Computer Science and Information Technologies
Vol. 6, No. 2, July 2025, pp. 102~111
ISSN: 2722-3221, DOI: 10.11591/csit.v6i2.p102-111

An ensemble learning approach for diabetes prediction using the stacking method

Elliot Kojo Attipoe, Alimatu-Saadia Yussiff, Maame Gyamfua Asante-Mensah, Emmanuel Dortey Tetteh, Regina Esi Turkson
Department of Computer Science and Information Technology, Faculty of Physical Sciences, University of Cape Coast, Cape Coast, Ghana

Article Info

Article history:

Received Feb 21, 2025
Revised May 8, 2025
Accepted May 23, 2025

Keywords:

Diabetes prediction
Ensemble methods
K-nearest neighbor
Logistic regression
Support vector machine

ABSTRACT

Diabetes is a severe illness characterized by high blood glucose levels. Machine learning algorithms, with their ability to detect and predict diabetes in its early stages, offer a promising avenue for research. This study sought to enhance the accuracy of predicting diabetes mellitus by employing the stacking method. The stacking method was chosen because it integrates predictions from various base models, resulting in a more precise final prediction. The stacking method enhances accuracy and generalization by utilizing the varied strengths of multiple base models. The Pima Indians diabetes dataset, a widely used benchmark dataset, was utilized in the study. The machine learning models used for the studies were logistic regression (LR), naïve Bayes (NB), extreme gradient boost (XGBoost), K-nearest neighbor (KNN), decision tree (DT), and support vector machine (SVM). LR, KNN, and SVM were the best-performing models based on accuracy, F1-score, precision, and area under the curve (AUC) score, and were consequently used as the base model for the stacking method. The LR model was utilized for the meta-model. The proposed ensemble approach using the stacking method demonstrated a high accuracy of 82.4%, better than the individual models and other ensemble techniques such as bagging or boosting. This study advances diabetes prediction by developing a more accurate early-stage detection model, thereby improving clinical management of the disease.

Corresponding Author:

Elliot Kojo Attipoe
Department of Computer Science and Information Technology, University of Cape Coast
Cape Coast, Ghana
Email: eattipoe@ucc.edu.gh

1. INTRODUCTION

Diabetes mellitus, a chronic metabolic disorder, presents significant global health challenges, affecting millions worldwide [1]. Characterized by increased blood glucose levels resulting from inadequate insulin production or the body's ineffective use of insulin, diabetes can lead to severe complications, including kidney failure, blindness, and cardiovascular diseases [2]. The urgency of early and accurate diabetes prediction cannot be overstated. It is a crucial and immediate step towards effective intervention and management, potentially reducing health risks and mortality rates. Our study, with its potential to contribute to the development of improved predictive models, holds promise for the future of diabetes management.

The emergence of machine learning has transformed healthcare research, introducing innovative methodologies and a wide range of applications. Our study, driven by the potential of machine learning, takes a step further in this innovative trend, proposing an ensemble machine learning approach for predicting diabetes.

This innovative approach, breaking away from conventional diabetes prediction models, holds promise for the future of diabetes management, sparking intrigue and engagement in the field.

Machine learning approaches have gained attention as viable options for disease predictions because it can process extensive datasets and identify complex patterns [3]. However, it is essential to note that no machine learning model can be considered universally optimum for all prediction tasks. This highlights the necessity of employing an ensemble method combining multiple models to achieve superior prediction performance. Harnessing the power of an ensemble machine learning methods boosts prediction accuracy and enhances the model's robustness and generalizability [4]. Ensemble approaches provide a comprehensive approach to diabetes prediction research by utilizing the strengths of individual algorithms and limiting their weakness [5]. Despite the significant progress in diabetes prediction using ensemble techniques such as stacking, bagging, boosting, and soft voting, different studies lack consistent performance metrics. Most research focuses on model accuracy, with limited attention to real-world deployment and handling imbalanced datasets. This study fills these gaps by creating a more reliable, interpretable, and generalizable ensemble machine learning framework for diabetes prediction, incorporating feature selection techniques and advanced model evaluation metrics.

Several studies have used various datasets, different combinations of algorithms, other types of ensemble methods, and different prediction metrics. In their work, Liu *et al.* [5] proposed an early diabetes prediction model using a stacking ensemble learning approach, integrating gradient boosting decision tree (DT), AdaBoost, random forest (RF), and logistic regression (LR). Their model exhibited enhanced predictive performance relative to individual machine learning models, with better accuracy and recall rates. By selecting key early symptoms such as polyuria, polydipsia, and sudden weight loss, they could effectively enhance the early detection of diabetes, highlighting the advantages of ensemble methods in medical diagnosis applications. In a similar study, Dutta *et al.* [6] proposed an ensemble machine learning approach for early diabetes prediction using a newly labeled dataset from Bangladesh. Their model combined naive Bayes (NB), RF, DT, extreme gradient boost (XGBoost), and LightGBM classifiers, achieving an accuracy of 73.5% and an area under the curve (AUC) of 0.832. Utilizing extensive data preprocessing, including missing value imputation, feature selection, and hyperparameter optimization, the study demonstrated the advantages of ensemble models in enhancing predictive performance for early diabetes detection. This work highlights the importance of robust preprocessing techniques and multiple classifiers in medical prediction tasks.

Ganie and Malik [7] used the diabetes dataset on the University of California repository for their research. The authors used an ensemble learning-based framework utilizing techniques like bagging, boosting, and voting. Ultimately, the bagged DT was the most effective classifier, with an accuracy of 99.41%. In another study, Gourisaria *et al.* [8] utilized the diabetes datasets from Frankfurt Hospital in Germany. The authors used 14 machine learning classification algorithms to develop a predictive model for diabetes. Utilizing the soft voting method, the top five performing algorithms were used to create the classifier. The proposed ensemble classifier achieved an accuracy of 97.3%.

In a related work, Jain [9] implemented a model based on three algorithms using the Indian diabetes dataset. The algorithms were evaluated based on accuracy, and the RF model was the most effective. The RF model was recommended for diabetes prediction over the LR and K nearest neighbor (KNN). Using the same Pima Indian diabetes dataset (PIDD), Charitha *et al.* [10] introduced a comprehensive framework for diabetes prediction employing various machine learning classifiers, including KNN, DT, RF, AdaBoost, NB, XGBoost, and multilayer perceptron (MLP). The authors utilized a weighted ensemble of diverse machine learning models to improve the precision of diabetes prediction. The suggested ensemble classifier performed superior to prior publications, with a 2.00% increase in the AUC score.

Abnoosian *et al.* [11] proposed an ensemble machine learning model for predicting diabetes using a multi-classification framework based on an imbalanced dataset of Iraqi patients. Their approach incorporated various preprocessing techniques, such as missing value imputation and feature selection, along with a combination of models, including KNN, support vector machine (SVM), DT, and RF. The ensemble method achieved impressive performance, with an accuracy of 98.87% and an AUC of 0.999, demonstrating the effectiveness of ensemble learning in managing dataset imbalance and improving diabetes prediction accuracy. Utilizing a hybrid stacked ensemble approach with genetic algorithms [12] proposed a new model for diabetes prediction. Their method integrated multiple machine learning models, such as RF, SVM, and neural networks, to improve diagnostic accuracy. By using genetic algorithms for feature selection, they achieved an accuracy of 98.8% and 99% on two different diabetes datasets, significantly outperforming other individual machine learning methods. This study highlights the effectiveness of combining ensemble learning with genetic algorithms to enhance predictive medical diagnosis performance.

In another study, Singh and Gupta [13] applied an ensemble learning technique to the Indian diabetics' dataset, utilizing five models: RF, light gradient boost (LG Boost), XGBoost, gradient boost, and AdaBoost. The bagging and boosting ensemble method is employed to classify patients as diabetic or non-diabetic. Although the authors did not explicitly state the performance criterion used to evaluate their model, they did

suggest that recording outperforms the current methods. In their work, Kumari *et al.* [3] employed an ensemble learning technique utilizing a soft voting classifier. The proposed model employed an ensemble of three algorithms for classification, specifically RF, LR, and NB. The suggested ensemble strategy achieves the highest accuracy, precision, recall, and F1-score levels, with values of 79.04%, 73.48%, 71.45%, and 80.6%, respectively.

In a more recent study, Fahim *et al.* [14] used machine learning algorithms to predict the possibility of diabetes in women. Utilizing the Indian diabetes dataset, the authors applied the hard voting technique to the following algorithms: XGBoost, KNN, and RF. According to their classification reports, the precision, recall, and F1-scores were perfect, with an accuracy of 86%. Oliullah *et al.* [15] also proposed a stacked ensemble machine learning model for diabetes prediction, incorporating classifiers such as RF, XGBoost, natural gradient boosting (NGBoost), AdaBoost, and LightGBM. Through feature engineering and data preprocessing, their proposed model achieved a high accuracy of 92.91%, significantly improving baseline models. The authors also employed Shapley additive explanations (SHAP) to interpret the model, identifying insulin and glucose levels as key predictors. Their findings highlight the effectiveness of ensemble learning in enhancing prediction accuracy and model transparency for early diabetes detection.

The stacking ensemble machine learning method in diabetes prediction is dynamic and rapidly evolving. Recent research underscores the potential of this method, emphasizing its accuracy, versatility, interpretability, and real-time application. By harnessing the power of many algorithms, this ensemble technique offers a potential approach to enhance diabetes prediction. Utilizing the PIMA dataset, multiple authors have validated the effectiveness of ensemble methods and introduced a novel algorithm combination that other researchers can further explore. This study addresses the gaps in diabetes prediction research by demonstrating the effectiveness of ensemble machine learning techniques, specifically stacking, over individual classifiers. Previous models that utilized classifiers such as RF and NB underperformed in key metrics like precision, recall, and accuracy when applied to the PIMA diabetes dataset. This study, therefore, aims to: i) develop a new stacking ensemble machine learning model for diabetes prediction using LR, KNN, and support vector classifier; and ii) compare the proposed ensemble approach's performance against traditional machine learning models regarding accuracy, precision, F1-score, and AUC-ROC.

2. METHOD

The research sought to enhance the performance of current machine learning models in predicting diabetes. We have presented an ensemble machine learning approach utilizing the stacking technique. The stacking method was chosen above alternative approaches because of its ability to combine the predictions of several base models, resulting in a more precise final prediction. The stacking strategy enhances accuracy and generalization by utilizing the varied strengths of distinct base models. Figure 1 illustrates the framework used for the study.

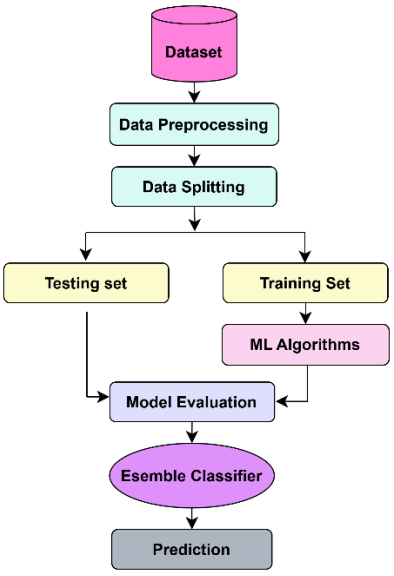

Figure 1. Proposed framework for the study

The proposed diabetic predictive model utilized a series of steps, including gathering data, preparing the data, dividing the data into subsets, employing several machine learning classification methods, and using models for the ensemble model. Seven classification machine learning algorithms were trained on the dataset to achieve this goal. The descriptions of how the steps were employed in this research are presented in sub-sections 2.1 to 2.4.

2.1. Dataset collection

The study used the PIMA Indians diabetes dataset from the Kaggle repository. The data comprises 768 instances with eight features. Table 1 describes the attributes of the datasets.

Table 1. Attributes of the PIMA diabetes dataset

Attribute	Description
Pregnancies	Number of pregnancies
Glucose	Plasma glucose concentration at 2 hours in an oral glucose tolerance test
blood pressure	Diastolic blood pressure (mm Hg)
Skin thickness	Triceps skin fold thickness (mm)
Insulin	2-Hour serum insulin (µ/ml)
BMI	Body mass index (weight in kg/(height in m)^2)
Diabetes pedigree function	Diabetes pedigree function
Age	Age (years)
Target	Class variable (0 or 1) 268 of 768 are 1; the others are 0

2.2. Data preprocessing

Data preparation is an important step in machine learning because it converts data into a format suitable for input into machine learning algorithms. Data preprocessing involves several techniques, such as cleansing, normalization, addressing missing values, label encoding, and dataset partitioning. The data cleaning process eliminates extraneous disruptions and inconsistencies, hence improving data quality. To achieve precise and efficient outcomes, it is imperative to eliminate data containing irrelevant information and replace any missing values. The pregnancy attribute was less relevant to the study, so it was dropped. We further explored the relevance of the remaining attributes by checking their correlation values. Figure 2 shows a heatmap distribution for the dataset. The blood pressure and skin thickness attributes were subsequently removed from the dataset due to their correlation values.

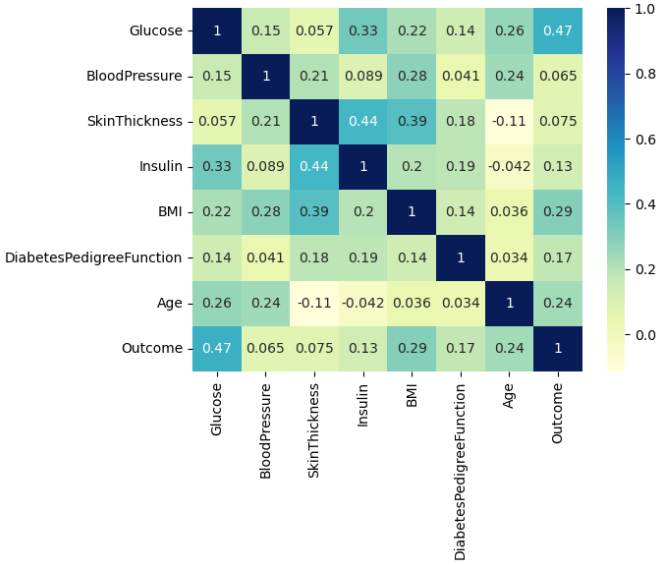

Figure 2. Heatmap of the dataset

2.3. Machine learning classification methods

Seven machine learning classification algorithms were applied to the diabetes dataset to train and build the model. The classification algorithms considered for this study were LR, NB, RF classifier, XGBoost, KNN, DT, and SVM.

2.3.1. Logistic regression

LR is a statistical and machine learning method for solving classification problems. This method is employed when the dependent variable is categorical [16]. Binary classification is the most prevalent use case, where the output is limited to either 0 or 1.

2.3.2. Naïve Bayes

NB is a machine learning classifier that uses Bayes' theorem to calculate probabilities [17]. It is used for classification tasks. The term "naïve" is used because it presupposes that each feature is independent of the others, given the class variable. Although independence is frequently violated in real-world datasets, NB classifiers exhibit excellent performance in numerous contexts.

2.3.3. Random forest

RF is a widely used ensemble learning technique for classification and regression tasks. As an ensemble method, it amalgamates numerous separate models to generate a more resilient and precise composite model [18]. An RF algorithm constructs many DT and combines their results.

2.3.4. Extreme gradient boost

The XGBoost algorithm aims to provide a remarkably efficient, adaptable, and portable implementation of gradient boosting techniques [19]. It has gained significant popularity due to its fast and efficient performance, especially in jobs involving classification and regression. The process commences with a weak prediction, such as the average of the goal values. Subsequently, each successive tree predicts the residual from the previous prediction. Each tree is constructed to correct the faults of its predecessor, and the trees are sequentially integrated, with each tree addressing the residual errors.

2.3.5. K-nearest neighbor

The KNN algorithm is a simple and intuitive non-parametric method for classification and regression tasks. KNN forecasts the result for a new data point by locating the 'k' nearest training instances according to a selected distance metric [17]. The algorithm uses the majority voting among neighbors or weighted averages of their values.

2.3.6. Decision tree

A DT is a model constructed like a flowchart with a tree-like structure. It is used for tasks involving classification and regression. The process consists of partitioning datasets into smaller groups according to the values of input features, leading to the creation of a DT model [20]. The dataset is divided into subsets repeatedly, using the feature that provides the highest increase in information or the lowest level of uncertainty. This process continues until a specified stopping requirement, such as reaching a specific tree depth, is satisfied. The terminal nodes correspond to the ultimate predictions.

2.3.7. Support vector machine

The SVM is a resilient and adaptable method for classification and regression tasks. It is particularly suitable at classifying complex datasets with clear delineating boundaries. The SVM method seeks to identify the best hyperplane that optimizes the separation between two classes.

2.4. Implementation of the stacking techniques

The machine learning algorithms were implemented to automatically identify observations that reveal the presence or absence of diabetes disease. One key objective of the study was to use the top-performing algorithms to implement the stacking method. The first stage was to evaluate the performance of models and choose the best-performing models as the base model for the stacking ensemble classifier. Figure 3 illustrates the proposed ensemble method to be employed. In this method, the results from the base models would be passed to a meta-model for the final prediction.

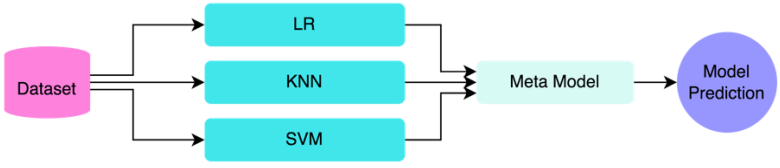

Figure 3. Proposed ensemble method

3. RESULTS AND DISCUSSION

In this study, multiple machines learning classifiers, including LR, NB, RF, XGBoost, KNN, DT, and SVM, were employed to predict the likelihood of diabetes. The dataset comprises 768 samples, with 268 individuals diagnosed with diabetes and 500 without the condition, distributed across eight unique features. Various performance metrics, such as accuracy, precision, recall, F1-score, and area under the curve (AUC), were utilized to evaluate the efficacy of these models. The results are presented in two parts; the performance of individual algorithms and the performance of the stacking method.

3.1. Performance of individual algorithms

The performance of the algorithms based on accuracy, precision, F1-score, and AUC scores are presented in Figures 4 to 7. The results indicated that LR, KNN, and SVM classifiers were the best-performing algorithms. Therefore, used as the base model for the proposed stacking method.

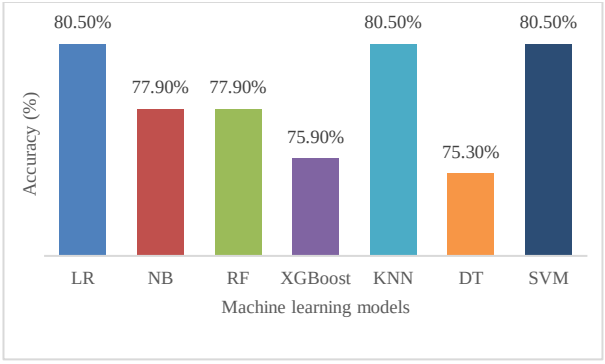

Figure 4. Comparison of the accuracy of each model

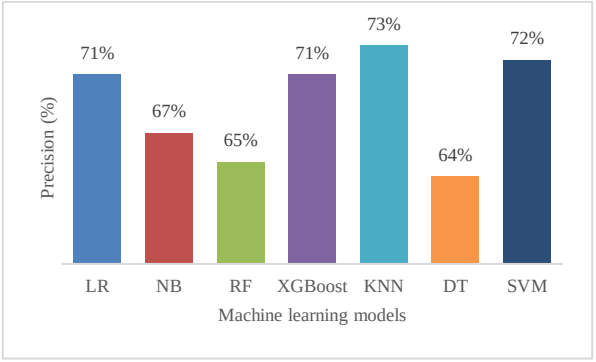

Figure 5. Comparison of the precision of each model

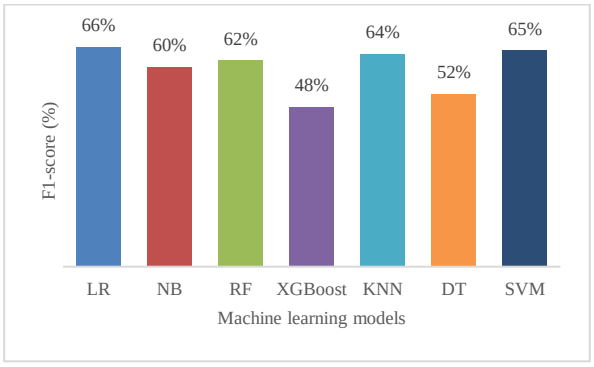

Figure 6. Comparison of the F1-score of each model

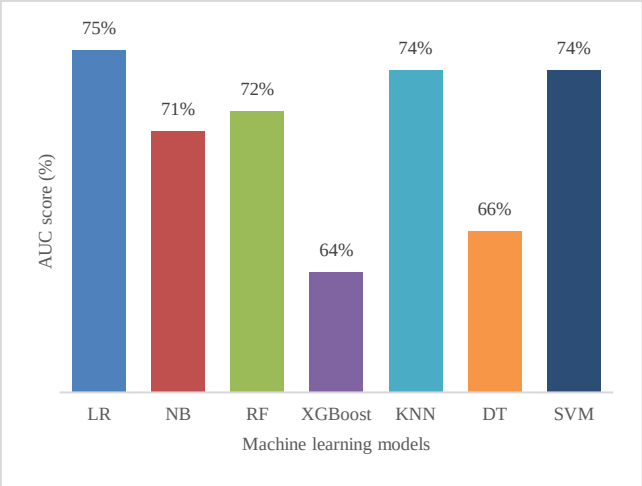

Figure 7. Comparison of the AUC score of each model

3.2. Performance of the stacking method

The results from the analysis indicate that, stacked ensemble method performed better than individual classifiers. The proposed model demonstrated an accuracy of 82.4%, outperforming individual classifiers, as shown in Table 2. The model achieved a precision of 78%, a recall of 60%, an F1-score of 67%, and an AUC score of 76%. These metrics illustrate the effectiveness of the ensemble approach in handling the classification task, particularly in balancing precision and recall.

Table 2. Model evaluation

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	AUC score (%)
LR	80.5	71	62	66	75
KNN	80.5	73	57	64	74
SVM	80.5	72	60	65	74
Ensemble classifier	82.4	78	60	67	76

The stacking method offers a unique advantage over other ensemble techniques like bagging, boosting, and voting because of its capacity to learn the best combination of predictions from many models adaptively [21], [22]. While techniques like bagging and boosting aim to reduce variation through consistent combination strategies, the stacking method employs a meta-model trained to enhance the final prediction by utilizing the outputs of base models. The meta-model prioritizes more dependable outputs, resulting in improved performance and generality of the final prediction compared to previous ensemble approaches [23], [24].

The proposed model, a combination of LR, KNN, and SVM, yielded higher accuracy than previous models, as shown in Table 3. For example, Priya *et al.* [25] achieved an accuracy of 81% using gradient boosting, RF, and DT. In contrast, Tasin *et al.* [26] reported an accuracy of 81% with a broader range of classifiers, including DT, SVM, RF, LR, and KNN.

Table 3. Performance comparison with existing models

Authors	Models	Accuracy (%)
Kumari *et al.* [3]	RF, LR, NB	79.04
Dutta *et al.* [6]	NB, RF, DT, XGBoost, LightGBM	73.50
Priya *et al.* [25]	Gradient boosting, RF, DT	81
Tasin *et al.* [26]	DT, SVM, RF, LR, KNN	81
Proposed model	LR, KNN, SVM	82.4

The superior performance of our model can be attributed to the stacking method, which combines the strengths of several classifiers to improve the overall prediction accuracy. Compared to standalone models, ensemble methods generally provide a more robust solution by minimizing the weaknesses inherent in individual models. As shown in Table 2, the proposed ensemble approach surpasses the performance of

previous models in accuracy, precision, and recall, which are critical in medical diagnosis, where the cost of false positives and false negatives can be high.

4. CONCLUSION

Diabetes mellitus still presents significant global health challenges affecting millions worldwide. It is estimated to cause about two million deaths annually. However, early detection leads to effective intervention and management practices, reducing the high risk of mortality rates. Using the PIMA diabetes dataset, this paper presented an ensemble classifier using the stacking method to predict diabetes. The classifier uses LR, KNN, and a SVM and achieves accuracy, precision, recall, F1-score, and AUC scores of 82.4%, 78%, 60%, 67%, and 76%, respectively. This research supports the notion that combining classifiers can produce better outcomes than using a single model, offering promising insights for applying machine learning to healthcare data. The result also indicated that the stacking method improves the final prediction's performance and generality compared to previous ensemble approaches. Future studies could explore other base models to improve the overall performance metrics, especially recall and F1-scores, which remain relatively lower. Additionally, incorporating feature engineering techniques and dimensionality reduction methods, such as principal component analysis, could refine the input data and improve classifier performance. Finally, expanding the study to include more extensive, diverse datasets or real-world clinical data could enhance the model's generalizability across different populations and healthcare settings, offering more profound insights into global diabetes management.

FUNDING INFORMATION

No funding involved.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Elliot Kojo Attipoe	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓			
Alimatu-Saadia Yussiff		✓				✓		✓	✓	✓		✓	✓	
Maame Gyamfua Asante-Mensah	✓		✓	✓			✓			✓	✓		✓	
Emmanuel Dortey Tetteh		✓		✓			✓	✓	✓			✓		
Regina Esi Turkson	✓		✓		✓	✓	✓			✓	✓			

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

No conflict of interest.

DATA AVAILABILITY

The data that support the findings of this study are openly available in [Kaggle] at https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database, reference number [16].

REFERENCES

[1] WHO, "Diabetes," *World Health Organization*. 2024. Accessed: Jun. 03, 2024. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/diabetes
[2] American Diabetes Association, "2. Classification and diagnosis of diabetes: standards of medical care in diabetes-2021," *Diabetes Care*, vol. 44, no. 1, pp. S15–S33, 2021, doi: 10.2337/dc21-S002.
[3] S. Kumari, D. Kumar, and M. Mittal, "An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier," *International Journal of Cognitive Computing in Engineering*, vol. 2, pp. 40–46, 2021, doi: 10.1016/j.ijcce.2021.01.001.
[4] P. Rani, R. Lamba, R. K. Sachdeva, P. Bathla, and A. N. Aledaily, "Diabetes prediction using machine learning classification algorithms," in *2023 International Conference on Smart Computing and Application (ICSCA)*, 2023, pp. 1–5, doi: 10.1109/ICSCA57840.2023.10087827.

[5] J. Liu, L. Fan, Q. Jia, L. Wen, and C. Shi, "Early diabetes prediction based on stacking ensemble learning model," in *2021 33rd Chinese Control and Decision Conference (CCDC)*, 2021, pp. 2687–2692, doi: 10.1109/CCDC52312.2021.9601932.
[6] A. Dutta *et al.*, "Early prediction of diabetes using an ensemble of machine learning models," *International Journal of Environmental Research and Public Health*, vol. 19, no. 19, 2022, doi: 10.3390/ijerph191912378.
[7] S. M. Ganie and M. B. Malik, "An ensemble machine learning approach for predicting type-II diabetes mellitus based on lifestyle indicators," *Healthcare Analytics*, vol. 2, 2022, doi: 10.1016/j.health.2022.100092.
[8] M. K. Gourisaria, G. Jee, G. M. Harshvardhan, V. Singh, P. K. Singh, and T. C. Workneh, "Data science appositeness in diabetes mellitus diagnosis for healthcare systems of developing nations," *IET Communications*, vol. 16, no. 5, pp. 532–547, 2022, doi: 10.1049/cmu2.12338.
[9] V. Jain, "Performance analysis of supervised machine learning algorithm for prediction of diabetes," in *2022 International Conference on Edge Computing and Applications (ICECAA)*, 2022, pp. 1162–1165, doi: 10.1109/ICECAA55415.2022.9936503.
[10] C. Charitha, A. Devi Chaitrasree, P. C. Varma, and C. Lakshmi, "Type-II diabetes prediction using machine learning algorithms," in *2022 International Conference on Computer Communication and Informatics (ICCCI)*, 2022, pp. 1–5, doi: 10.1109/ICCCI54379.2022.9740844.
[11] K. Abnoosian, R. Farnoosh, and M. H. Behzadi, "Prediction of diabetes disease using an ensemble of machine learning multi-classifier models," *BMC Bioinformatics*, vol. 24, no. 1, 2023, doi: 10.1186/s12859-023-05465-z.
[12] J. Abdollahi and B. Nouri-Moghaddam, "Hybrid stacked ensemble combined with genetic algorithms for diabetes prediction," *Iran Journal of Computer Science*, vol. 5, no. 3, pp. 205–220, 2022, doi: 10.1007/s42044-022-00100-1.
[13] S. Singh and S. Gupta, "Prediction of diabetes using ensemble learning model," in *Machine Intelligence and Soft Computing*, Singapore: Springer, 2021, pp. 39–59, doi: 10.1007/978-981-15-9516-5_4.
[14] F. Fahim, M. T. Ahmed, M. N. M. Shuvo, and M. R. Islam, "A comparison between different kernels of support vector machine to predict cardiovascular diseases using phonocardiogram signal," in *2022 Second International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT)*, 2022, pp. 1–4, doi: 10.1109/ICAECT54875.2022.9808063.
[15] K. Oliullah, M. H. Rasel, M. M. Islam, M. R. Islam, M. A. H. Wadud, and M. Whaiduzzaman, "A stacked ensemble machine learning approach for the prediction of diabetes," *Journal of Diabetes and Metabolic Disorders*, vol. 23, no. 1, pp. 603–617, 2024, doi: 10.1007/s40200-023-01321-2.
[16] M. Martínez-García, I. Inza, and J. A. Lozano, "Learning a logistic regression with the help of unknown features at prediction stage," in *2023 IEEE Conference on Artificial Intelligence (CAI)*, 2023, pp. 298–299, doi: 10.1109/CAI54212.2023.00133.
[17] M. R. Romadhon and F. Kurniawan, "A comparison of naive Bayes methods, logistic regression, and KNN for predicting healing of covid-19 patients in Indonesia," in *2021 3rd East Indonesia Conference on Computer and Information Technology (EIConCIT)*, 2021, pp. 41–44, doi: 10.1109/EIConCIT50028.2021.9431845.
[18] V. K. G. Kalaiselvi, H. Shanmugasundaram, E. Aishwarya, M. Ragavi, C. Nandhini, and S. J. Bhuvaneshwari, "Analysis of Pima Indian diabetes using KNN classifier and support vector machine technique," in *2022 Third International Conference on Intelligent Computing Instrumentation and Control Technologies (ICICICT)*, 2022, pp. 1376–1380, doi: 10.1109/ICICICT54557.2022.9917992.
[19] V. S. Narayana, L. Chennagiri, B. D. P. Kumar, S. K. R. Mallidi, and T. S. R. Sai, "Prediction of COVID-19 victim's well-being using extreme gradient boost algorithm," in *2023 2nd International Conference on Edge Computing and Applications (ICECAA)*, 2023, pp. 958–963, doi: 10.1109/ICECAA58104.2023.10212406.
[20] F. Aaboub, H. Chamlal, and T. Ouaderhman, "Analysis of the prediction performance of decision tree-based algorithms," in *2023 International Conference on Decision Aid Sciences and Applications (DASA)*, 2023, pp. 7–11, doi: 10.1109/DASA59624.2023.10286809.
[21] X. Dong, Z. Yu, W. Cao, Y. Shi, and Q. Ma, "A survey on ensemble learning," *Frontiers of Computer Science*, vol. 14, no. 2, pp. 241–258, 2020, doi: 10.1007/s11704-019-8208-z.
[22] C. Cai *et al.*, "Using ensemble of ensemble machine learning methods to predict outcomes of cardiac resynchronization," *Journal of Cardiovascular Electrophysiology*, vol. 32, no. 9, pp. 2504–2514, 2021, doi: 10.1111/jce.15171.
[23] S. Asif, Y. Wenhui, Y. Tao, S. Jinhai, and H. Jin, "An ensemble machine learning method for the prediction of heart disease," in *2021 4th International Conference on Artificial Intelligence and Big Data (ICAIBD)*, 2021, pp. 98–103, doi: 10.1109/ICAIBD51990.2021.9459010.
[24] C. A. T. Stevens *et al.*, "Ensemble machine learning methods in screening electronic health records: A scoping review," *Digital Health*, vol. 9, 2023, doi: 10.1177/20552076231173225.
[25] B. K. Priya, V. S. A. K. Tanniru, and M. Katamaneni, "Ensemble learning model for diabetes prediction," in *2023 International Conference on Innovative Data Communication Technologies and Application (ICIDCA)*, 2023, pp. 33–36, doi: 10.1109/ICIDCA56705.2023.10099617.
[26] I. Tasin, T. U. Nabil, S. Islam, and R. Khan, "Diabetes prediction using machine learning and explainable AI techniques," *Healthcare Technology Letters*, vol. 10, no. 1–2, pp. 1–10, 2023, doi: 10.1049/htl2.12039.

BIOGRAPHIES OF AUTHORS

Elliot Kojo Attipoe received a master's degree in Information Management and Technology from Cranfield University, UK, in 2011. He is a lecturer at the Department of Computer Science and Information Technology at the University of Cape Coast, Ghana. His main research interests are machine learning, data analytics, software engineering, and programming languages. He can be contacted at email: eattipoe@ucc.edu.gh.

Alimatu-Saadia Yussiff received Ph.D. in Information Technology from the Universiti Teknologi Petronas, Malaysia, in 2016. She is a Senior Lecturer in the Department of Computer Science and Information Technology at the University of Cape Coast. Her research interests are human-computer interaction, internet and web technologies, e-learning, and software engineering. She is a dedicated, professional, and accomplished lecturer with knowledge of teaching computer science and information technology courses. She can be contacted at email: asyussiff@ucc.edu.gh.

Maame Gyamfua Asante-Mensah is a researcher and academic specializing in computational science and biomedical signal processing. She earned her Ph.D. in Computational and Data Science and Engineering from the Skolkovo Institute of Science and Technology (Skoltech), Moscow, Russia. Her primary research interests encompass deep learning, tensor completion, and randomized algorithms, with applications in neuroscience, image denoising, and biomedical signal processing. She currently serves as a lecturer in the Department of Computer Science and Information Technology at the University of Cape Coast, Ghana. She can be contacted at email: gasante-mensah@ucc.edu.gh.

Emmanuel Dortey Tetteh is a lecturer at the Department of Computer Science and Information Technology, University of Cape Coast, Ghana. He obtained Ph.D. in Information and Communication Engineering from the University of Electronic Science and Technology of China. His research interests include software engineering, computer networking, and information systems. He can be contacted at email: etetteh@ucc.edu.gh.

Regina Esi Turkson received her Ph.D. of Engineering in Computer Science and Technology at the University of Electronic Science and Technology of China (UESTC); she is currently a lecturer/researcher at the Department of Computer Science and Information Technology, University of Cape Coast, Ghana. Her research interests include machine learning, artificial intelligence, computational intelligence, federative learning, computer security, and cryptography. She can be contacted at email: rturkson@ucc.edu.gh or regina.turkson@ucc.edu.gh.