International Journal of Electronic Healthcare

ISSN online: 1741-8461 - ISSN print: 1741-8453
https://www.inderscience.com/ijeh

Predicting diabetes using Cohen's Kappa blending ensemble learning

Isaac Kofi Nti, Owusu Nyarko-Boateng, Adebayo Felix Adekoya, Benjamin Asubam Weyori, Henrietta Pokuaa Adjei

DOI: 10.1504/IJEH.2023.10052670

Article History:

Received:	07 February 2022
Accepted:	28 October 2022
Published online:	27 January 2023

Predicting diabetes using Cohen's Kappa blending ensemble learning

Isaac Kofi Nti*

Department of Computer Science and Informatics,
University of Energy and Natural Resources,
Sunyani, Ghana
and
School of Information Technology,
University of Cincinnati,
OH, USA
Email: ntious1@gmail.com
*Corresponding author

Owusu Nyarko-Boateng,
Adebayo Felix Adekoya and
Benjamin Asubam Weyori

Department of Computer Science and Informatics,
University of Energy and Natural Resources,
Sunyani, Ghana
Email: owusu.nyarko-boateng@uenr.edu.gh
Email: adebayo.adekoya@uenr.edu.gh
Email: benjamin.weyori@uenr.edu.gh

Henrietta Pokuaa Adjei

Department of Computer Science,
Sunyani Technical University,
Sunyani, Ghana
Email: herttydonkor@gmail.com

Abstract: Diabetes is a well-known risk factor for early mortality and disability. As signatories to the 2030 Agenda for Sustainable Development, Member States set an ambitious objective of a one-third reduction in early death due to non-communicable diseases (NCDs), which includes diabetes. Nonetheless, the current economic impact of diabetes on countries, individuals, and healthcare requires an agent means of its early detection. However, early detection of diabetes with conventional techniques is a considerable challenge for the healthcare industry and physicians. This study proposed a blended ensemble predictive model with Cohen's Kappa correlation-based base-learners selection to decrease unnecessary diabetes-related mortality through early detection. The empirical outcome shows that our proposed predictive model outperformed existing state-of-the-art approaches for predicting diabetes, thus resulting in enhanced diabetes prediction ability.

Keywords: diabetes; blended ensemble; diabetes prediction; Kappa statistic.

Reference to this paper should be made as follows: Nti, I.K., Nyarko-Boateng, O., Adekoya, A.F., Weyori, B.A. and Adjei, H.P. (2023) 'Predicting diabetes using Cohen's Kappa blending ensemble learning', *Int. J. Electronic Healthcare*, Vol. 13, No. 1, pp.57–70.

Biographical notes: Isaac Kofi Nti holds HND in Electrical and Electronic Engineering from Sunyani Technical University in 2007, a BSc in Computer Science from Catholic University College in 2011, an MSc in Information Technology from Kwame Nkrumah University of Science and Technology in 2016, and a PhD in Computer Science from the University of Energy and Natural Resources (UENR) in 2022. He has over 45 publications in highly peer-reviewed journals indexed on the web of science and Scopus. Currently, he is a Lecturer with the Department of Computer Science and Informatics, University of Energy and Natural Resources, Sunyani, Ghana, and a visiting Professor with the School of Information Technology, University of Cincinnati, Ohio, USA. His research interests include artificial intelligence, health informatics, energy system modelling, cybersecurity intelligent information systems and social and sustainable computing, business analytics, and data privacy and security.

Owusu Nyarko-Boateng holds a Higher National Diploma in Electrical and Electronic Engineering from Accra Technical University. He studied BSc Computer Science and PGDE at Catholic University College, Sunyani-Ghana, MSc Information Technology at Kwame Nkrumah University of Science and Technology, Kumasi-Ghana, and is currently enrolled as PhD Computer Science student at the University of Energy and Natural Resources, Sunyani-Ghana. He holds CISCO Fundamentals of Fibre Optics Technology (FFOT) and CISCO Optical Technology Advanced (OPT300) v2.0 certificates. His research interests are optical technology, submarine and underground fibre optics cable transmission, MIMO, WiMAX, spread spectrum technologies, 5G, data communication, cognitive science, intelligent transmission systems, deploying machine learning in tracing faults in network infrastructure, wireless sensor networks and telecommunication policy development.

Adebayo Felix Adekoya holds a BSc (1994), MSc (2002), and PhD (2010) in Computer Science, an MBA in Accounting and Finance (1998), and a Postgraduate Diploma in Teacher Education (2004). In addition, he has put in about 28 years of experience as a lecturer, researcher, and administrator at the higher educational institution levels in Nigeria and Ghana. He is an Associate Professor of Computer Science with the Department of Computer Science and Informatics, The University of Energy and Natural Resources, Sunyani, Ghana. His research interests include artificial intelligence, business and knowledge engineering, intelligent information systems, and social and sustainable computing.

Benjamin Asubam Weyori received his PhD and MPhil in Computer Engineering from the Kwame Nkrumah University of Science and Technology (KNUST), Ghana in 2016 and 2011, respectively. He obtained his Bachelor of Science in Computer Science from the University for Development Studies (UDS), Tamale, Ghana in 2006. He is currently a Senior Lecturer and the Acting Director of the distance education and online learning at the University of Energy and Natural Resources (UENR) in Ghana. His main research interest includes artificial intelligence, computer visions (image processing), machine learning and web engineering.

Henrietta Pokuaa Adjei holds an 'A' 3-year Post-Secondary in Education from St. Joseph College of Education, a BEd in Information Technology from the University of Education, Winneba in 2012, and an MPhil in Computer Science from University of Energy and Natural Resources (UENR) in 2021. She has three publications to her credit. Currently, she is a Lecturer with the department of computer science, Sunyani Technical University, Sunyani, Ghana. Her research interests include artificial intelligence and machine learning, human-computer interaction and information visualisation, computer vision and graphics.

1 Introduction

The first Word Health Organisation (WHO) Global Report on diabetes was released on 7 April 2016. Even though diabetes has been documented in olden texts and regarded as a dangerous ailment, it does not appear to have been frequently encountered by physicians or healers worldwide (Roglic, 2016). Nevertheless, the increasing prevalence of diabetes has had a detrimental effect on human health and development globally. According to the literature (Roglic, 2016; Yang et al., 2020) diabetes is a chronic condition characterised by high blood glucose levels and abnormal lipid and protein metabolism. Diabetes-related losses to gross domestic product, including direct and indirect expenses, are expected to equal US$ 1.7 trillion, with low-income and middle-income nations bearing the brunt of the loss at US$ 800 billion.

Apart from the economic strain diabetes places on the healthcare system and the national economy, it frequently results in catastrophic personal expenditures owing to out-of-pocket expenses and income loss due to disability and untimely mortality. In 2014 the number of diabetes cases had risen to 422 million from 180 million in 1980. According to the international diabetes federation's newest prognosis, over 600 million people will live with diabetes by 2035 (Zimmet et al., 2014). However, the increase in type 2 diabetes in children, adolescents, and young people is one of the most concerning aspects of this fast growth (Zimmet et al., 2014). Diabetes is one of Africa's most prevalent non-communicable illnesses, adding to the aging population's growing disease burden in regions such as Ghana (Gatimu et al., 2016; Amoah et al., 2002). Thus, the ramped increase in diabetes among developing countries with less physician-patient ratio calls for an innovative approach to early detection and treatment. Early detection of illnesses such as diabetes is critical since the number of diabetic patients of all ages grows. However, identifying the underlying causes of diabetes's early development has become difficult for medical practitioners. Hence, the continuous growth of diabetic patient data necessitates effective machine learning algorithms that learn from the underlying data's patterns and spot crucial circumstances in patients (Singh et al., 2021; Theis et al., 2021). Artificial intelligence and machine learning (ML) algorithms have opened-up innovative approaches for solving complex challenges in several areas (Nti, 2022). For example, healthcare (Akyeramfo et al., 2019), finance (2020), education (Nti et al., 2021), energy systems (Nti et al., 2020c) and more.

In this study, we proposed a blended ensemble ML predictive model for the early detection of diabetes among patients. Specifically, to:

a examine the health-related features and their impact on predicting diabetes

b propose a novel blended ensemble predictive model with Cohen's Kappa correlation coefficient for selecting base learners.

Cohen's kappa coefficient is adopted to select base learners for our blended ensemble predictive model, so we can balance the diversity and accuracy of the weak learners to improve performance in the ensemble model. Furthermore, that approach allows for a faster and more innovative way of picking base learners that are considerably diverse yet possess high predictive accuracy. Our work makes the following significant contributions:

1 an innovative approach for enhancing the ensemble model's classification power by efficiently picking base classifiers based on the Cohen's Kappa coefficient

2 we test our proposed prediction model and compare its performance to that of existing state-of-the-art research using the Pima Indian Diabetes (PID) Database.

We structured the remaining parts of this paper as follows: Section 2 presents related works, Section 3 methods, and Section 4 results and discussions. Finally, we summarised the work with conclusions and future work in Section 5.

2 Related works

Despite these challenges connected with a diabetes diagnosis, the medical and informatics research communities have premium diabetes prediction using ML approaches (Ahmad et al., 2021). Several works of literature applied ML to predict diabetes using the PID dataset. The PID has nine features and seven hundred 66 entries characterising female patients. This section discusses some works that are closely related.

Singh et al. (2021) proposed an ensemble-based diabetes predictive framework called eDiaPredict. The ML models included XGBoost, Neural Network (NN), Random Forest (RF), Support Vector Machine (SVM), and Decision tree (DT). They applied their framework to the PIMA Indian diabetes (PID) dataset and recorded predictive accuracy of 95%. The PID dataset was applied to different ML and data mining algorithms (NN, SVM, RF, and logistic regression) and achieved 88.6% accuracy (Khanam et al., 2021). Likewise, Sisodia and Sisodia (2018) applied three ML algorithms, DT, SVM, and Naive Bayes, to predict diabetes based on the PID dataset. The study reported an accuracy of 76.30%; however, it concluded that one of the most significant real-world medical challenges is the early identification of diabetes with high precision.

An eXtreme Gradient Boosting (XGBoost) was proposed for predicting diabetes and achieved an area under the receiver operating characteristic curve (AUC) of 0.8768 (Yang et al., 2020). Ayon and Islam (2019) proposed a deep neural network for predicting diabetes using the PID dataset and achieved a success rate of 98.35% accuracy. Similarly, artificial neural network (ANN) was proposed for the early detection of diabetes (Pradhan et al., 2020). However, the study recorded an accuracy of 87.3% and concluded that accuracy enhancement is needed in diabetes prediction with more robust techniques. In Hasan et al. (2020), seven ML algorithms were ensembled to predict diabetes. The algorithm includes k-nearest Neighbour, DT, RF, AdaBoost, Naive Bayes, XGBoost, and Multilayer Perceptron (MLP). They recorded an AUC of 95%, although

the authors believed that enhancement in diabetes prediction is the way forward in future works. Similarly, five different ML classifiers were put together to study the prediction of diabetic patients (Ahmad et al., 2021).

Considering the discussion above, it is evident that most of the existing literature

1 employs non-generalisable features

2 reports only one or two evaluation metrics while overlooking other metrics that may perform worse owing to a trade-off between the several evaluation metrics

3 employs a sizeable dimensional feature size that is unpracticable in many real-world scenarios

4 proposed explicit models that may be non-generalisable.

3 Methodology

Figure 1 shows the proposed block diagram of the novel diabetes prediction model with three phases. Namely,

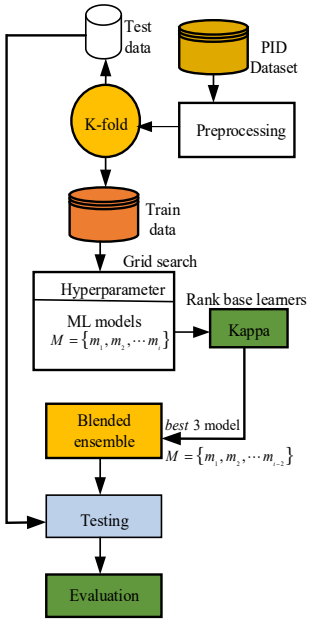
1 data preprocessing

2 machine learning models

3 evaluation metrics.

We explain the details of each phase in the following section.

Figure 1 The proposed block diagram of the novel diabetes prediction model (see online version for colours)



3.1 Study dataset preprocessing

The PID dataset downloaded from Kaggle (https://www.kaggle.com/uciml/Pima-indians-diabetes-database) was used in this study. We preprocessed to remove the noise for higher accuracy achievement. In this work, missing were replaced with the corresponding mean value, and outliers were removed. The mean-based replacement is advantageous since it assigns continuous data without outliers (Hasan et al., 2020). We normalised the dataset with the maximum-minimum function. The clean dataset was of size nine columns with 699 observations. Feature selection was applied to identify the most significant predictors of diabetes. Table 1 shows the details of the study dataset. The dataset was portioned into 85% training and 15% testing; a 10-fold cross-validation was adopted to train the proposed model.

Table 1 Feature of study dataset

Feature	*Description*
BloodPressure	Diastolic blood pressure (mm Hg)
BMI	Body mass index (kg/m2)
SkinThickness	Triceps skinfold thickness (mm)
DiabetesPedigreeFunction	Diabetes pedigree function
Age	Age (years)
Insulin	2-hour serum insulin (µlU/mL)
Glucose	Plasma glucose concentration two h in an oral glucose tolerance test
Pregnancies	The number of times pregnant.
Target	Diabetes diagnoses results (tested_positive: 1, tested_negative: 0)

3.2 Machine learning models

Based on the literature, we selected six well-known ML algorithms. A brief description of these algorithms is given in this section.

- *CatBoost* is an ML algorithm developed by Yandex; it integrates well with popular deep learning frameworks such as Apple's Core ML and Google's TensorFlow (Hancock and Khoshgoftaar, 2020)**.** In addition, it is a member of the Gradient Boosted Decision Trees (GBDT's) ML ensemble techniques.
- *Random forest (RFC)* is a supervised ML technique built on the foundation of DT algorithms for classification, regression, and other glitches that functions by training many DTs (Nti et al., 2019).
- *Gradient boosting (GBC)* is an ML approach for regression and classification applications. It generates a prediction model in a collection of feeble prediction models, most often DTs.
- *Extreme gradient boosting (Xgboost)* is an open-source model for building the gradient boosting (GB) method efficiently and effectively for regression or classification predictive modeling problems

- *Light gradient boosting machine (Lgbm)* is a general open-source distributed GB framework for ML, created initially by Microsoft. It is based on DT algorithms and is used to perform tasks such as classification, regression, and other types of ML.
- *Ada boost classifier (ADA),* or Adaptive Boosting, is a classifier based on ensemble boosting that combines several classifiers to enhance classifier accuracy (Cao et al., 2014). AdaBoost is a technique for iterative ensemble construction. The AdaBoost classifier constructs a robust classifier by merging multiple underperforming classifiers, leading to high accuracy (Zhou et al., 2022).

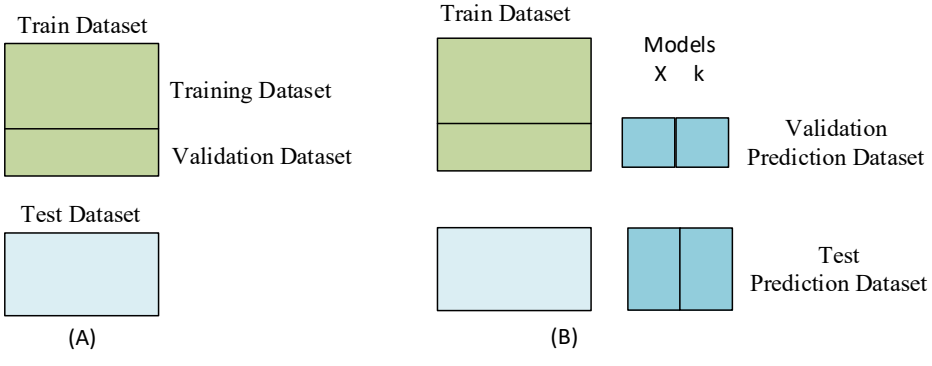
3.2.1 Blended ensemble

The Cohen's Kappa correlation coefficient method was adopted for selecting the best three (3) base learners for the proposed ensemble model based on the blending technique to achieve higher prediction accuracy. Cohen's Kappa statistic is handy but not widely used in ML applications. On the other hand, Cohen's kappa statistic is an excellent metric that is equally adept at handling difficulties involving many unbalanced classes. In summary, the kappa statistic indicates how near the ML classifier categorised examples to the labeled data. Thus, the best base learners can be picked for the ensemble model using the Kappa metric. Cohen's Kappa is defined in equation (1).

$$k = \left(\frac{\hat{y} - y}{1 - y}\right) = \left(1 - \frac{1 - \hat{y}}{1 - y}\right) \tag{1}$$

where y' is the observed value and y the expected value.

Figure 2 Blending ensemble, evaluation of stacking and blending ensemble (see online version for colours)



Blending is a method for combining machine learning algorithms. It is a colloquial term for stacked generalisation or stacking ensemble. The meta-model is fitted on holdout data rather than on out-of-fold predictions produced by the underlying model [23], [24]. The following steps outline in detail the blending ensemble technique.

1 split the training dataset into a training and validation dataset [see Figure 3(a)]

2 train model (x) on the training set

3 make predictions (test models) on the validation and test datasets, as shown in Figure 3(b)

4 use the validation dataset and its predictions as features to build a fresh model (L)

5 make final predictions with the model (L) on the test dataset and meta-features.

3.3 Evaluation metrics

We adopted six (6) commonly used evaluation metrics to examine the proposed model's performance. They are

1 accuracy

2 precision

3 Cohen's Kappa

4 recall

5 area under the receiver operating characteristic curve (AUC)

6 F-score.

We refer readers to Nti et al. (2019a) for detailed definitions of these metrics.

4 Results and discussion

All experiment in this paper was carried out using Python programming language, Scikit-learn, NumPy, Pandas, and the Seaborn. In addition, the grid search technique was adopted for hyperparameter tuning of all ML algorithms. The implementation was done in the cloud using Google Colab. The outcome of our experiments is presented in this section.

4.1 Exploratory data analysis

Figure 3 shows the correlation plot among nine features of the study dataset. It was observed that is a high positive association between pregnancies and age. Figure 4 shows a distribution between observation with tested positive (1) and tested negative (0). It was observed that there is a high imbalance in the dataset, i.e., tested negative (66.7%) and tested positive (33.3%). Imbalance datasets in classification affect the accuracy of the models. Hence, we adopted the SMOTE to balance the study data in this study.

Figure 3 Correlation plot between features (see online version for colours)

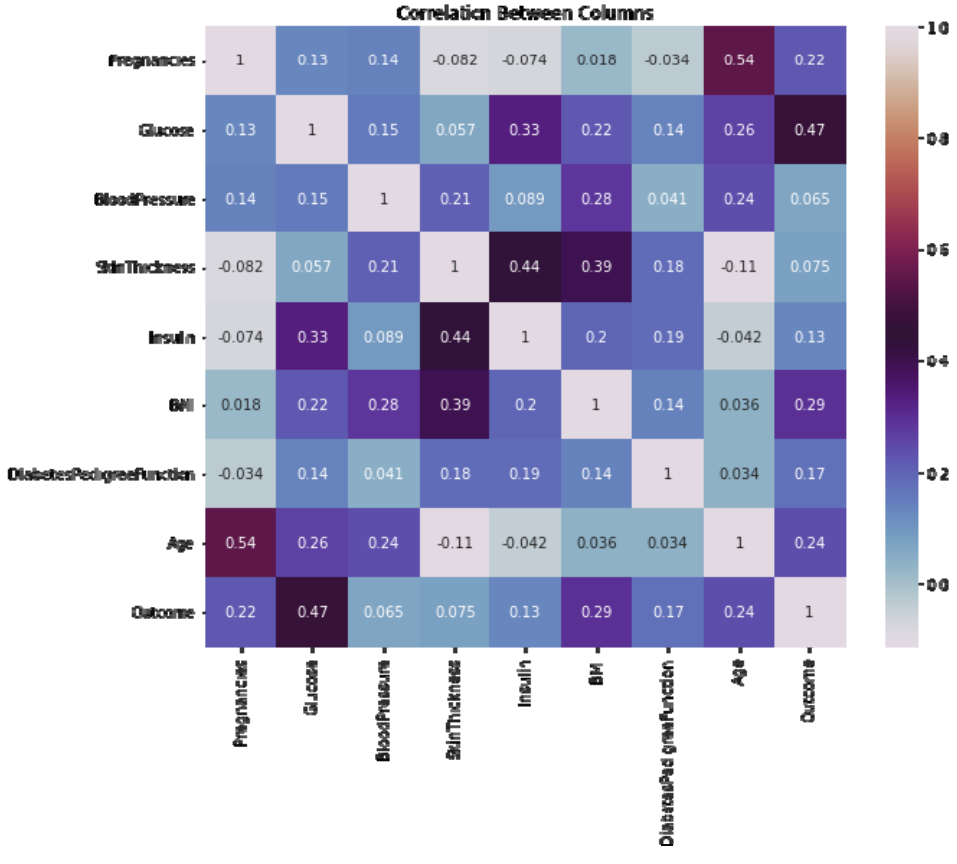
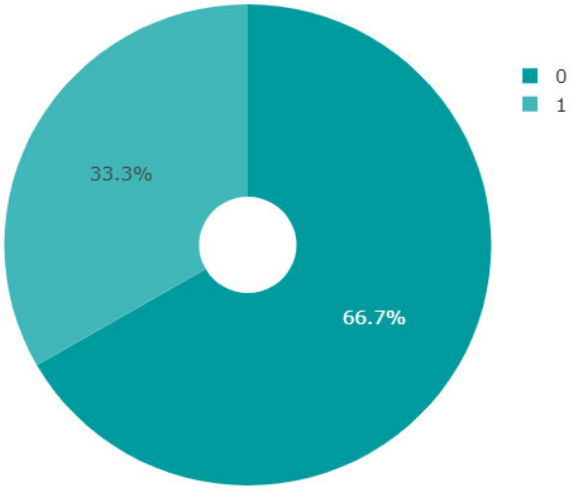


Figure 4 Distribution of the target feature-outcome (see online version for colours)



4.2 Feature selection

The quality of features (inputs) to an ML model deeply affects its performance. Here we aim to identify the features among the diabetes dataset that are highly significant in predicting diabetes. Figure 5 shows the vital ranking of the features in the study dataset. It was observed that glucose was the feature with the highest importance, followed by BMI, age, and DiabetesPedigreeFunction. These results suggest that patients with high glucose levels are at higher risk of diabetes.

Figure 5 Feature importance ranking (see online version for colours)

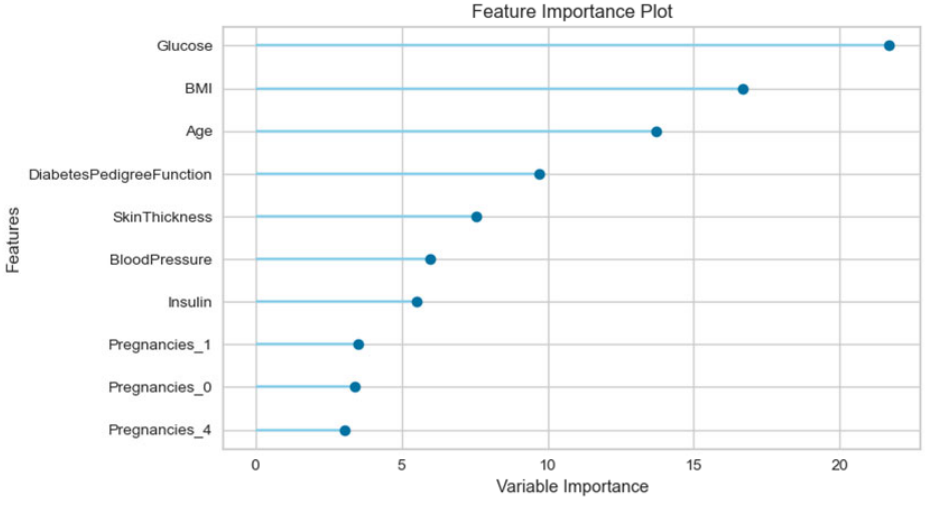
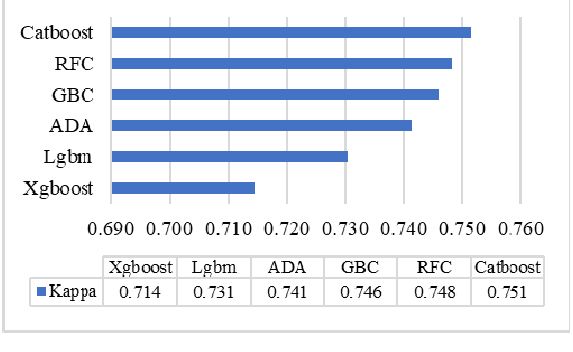


Figure 6 Base learner ranking based on Cohen's Kappa statistic (see online version for colours)



4.3 Base learner selection

Figure 6 shows the ranking of the base learner based on Cohen's Kappa statistic. All six base learners were applied to the training dataset; we ranked their performance based on Cohen's Kappa statistic. It was observed that the Catboost (0.751) outperformed the RFC (0.748), GBC (0.746), ADA (0.741), Lgbm (0.731) and Xgboost (0.741). From the

outcome, the best three performing models (i.e., Catboost, RFC, and GBC) were picked and ensembled using the blending technique.

4.4 Blended ensemble

Table 2 shows the performance of our proposed model with other models. It was observed that the proposed blended ensemble classifier achieved an accuracy of 89%, AUC (94.9%), recall (80.7%), precision (87.4%), and F1-score (83.3%). The results show that our proposed model outperformed the random forest classifier and the Catboost classifier, and the gradient boosts classier with an accuracy rate of approximately 19% higher than what was achieved by the Catboost, GBC, and the RF classifiers. Also, the high precision (87.4%) and recall (80.7) show that the proposed model can efficiently distinguish between positive and negative classes. Figure 7 shows the AUC curve of the proposed model. Figure 8 shows the proposed model's Kolmogorov–Smirnov (KS) statistic plot.

Table 2 Proposed model performance with other models

Model	*Accuracy*	*AUC*	*Recall*	*Prec.*	*F1*
Catboost	0.7464	0.8164	0.6790	0.6239	0.6481
GBC	0.7448	0.8126	0.6883	0.6145	0.6469
RF	0.7496	0.8227	0.6606	0.6301	0.6418
Our model	*0.890*	*0.949*	*0.807*	*0.874*	*0.833*

Figure 7 AUC curve of the proposed model (see online version for colours)

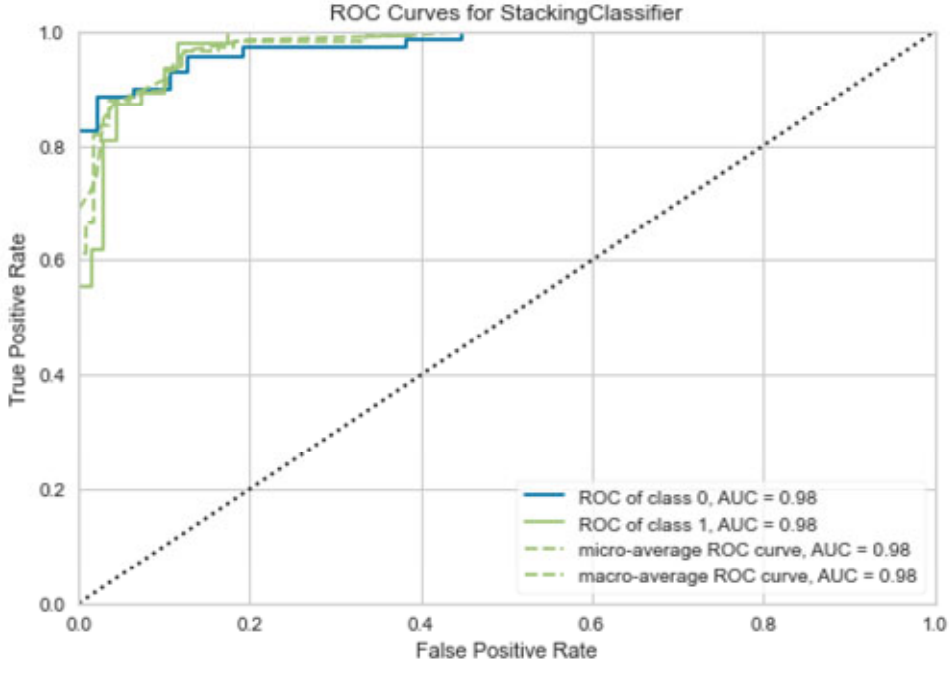
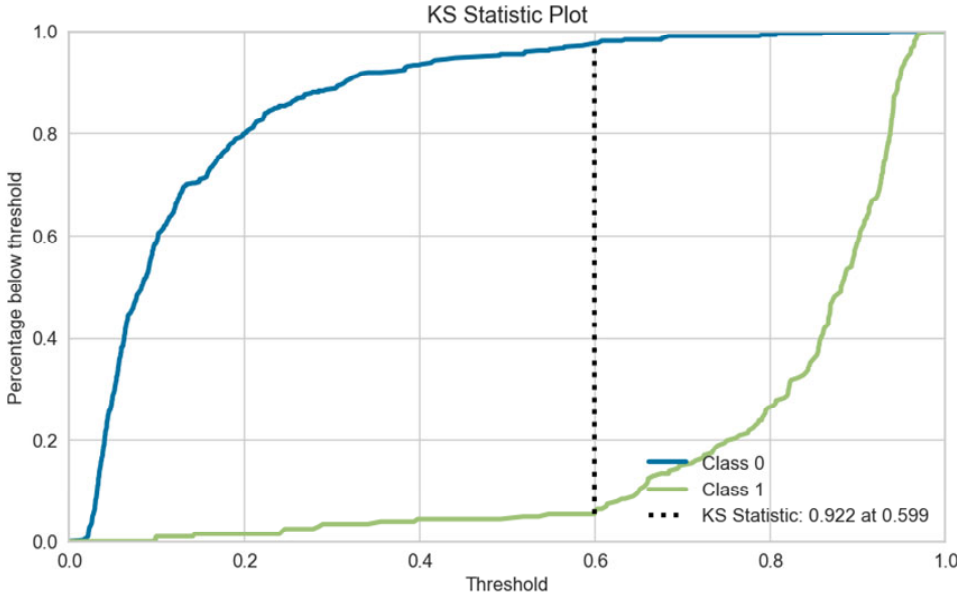


Figure 8 KS statistic plot of the proposed model (see online version for colours)



4.5 *Comparison of the proposed model with other studies on the PID dataset*

Table 3 compares our proposed model with other state-of-the-art studies on predicting diabetes on the PID dataset. The aim was to examine the efficiency and robustness of the proposed model. From Table 3, our proposed model outperformed approximately 67% of related studies discussed in this paper. This revelation shows that this study substantially improved predicting diabetes, which is further knowledge. Also, our come has shown that the performance of machine learning algorithms in predicting diabetes can be improved beyond the existing literature with adequate features and based on learners' selection techniques.

Table 3 Comparison of the proposed model with other studies on the PID dataset

Ref.	*Accuracy*
Singh et al. (2021)	95%
Khanam and Foo (2021)	88.6%
Sisodia and Sisodia (2018)	76.30%
Ayon and Islam (2019)	98.35
Pradhan et al.[17]	87.3%
Our model	*89%*

5 Conclusions

Diabetes is a long-lasting illness with no known treatment; hence, early identification is critical. In this study, we proposed a blended ensemble predictive model using the Kappa correlation coefficient as the basis to select our base learning. We experimented with our model on the PID dataset and assessed its performance with six evaluation metrics. Namely, accuracy, precision, Cohen's Kappa, recall, area under the receiver operating characteristic curve (AUC), and F-score. The experimental results demonstrate the system's suitability, with an accuracy of 89%. Also, it was observed that in predicting diabetes, glucose, BMI, age, and DiabetesPedigreeFunction were the most significant features.

In future research, we intend to advance algorithms in related disciplines and tackle present issues using unique and modern methods, such as deep neural networks. In addition, we want to gather additional data, such as picture and lifestyle data, to enhance the data collecting process's quality, update the system, and develop more robust and highly accurate models.

References

Ahmad, H.F., Mukhtar, H., Alaqail, H., Seliaman, M. and Alhumam, A. (2021) 'Investigating health-related features and their impact on the prediction of diabetes using machine learning', *Appl. Sci.*, Vol. 11, No. 3, pp.1–18, doi: 10.3390/app11031173.

Akyeramfo-Sam, S., Addo Philip, A., Yeboah, D., Nartey, N.C. and Kofi Nti, I. (2019) 'A web-based skin disease diagnosis using convolutional neural networks', *Int. J. Inf. Technol. Comput. Sci.*, November, Vol. 11, No. 11, pp.54–60, doi: 10.5815/ijitcs.2019.11.06.

Amoah, A.G.B., Owusu, S.K. and Adjei, S. (2002) 'Diabetes in Ghana: a community based prevalence study in Greater Accra', *Diabetes Res. Clin. Pract.*, June, Vol. 56, No. 3, pp.197–205, doi: 10.1016/S0168-8227(01)00374-6.

Ayon, S.I. and Islam, M. (2019) 'Diabetes prediction: a deep learning approach', *Int. J. Inf. Eng. Electron. Bus.*, Vol. 11, No. 2, pp.21–27, doi: 10.5815/ijieeb.2019.02.03.

Cao, J., Chen, J. and Li, H. (2014) 'An adaboost-backpropagation neural network for automated image sentiment classification', *Sci. World J.*, Vol. 2014, doi: 10.1155/2014/364649.

Gatimu, S.M., Milimo, B.W. and Sebastian, M.S. (2016) 'Prevalence and determinants of diabetes among older adults in Ghana', *BMC Public Health*, December, Vol. 16, No. 1, p.1174, doi: 10.1186/s12889-016-3845-8.

Hancock, J.T. and Khoshgoftaar, T.M. (2020) 'CatBoost for big data: an interdisciplinary review', *J. Big Data*, December, Vol. 7, No. 1, p.94, doi: 10.1186/s40537-020-00369-8.

Hasan, M.K., Alam, M.A., Das, D., Hossain, E. and Hasan, M. (2020) 'Diabetes prediction using ensembling of different machine learning classifiers', *IEEE Access*, Vol. 8, pp.76516–76531, doi: 10.1109/ACCESS.2020.2989857.

Khanam, J.J. and Foo, S.Y. (20212021) 'A comparison of machine learning algorithms for diabetes prediction', *ICT Express*, Vol. 7, No. 4, pp.432–439, doi: 10.1016/j.icte.2021.02.004.

Nti, I.K., Adekoya, A.F. and Weyori, B.A. (2019a) 'A systematic review of fundamental and technical analysis of stock market predictions', *Artif. Intell. Rev.*, April, Vol. 53, No. 4, pp.3007–3057, doi: 10.1007/s10462-019-09754-z.

Nti, I.K., Adekoya, A.F. and Weyori, B.A. (2019b) 'Random forest based feature selection of macroeconomic variables for stock market prediction', *Am. J. Appl. Sci.*, July, Vol. 16, No. 7, pp.200–212, doi: 10.3844/ajassp.2019.200.212.

Nti, I.K., Adekoya, A.F. and Weyori, B.A. (2020a) 'A comprehensive evaluation of ensemble learning for stock-market prediction', *J. Big Data*, December, Vol. 7, No. 1, p.20, doi: 10.1186/s40537-020-00299-5.

Nti, I.K., Adekoya, A.F. and Weyori, B.A. (2020b) 'Efficient stock-market prediction using ensemble support vector machine', *Open Comput. Sci.*, July, Vol. 10, No. 1, pp.153–163, doi: 10.1515/comp-2020-0199.

Nti, I.K., Akyeramfo-Sam, S., Bediako-Kyeremeh, B. and Agyemang, S. (2021) 'Prediction of social media effects on students' academic performance using Machine Learning Algorithms (MLAs)', *J. Comput. Educ.*, Aug., doi: 10.1007/s40692-021-00201-z.

Nti, I.K., Quarcoo, J.A., Aning, J. and Fosu, G.K. (2022) 'A mini-review of machine learning in big data analytics: applications, challenges, and prospects', *Big Data Min. Anal.*, June, Vol. 5, No. 2, pp.81–97, doi: 10.26599/BDMA.2021.9020028.

Nti, I.K., Teimeh, M., Adekoya, A.F. and Nyarko-boateng, O. (2020c) 'Forecasting electricity consumption of residential users based on lifestyle data using artificial neural networks', *ICTACT J. Soft Comput.*, Vol. 10, No. 03, pp.2107–2116, doi: 10.21917/ijsc.2020.0300.

Pradhan, N., Rani, G., Dhaka, V.S. and Poonia, R.C. (2020) 'Diabetes prediction using artificial neural network', *Deep Learn. Tech. Biomed. Heal. Informatics*, Vol. 121, pp.327–339, doi: 10.1016/B978-0-12-819061-6.00014-8.

Roglic, G. (2016) 'WHO global report on diabetes: a summary', *International Journal of Noncommunicable Diseases*, Vol. 1, No. 1, p.3, https://www.ijncd.org/text.asp?2016/1/1/3/184853.

Singh, A., Dhillon, A., Kumar, N., Hossain, M.S., Muhammad, G. and Kumar, M. (2021) 'eDiaPredict: an ensemble-based framework for diabetes prediction', *ACM Trans. Multimed. Comput. Commun. Appl.*, June, Vol. 17, No. 2s, pp.1–26, doi: 10.1145/3415155.

Sisodia, D. and Sisodia, D.S. (2018) 'Prediction of diabetes using classification algorithms', *Procedia Comput. Sci.*, Vol. 132, No. Iccids, pp.1578–1585, doi: 10.1016/j.procs.2018.05.122.

Theis, J., Galanter, W., Boyd, A. and Darabi, H. (2021) 'Improving the in-hospital mortality prediction of diabetes ICU patients using a process mining/deep learning architecture', *IEEE J. Biomed. Heal. Informatics*, Vol. 26, No. 1, pp.388–399, doi: 10.1109/JBHI.2021.3092969.

Wu, T., Zhang, W., Jiao, X., Guo, W. and Alhaj Hamoud, Y. (2021) 'Evaluation of stacking and blending ensemble learning methods for estimating daily reference evapotranspiration', *Computers and Electronics in Agriculture*, Vol. 184, p.106039, https://doi.org/10.1016/j.compag.2021.106039.

Yang, H. et al. (2020) 'Risk prediction of diabetes: big data mining with fusion of multifarious physical examination indicators', *Inf. Fusion*, November, Vol. 75, pp.140–149, 2021, doi: 10.1016/j.inffus.2021.02.015.

Zhou, Y., Mazzuchi, T.A. and Sarkani, S. (2020) 'M-AdaBoost – a based ensemble system for network intrusion detection', *Expert Syst. Appl.*, April, Vol. 162, p.113864, doi: 10.1016/j.eswa.2020.113864.

Zimmet, P.Z., Magliano, D.J., Herman, W.H. and Shaw, J.E. (2014) 'Diabetes: a 21st-century challenge', *Lancet Diabetes Endocrinol.*, January, Vol. 2, No. 1, pp.56–64, doi: 10.1016/S2213-8587(13)70112-8.