RESEARCH ARTICLE

CYPRUS
JOURNAL OF MEDICAL SCIENCES

DOI: 10.4274/cjms.2025.2024-126
Cyprus J Med Sci 2025;10(Suppl 1):111-115

Ensemble Learning for Diabetes Classification Using Voting Classifier on CDC Health Indicators Dataset

Bardia Arman[1], Kian Jazayeri[2], Erbug Celebi[3], Kezban Alpan[4], Kamil Dimililer[5]

[1]Artificial Intelligence Application and Research Center, Cyprus International University Faculty of Engineering, Nicosia, North Cyprus
[2]Department of Information Technology, Cyprus International University School of Applied, Nicosia, North Cyprus
[3]Artificial Intelligence Application and Research Center, Cyprus International University, Nicosia, North Cyprus
[4]HND Digital Technologies Programme, David Game Higher Education, London, United Kingdom
[5]Department of Electrical and Electronic Engineering Applied Artificial Intelligence Research Centre Near East University, Nicosia, North Cyprus

Abstract

BACKGROUND/AIMS: Diabetes is one of the paramount public health challenges, affecting millions worldwide. Classification models can boost early detection and aid in treatment, particularly for diabetes type 2. This study, therefore, uses an ensemble learning approach for classifying diabetes type 2, utilizing a soft voting classifier, using multiple machine learning techniques on the Centers for Disease Control and Prevention Health Indicators Dataset.

MATERIALS AND METHODS: An ensemble model was developed in which the predictions of five machine learning algorithms were combined: XGBoost, Random Forest, Gradient Boosting, Support Vector Machine, and convolutional neural network-long short-term memory. Each model is trained using bootstrapped re-sampling, and predictions are aggregated through soft voting to improve classification performance on the test set.

RESULTS: On the test set, it achieved a classification accuracy of 87.8%, precision of 99.5%, recall of 99.51%, and an F1 score of 99.2%, hence proving high efficacy in identifying diabetes type 2 cases.

CONCLUSION: It follows that the proposed ensemble model efficiently classifies diabetes type 2 with high precision and recall; hence, it underpins the importance of ensemble learning in boosting the accuracy of classification. This may provide a reliable tool for early detection of diabetes, contributing to better patient outcomes through timely intervention.

Keywords: Convolutional neural network-long short-term memory, disease control and prevention, Support Vector Machin, XGBoost

INTRODUCTION

Diabetes mellitus is a chronic disease that impairs the body's ability to convert food into energy, resulting in elevated blood sugar levels due to insufficient or ineffective insulin production.[1] Type 2 diabetes accounts for 90-95% of cases, in which insulin is produced but is inadequate in its action. High blood sugar can damage blood vessels and organs, leading to severe complications such as cardiovascular disease, kidney failure, and neuropathy. Moreover, undiagnosed diabetes can reduce life expectancy by up to 8 years, highlighting the urgent need for early detection and intervention.[2]

Common symptoms include frequent thirst, nighttime urination, fatigue, unplanned weight loss, slow wound healing, increased hunger,

To cite this article: Arman B, Jazayeri K, Çelebi E, Alpan K, Dimiller L. Ensemble learning for diabetes classification using voting classifier on CDC health indicators dataset. Cyprus J Med Sci. 2025;10(Suppl 1):111-115

ORCID IDs of the authors: B.A. 0009-0001-9449-2688; K.J. 0000-0003-2843-7354; E.Ç. 0000-0003-3289-3328; K.A. 0000-0002-0492-1275; K.D. 0000-0002-2751-0479.

Corresponding author: Bardia Arman
E-mail: bardia.arman@gmail.com
ORCID ID: orcid.org/0009-0001-9449-2688

Received: 18.11.2024
Accepted: 15.04.2025
Publication Date: 04.06.2025

and blurred vision.[3] Diagnosis primarily relies on blood glucose measurement,[4] influenced by various health indicators. Research indicates that obesity, high blood pressure (HighBP), high cholesterol (HighCol), stroke history, and cardiovascular diseases are significant risk factors for diabetes.[5-7] These health-related factors complicate diabetes management, necessitating effective predictive models.

In public health, classifying patients as diabetic or non-diabetic using advanced machine learning techniques can significantly enhance early detection and treatment strategies. This study aims to leverage ensemble learning models to improve classification accuracy for type 2 diabetes, integrating multiple machine learning algorithms through a voting classifier. This approach seeks to provide a reliable tool for healthcare professionals to identify at-risk individuals, ultimately contributing to better patient outcomes.

Previous studies have explored various machine learning approaches for diabetes prediction. For instance, Singh and Singh[8] achieved 83.6% accuracy with a stacking-based ensemble framework. Kibria et al.[9] reached 90% accuracy using a soft voting classifier. Dogru et al.[10] developed a hybrid model achieving 99.6% accuracy. Sunny et al.[11] also proposed a soft voting ensemble method for accurate diabetes risk diagnosis.

MATERIALS AND METHODS

Statistical Analysis

To conduct this study and the proposed method, the Centers for Disease Control and Prevention (CDC) Diabetes Health Indicators Dataset was selected. The dataset contains 253,680 samples with 35 features, consisting of medical and behavioural data of individuals. The proposed method is implemented in the feature group shown in Table 1. Since this study primarily focuses on the classification of diabetes based on medical data, five medical features were selected from the dataset: including individuals' HighBP, HighCol, body mass index (BMI), which determines whether they are at a healthy weight, whether they have had a stroke in their medical history, and whether they have any history of heart disease or heart attack. The age and gender of individuals were used as demographic data.

Machine Learning Algorithms

Random forest: Random Forest allows the generation of various models and classifications by training each decision tree on a different observation sample. The algorithm creates a decision tree for each example and determines the estimated value of each decision tree.[12]

Table 1. Selected features for the study 1= diabetes

Feature	Description	Definition
HighBP	High blood pressure	0= No, 1= Yes
HighCol	High cholesterol	0= No, 1= Yes
BMI	Body mass index	Numerical
Stroke	Stroke history	0= No, 1= Yes
Heart diseaseor attack	Heart disease history	0= No, 1= Yes
Age	Age of individual	Numerical
Sex	Biological classification	0= Female, 1= Male
Education	Education level	Scale 1 to 6
Diabetes binary	Target value	0= No diabetes, 1= Diabetes

Gradient boosting: In the first stage, an initial tree is created. It then calculates an error based on the difference between the actual value of the target variable and the value predicted by the tree a second tree is created to reduce this error. The second tree is used to estimate the negative gradient of the error predicted by the first tree.[14]

eXtreme gradient boosting: XGBoost is one of the ensemble methods that operates on decision trees, unlike traditional GB, which aims to minimize the errors of the model.

$L(y, y^{\wedge}) + \Omega(f)$ (1)

Support vector machines: As represented in Figure 1, support vector machine (SVM) is a supervised learning technique used for classification and regression. The SVM algorithm draws lines to separate sets of two or more points that are placed on a plane. Considering two data sets, this line aims to maximize the distance between the points of both sets. The decision boundary that needs to be determined for separation finds the best margin between classes and defines the hyperplane.[15]

Convolutional-longitudinal-short-term neural network: A deep learning architecture formed by the combination of CNN and LSTM networks is known as CNN-LSTM, and represented in Figure 2. CNNs capture spatial relationships within the dataset through convolution. LSTM, a type of recurrent neural network, is successful in capturing long-term dependencies. The combination of CNN-LSTM enables learning both spatial and temporal features of the data.[16]

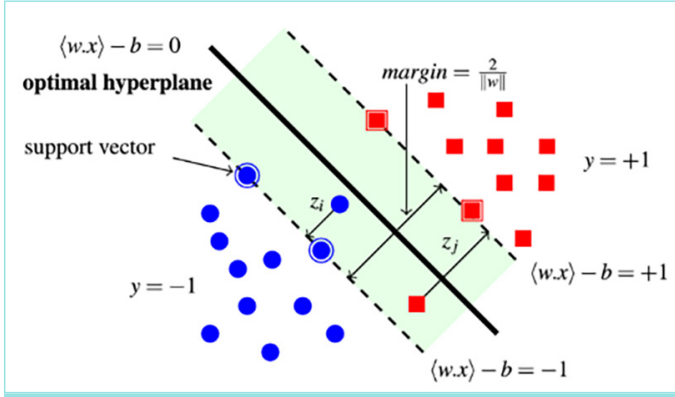

Figure 1. Classification of the datapoints into two classes with SVM.[17]

SVM: Support vector machine.

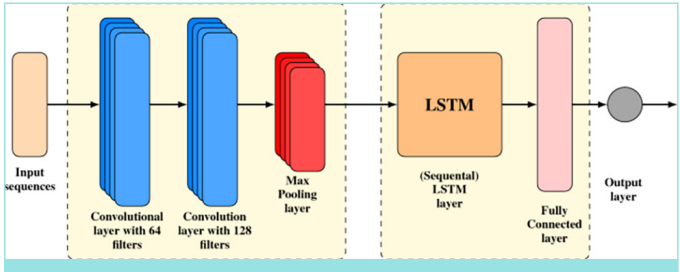

Figure 2. Example CNN-LSTM model.[18]

CNN-LSTM: Convolutional neural network-long short-term memory.

Correlation: Figure 3 illustrates the correlation between Diabetes Binary and eight characteristic variables. Health-related factors such as HighBP, BMI, HighCol, and heart disease show a positive correlation with diabetes. Among demographic variables, education level shows a negative correlation and stands out as significant, even when compared among medical factors. Examining both positive and negative correlations provides a comprehensive view of the dataset.

Proposed model: The dataset underwent an loading, preprocessing, and an 80-20 train-test split, followed by feature standardization. Bootstrap resampling was applied to the training set to enhance model robustness. Multiple models (XGBoost, Random Forest, Gradient Boosting, SVM, and CNN-LSTM) were trained on resampled datasets (Figure 4). The predictions were combined using soft voting, with the probabilities averaged for the predictions. The model's performance was evaluated using accuracy, precision, recall, and F1 score. This ensemble approach improves robustness and mitigates overfitting.

RESULTS

Performance Analysis

The effectiveness of different machine learning algorithms in binary classification varies. The accuracy of measurements evaluates models such as Decision Tree, Random Forest, K-nearest neighbor (KNN), CatBoost, Gaussian Naive Bayes, Logistic Regression, Linear Discriminant, Gradient Boosting, and the proposed model. As shown in Table 2 and Figure 5, the proposed model achieved the highest accuracy (87.8%) in classifying the CDC Diabetes Indicator Dataset.

Beyond accuracy, precision, recall, and F1 score provide deeper insight into model performance. Random Forest and KNN excel in these metrics, demonstrating strong predictive power and minimizing false positives and false negatives. While the proposed model has the highest accuracy, its recall and F1 score confirm its ability to correctly classify positive cases. In contrast, Gaussian naive Bayes and logistic regression show moderate accuracy with lower precision, recall, and F1 scores,

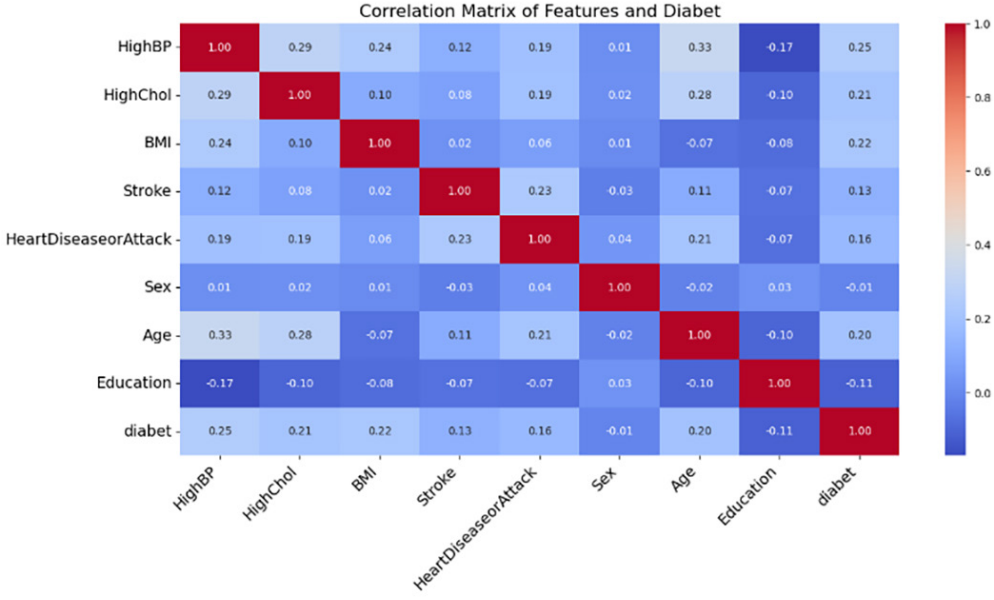

Figure 3. Correlation diagram.

HighBP: High blood pressure, HighCol: High cholesterol, BMI: Body mass index.

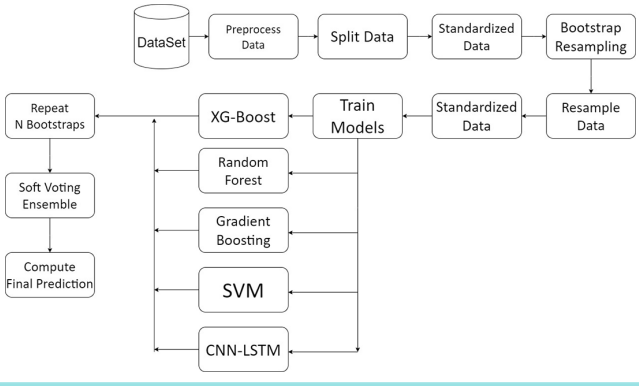

Figure 4. Model flowchart.

SVM: Support vector machine, CNN-LSTM: convolutional neural network-long short-term memory.

Table 2. Confusion metrics of different models

Model	Accuracy	Precision	Recall	F1 score
Decision tree	80.2%	97%	97.2%	97.8%
Random forest	84.9%	99%	99%	99%
KNN	85.4%	98%	98.5%	98.48%
Cat boost	86.6%	94.3%	94%	95.1%
Gaussian Naïve Bayes	74.1%	95.4%	96.5%	93.4%
Logistic regression	72.9%	89.7%	92.1%	91%
Linear discriminant	72.1%	96.3%	96.3%	97%
Gradient boosting	83.8%	97.4%	96.8%	97.2%
Proposed model	87.8%	99.5%	99.51%	99.2%

KNN: K-nearest neighbor

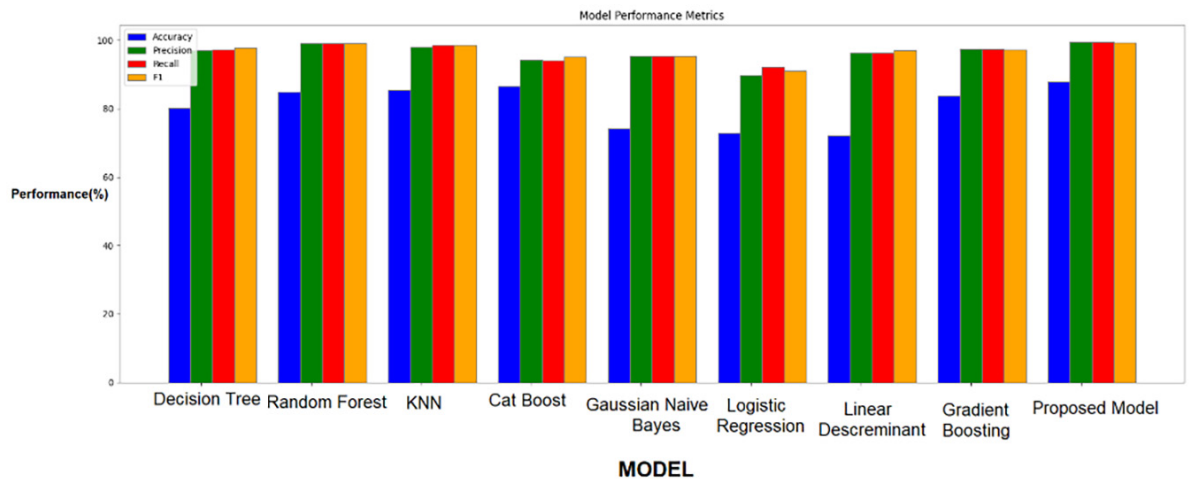

Figure 5. Model performance comparison.

KNN: K-nearest neighbor

indicating a higher rate of misclassification. Decision Trees and Linear Models, though less accurate, outperform Gaussian Naive Bayes and Logistic Regression in precision, recall, and F1 score.

Finally, Gradient Boosting shows competitive performance with high precision, recall, and F1 score, although the sentence is incomplete.

It is slightly weaker than Random Forest and KNN. Overall, this analysis emphasizes the importance of examining different metrics beyond accuracy to fully evaluate model performance, especially in scenarios where the costs of false positives and false negatives are different.

DISCUSSION

This study presents a robust ensemble learning model for classifying type 2 diabetes, utilizing five machine learning algorithms: XGBoost, Random Forest, Gradient Boosting, SVM, and CNN-LSTM. The ensemble approach, employing soft voting, achieved a classification accuracy of 87.8%; with precision, recall, and F1 scores of 99.5%, 99.51%, and 99.2%, respectively. These results demonstrate the model's effectiveness in accurately identifying diabetic patients while minimizing false positives and negatives, which is crucial in clinical settings.

The high performance of the ensemble model is attributed to its ability to leverage the strengths of diverse algorithms. Each algorithm contributes unique insights, allowing the ensemble to capture complex patterns in the data. For instance, Random Forest handles overfitting through its ensemble of decision trees, while XGBoost optimizes predictive accuracy via gradient boosting. The integration of CNN-LSTM captures both spatial and temporal features, which is beneficial for analyzing health-related time-series data.

However, the study has limitations. Reliance on a single dataset may restrict the generalizability of findings across different populations. Future research should validate the model on diverse datasets to ensure broader applicability. Additionally, the complexity of the ensemble model may pose challenges in real-time clinical implementation and interpretability. Simplifying the model or employing explainable AI techniques could enhance usability for healthcare professionals. In conclusion, this study highlights the potential of ensemble learning in diabetes prediction, offering a powerful tool for early detection and intervention.

Study Limitations

This study's reliance on a single dataset may limit the generalizability of findings across diverse populations. Additionally, the model's complexity could present challenges in real-time implementation and interpretability in clinical settings.

CONCLUSION

Diabetes prediction remains essential in public health for early intervention and management. This study's ensemble model, leveraging the strengths of various machine learning techniques, achieved superior predictive accuracy and reliability in classifying diabetes based on health indicators. With high scores in precision, recall, and F1 metrics, the model proves valuable for accurately identifying diabetic patients and minimizing classification errors. Although further work is needed to enhance interpretability and generalizability, the findings suggest that ensemble learning offers a powerful approach for diabetes prediction, potentially aiding clinicians in early detection and reducing the impact of diabetes on patient health outcomes.

MAIN POINTS

- An ensemble learning model combining XGBoost, Random Forest, Gradient Boosting, SVM, and CNN-LSTM was developed for diabetes type 2 classification.
- The model achieved high performance with 87.8% accuracy, 99.5% precision, 99.51% recall, and 99.2% F1 score.

- The results highlight the potential of ensemble models in improving early detection and management of diabetes.

ETHICS

Ethics Committee Approval: Not available.

Informed Consent: Not available.

Footnotes

Authorship Contributions

Concept: B.A., K.J., E.Ç., K.A., K.D., Design: B.A., K.J., E.Ç., K.A., Data Collection and/or Processing: B.A., E.Ç., K.A., K.D., Analysis and/or Interpretation: B.A., K.J., K.D., Literature Search: B.A., K.J., E.Ç., K.A., K.D., Writing: B.A., K.J., E.Ç., K.A., K.D.

DISCLOSURES

Conflict of Interest: No conflict of interest was declared by the authors.

Financial Disclosure: The authors declared that this study had received no financial support.

REFERENCES

1. Alberti KG, Zimmet PZ. Definition, diagnosis and classification of diabetes mellitus and its complications. Part 1: Diagnosis and classification of diabetes mellitus. Provisional report of a WHO consultation. Diabet Med. 1998; 15(7): 539-53.
2. Halim R, Dahi Z, Halim NM. Tip 2 diyabette devamlı egzersiz ve safran kullanımının insulin direnci ve glikozun hücre içine alımına etkisi. In 5th International Students Science Congress. (2021).
3. Tachkov K, Mitov K, Koleva Y, Mitkova Z, Kamusheva M, Dimitrova M, et al. Life expectancy and survival analysis of patients with diabetes compared to the non-diabetic population in Bulgaria. PloS one. 2020; 15(5): 0232815.
4. American Diabetes Association. Diagnosis and classification of diabetes mellitus. Diabetes Care. 2010; 33(Suppl 1): 62-9.
5. Inzucchi SE. Diagnosis of diabetes. N Engl J Med. 2012; 367(6): 542-50.
6. Lu W, Resnick HE, Jablonski KA, Jones KL, Jain AK, Howard WJ, et al. Non-HDL cholesterol as a predictor of cardiovascular disease in type 2 diabetes: the strong heart study. Diabetes care. 2003; 26(1): 16-23.
7. Lehto S, Rönnemaa T, Pyörälä K, Laakso M. Cardiovascular risk factors clustering with endogenous hyperinsulinaemia predict death from coronary heart disease in patients with type 2 diabetes. Diabetologia. 2000; 43(2): 148-55.
8. Singh N, Singh P. A stacked generalization approach for diagnosis and prediction of type 2 diabetes mellitus. Computational Intelligence in Data Mining: Proceedings of the International Conference on ICCIDM. 2020; 559-70.
9. Kibria HB, Nahiduzzaman M, Goni MOF, Ahsan M, Haider J. An ensemble approach for the prediction of diabetes mellitus using a soft voting classifier with an explainable AI. Sensors. 2022; 22(19): 7268.
10. Dogru A, Buyrukoglu S, Arı M. A hybrid super ensemble learning model for the early-stage prediction of diabetes risk. Med Biol Eng Comput. 2023; 61(3): 785-97.
11. Sunny S, Pinky S, Jalal S, Kayser M, Wadud M, Mansoor N. Soft voting ensemble-based approach for diagnosing diabetes mellitus. 2024 International Conference on Advances in Computing, Communication, Electrical, and Smart Systems. 2024; 1-6.
12. Biau G, Scornet E. A random forest guided tour. 2015; 25(2): 197-227.
13. Sahour H, Gholami V, Torkaman J, Vazifedan M, Saeedi S. Random forest and extreme gradient boosting algorithms for streamflow modeling using vessel features and tree-rings. Environmental Sciences. 2021; 80: 1-14.
14. Natekin A, Knoll A. Gradient boosting machines, a tutorial. Front Neurorobot. 2013; 7: 21.
15. Aldino A, Saputra A, Nurkholis A, Setiawansyah S. Application of Support Vector Machine (SVM) Algorithm in classification of low-cape communities in lampung timur. Building of Informatics, Technology and Science (BITS). 2021; 3(3): 325-30.
16. Mutegeki R, Han DS. A CNN-LSTM approach to human activity recognition. In 2020 international conference on artificial intelligence in information and communication (ICAIIC). 2020; 362-6.
17. Do TN. Automatic learning algorithms for local Support Vector Machines. SN Computer Science.
18. World Health Organization. Diabetes [Internet]. Geneva: World Health Organization; 2024 Aug 8 [cited 2025 Jun 3]. Available from: https://www.who.int/healthtopics/diabetes#tab=tab1