabetes in Females of Pima Indian Heritage: A Complete Supervised Learning Approach,” *Turkish J. Comput. Math. Educ.*, vol. 12, no. 10, pp. 3074–3084, 2021.

[19] M. Li, X. Fu, and D. Li, “Diabetes Prediction Based on XGBoost Algorithm,” *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 768, no. 7, 2020, doi: 10.1088/1757-899X/768/7/072093.

[20] Y. Tan, H. Chen, J. Zhang, R. Tang, and P. Liu, “Early Risk Prediction of Diabetes Based on GA-Stacking,” *Appl. Sci.*, vol. 12, no. 2, 2022, doi: 10.3390/app12020632.

[21] C. Umam, L. B. Handoko, and F. O. Isinkaye, “Performance Analysis of Support Vector Classification and Random Forest in Phishing Email Classification,” *Sci. J. Informatics*, vol. 11, no. 2, pp. 367–374, 2024, doi: 10.15294/sji.v11i2.3301.

[22] P. Sheth, M. Shah, N. Joisher, and R. Kotecha, “Early Prediction of Pre-Diabetes Using Machine Learning,” *SSRN Electron. J.*, no. January 2020, 2020, doi: 10.2139/ssrn.3572743.

[23] T. Zueger, S. Schallmoser, M. Kraus, M. Saar-Tsechansky, S. Feuerriegel, and C. Stettler, “Machine Learning for Predicting the Risk of Transition from Prediabetes to Diabetes,” *Diabetes Technol. Ther.*, vol. 24, no. 11, pp. 842–847, 2022, doi: 10.1089/dia.2022.0210.

[24] M. Akbar, P. C. Riau, R. Komputer, P. C. Riau, and A. Info, “Diabetes Risk Prediction Using Extreme Gradient Boosting (XGBoost),” vol. 7, no. 2, pp. 244–250, 2022, doi: 10.15575/join.v7i2.970.

[25] Kartina Diah Kusuma Wardani and Memen Akbar, “Diabetes Risk Prediction using Feature Importance Extreme Gradient Boosting (XGBoost),” *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 7, no. 4, pp. 824–831, 2023, doi: 10.29207/resti.v7i4.4651.

[26] S. Kumar, “Efective Hedging Strategy For Us Treasury Bond Portofolio Using Principal Component Analysis,” *Acad. Account. Financ. Stud. J.*, no. January, 2022.

[27] H. Roopa and T. Asha, “A Linear Model Based on Principal Component Analysis for Disease Prediction,” vol. 7, pp. 4–8, 2019.

[28] M. S. Ejaz, M. R. Islam, M. Sifatullah, and A. Sarker, “Implementation of Principal Component Analysis on Masked and Non-masked Face Recognition,” *1st Int. Conf. Adv. Sci. Eng. Robot. Technol. 2019, ICASERT 2019*, vol. 2019, no. Icasert, pp. 1–5, 2019, doi: 10.1109/ICASERT.2019.8934543.

[29] F. M. Basysyar and G. Dwilestari, “House Price Prediction Using Exploratory Data Analysis and Machine Learning with Feature Selection,” *Acadlore Trans. AI Mach. Learn.*, vol. 1, no. 1, pp. 11–21, 2022, doi: 10.56578/ataiml010103.

[30] T. Sarwar *et al.*, “The Secondary Use of Electronic Health Records for Data Mining: Data Characteristics and Challenges,” *ACM Comput. Surv.*, vol. 55, no. 2, 2023, doi: 10.1145/3490234.

[31] A. Nurizki, A. Fitrianto, and A. M. Soleh, “Performance of Ensemble Learning in Diabetic Retinopathy Disease Classification Performance of Ensemble Learning in Diabetic Retinopathy Disease Classification,” *Sci. J. Informatics*, vol. 11, no. May, pp. 375–386, 2024, doi: 10.15294/sji.v11i2.4725.

[32] N. Romadoni, A. M. Siregar, D. S. Kusumaningrum, and T. Rohana, “Classification Model of Public Sentiments About Electric Cars Using Machine Learning,” *Sci. J. Informatics*, vol. 11, no. 2, pp. 303–314, 2024, doi: 10.15294/sji.v11i2.1309.

[33] M. Heydarian and T. E. Doyle, “MLCM : Multi-Label Confusion Matrix,” pp. 19083–19095, 2022.

[34] J. Heland *et al.*, “Analysis of Performance of Classification Algorithms in Mushroom Poisonous Detection using Confusion Matrix Analysis,” vol. 9, no. 1, 2020.

[35] N. Architecture, “Multi-lable Classifier Performance Evaluation With Confusion Matrix,” pp. 1–14, 2020, doi: 10.5121/csit.2020.100801.