OPEN ACCESS
Computing Open
Vol. 2 (2024) 2450006 (13 pages)
© The Author(s)
DOI: 10.1142/S2972370124500065

World Scientific
www.worldscientific.com

Machine Learning-Based Diabetes Risk Prediction Using Associated Behavioral Features

Ayodeji O. J. Ibitoye*
School of Computing and Mathematical Sciences
University of Greenwich, SE10 9LS, London, United Kingdom
a.o.ibitoye@greenwich.ac.uk

Joseph D. Akinyemi
Department of Computer Science
University of York, YO10 5DD, York, United Kingdom
joseph.akinyemi@york.ac.uk

Olufade F. W. Onifade
Department of Computer Science
University of Ibadan, Nigeria
ofw.onifade@ui.edu.ng

Received 29 November 2023
Revised 4 May 2024
Accepted 13 May 2024
Published 12 June 2024

Diabetes is a global health concern that affects people of all races. With different uncertainties in human lifestyles, it is difficult to predict diabetes while assuming that the risk patterns are the same for all. The likelihood of diabetes in a patient is mostly predicted using machine learning (ML) models on features explicitly available in datasets, while the intrinsic relationship between features viz-a-viz their potential relevance to the presence of diabetes is oftentimes neglected. In this work, we explored feature importance and correlation to derive the top 15 feature pairs from a dataset of 263,882 samples of anonymized patient information. These top-15 feature pairs were fed into five different ML models (decision tree (DT), neural networks (NN), random forest (RF), support vector machine (SVM) and extreme gradient boosting (XGB)) for predicting the likelihood of diabetes, while also feeding the direct features (without correlated pairing) separately into the same 5 ML models. The models' performances were evaluated using accuracy, precision, recall and $F1$-score and NN presented the best performance overall achieving an $F1$-score of 85% for the correlated feature pairs (CF) and 75% for the direct feature pairs.

*Corresponding author.

The results confirm the importance of the correlation/relationship between features in predicting the likelihood of diabetes in patients more accurately.

Keywords: Diabetes; machine learning; risk prediction; paired relationship; decision support.

1. Introduction

Diabetes is a medical ailment that is depicted by high quantities of blood glucose or sugar in the human body. With two basic types of diabetes, health experts explained that Type 1 diabetes is initiated by a deficiency in insulin generation by the pancreas, and Type 2 diabetes is characterized by the body's cells becoming defiant to the results of insulin.[1] Despite being a non-communicable disease, type 2 diabetes has recently gained the status of an epidemic silent killer.[2] It is not a respect of age, ethnicity, or nationality. If a person's blood sugar balance varies from 100 to 125 mg/dL, he is diagnosed with prediabetes since the actual normal range of glucose levels in the human body is 70–99 decimeters.[3] In 2021, the International Diabetes Federation (IDF) records showed that the world had over 536.6 million diabetics,[4] and predicted that 783.2 million people are expected to be living with diabetes in the year 2045. Since diabetes is a chronic condition that has no cure now, detecting diabetes early is crucial for effective and ongoing management and/or treatment, although it remains a challenging task.[5] While various interventions and lifestyle changes can help control and mitigate the effects of diabetes, with the manual prevention approach, hidden patterns in data may go unnoticed, leading to suboptimal decision-making and depriving patients of the care they need. To address this issue, the automated identification of diabetes with improved accuracy through data mining is essential. Data mining continues to grow in popularity in the healthcare industry because the vast amounts of information generated by clinical and human lifestyle transactions are too complex to manage and analyze using conventional methods.[6] Given this, machine learning (ML) models have been used extensively in the prediction and management of healthcare challenges like diabetes. They are designed to predict the likelihood of an individual becoming diabetic based on a set of demographics, clinical and lifestyle factors.[7] These models have been utilized for a variety of tasks, such as the prediction of diabetic complications,[8] the exposure to early-stage diabetes, risk factors associated with the disease and the optimization of treatment plans.[9]

ML models have been developed to predict the likelihood of diabetes in patients based on information such as cholesterol level, blood pressure, demographic and socio-economic factors (e.g. age, income, etc.) and fruit intake among others.[10] These factors (or features) are often fed directly into ML models to learn their correlation with diabetes so that they can be used to predict it. While these methods have worked so far, they have not been able to uncover the hidden relationships between features which can significantly enhance predictive performance. Here, the research answers the question of how the correlation and relationship between features can impact the accuracy of predicting diabetes likelihood in patients using ML models. We believe the predictive power of ML-based diabetes prediction models can be enhanced by using the inter-relationships or correlations between input features. While many studies in the field of diabetes prediction utilize ML models, this study uniquely emphasizes the importance of considering the interplay between features and their correlations, rather than solely relying on individual feature importance. Traditional methods often treat features independently, overlooking potential dependencies and interactions among them. By investigating feature correlation, this approach acknowledges the complex nature of diabetes and recognizes that a holistic understanding of feature interactions may lead to more accurate predictive models. In contrast, other methods may focus solely on individual feature importance or use simpler modeling techniques that do not account for feature correlations. By explicitly addressing the impact of feature correlation on prediction accuracy, this study provides insights into how ML models can be optimized for diabetes prediction, potentially leading to more effective diagnostic and management strategies. Thus, by leveraging existing advances in this area, ML models were used in this work to uncover hidden patterns in human lifestyle to detect correlations between high-risk features, which contribute significantly to diabetes at an early stage for improved accuracy in diabetes prediction. The objective is to detect paired feature patterns that will enable timely intervention while improving patient outcomes. The specific contributions of this paper include the detection of high-risk factors through feature importance scoring and feature correlation and the uncovering of hidden personalized factors through ML models for improved diabetes prediction. The obtained correlated features serve as an enhancement of the original features thereby improving prediction performance. Experiments were performed on a dataset of 263,882 samples of anonymized patient information using five different ML models, decision tree (DT), neural networks (NN), random forest (RF), support vector machine (SVM) and extreme gradient boosting (XGB) to compare the predictive powers of these derived correlated feature pairs (CF) with those of the direct features. DT models, known for their interpretability, are utilized in understanding feature relationships. However, their susceptibility to overfitting could undermine the accuracy of predictions, especially in intricate feature correlations. NN excels in capturing complex feature relationships, which is crucial for understanding the nuanced interplay of features in diabetes prediction. Nevertheless, their requirement for extensive data and computational resources might pose challenges in fully exploiting feature correlations. RF models leverage ensemble learning to mitigate overfitting, thus offering robustness in handling correlated features. However, their computational demands might hinder their effectiveness in processing large datasets with intricate feature correlations. SVM models are adept at handling high-dimensional feature spaces, which is advantageous in capturing subtle correlations. Nonetheless, careful selection of kernel functions and hyperparameter tuning is essential to fully harness the benefits of feature relationships. XGB techniques are renowned for their efficiency in capturing intricate feature interactions, making them well-suited for exploiting correlations in diabetes prediction. However, their susceptibility to overfitting necessitates meticulous hyperparameter tuning to ensure optimal performance. Section 2 provides a sample of related works in the field of diabetes prognostication through ML techniques. Section 3 discusses the ML model and the processes employed in this research. In Sec. 4, sample experiments and the consequent results are presented. Section 5 concludes.

2. Related Works on Diabetes Risk Prediction

In more recent times, artificial intelligence has continued to revolutionize the health industry with progressive impacts. The authors in Ref. 11 conducted a detailed analysis of ML applications in healthcare, covering disease diagnosis, patient monitoring and healthcare management. A depression detection model was built by Islam *et al.*[12,13] used ML for hypertension prediction, while an ML model was developed by Yue *et al.*[14] for the purpose of diagnosing and predicting the prognosis of breast cancer. Similarly, an ML model was developed by Ibitoye and Nwosu[15] for analyzing sobriety and relapse in drug rehabilitation while a comparative analysis was conducted by Famutimi *et al.*[16] to evaluate the performance of single structure columnar in-memory and disk-resident data storage techniques using large healthcare datasets. However, when it comes to research works in diabetes, there are several types of diabetes prediction models, ranging from simple scoring systems that use a few clinical variables to complex ML models that integrate hundreds or thousands of variables.[17] Healthcare workers can utilize these models to identify individuals who are at high possibilities of developing diabetes, which can help guide targeted intervention aid in the prevention or postponement of the disease onset. Diabetes prediction using ML has been a popular research topic in current times. Various ML algorithms have been used to predict diabetes, including artificial neural networks (ANN), logistic regression (LR), RF, SVM and DT.[18] In Ref. 19, SVM was used to envisage the chances of diabetes on demographic and clinical features. The model attained an accuracy of 89.4% and an area under curve (AUC) of 0.932. A gradient-boosting model was used in Ref. 20 to predict diabetes in the Chinese population. The study recorded an accuracy of 82.68% and an area under the curve-receiver operating characteristic (AUC-ROC) of 0.87. By exploiting anthropometric and lifestyle factors, the authors of Ref. 21 used an LR model to calculate the risk of diabetes. The model accomplished an accuracy of 71% and an AUC of 0.763. Some studies have used demographic, clinical and lifestyle factors as predictors, while others have used genetic and biomarker data. With the use of a DT model, an accuracy of 81.6%, sensitivity of 79.2% and specificity of 82.7% were chronicled in the risk prediction of diabetic retinopathy using clinical and demographic features in Ref. 22. Using an ANN model in Ref. 23, an accuracy of 88.46% and an AUC of 0.954 were achieved when demographic and clinical features were used for diabetes risk prediction. Beyond ML, a study in Ref. 24 used a deep learning approach with a convolutional neural network (CNN). These models can be employed to forecast the likelihood of developing type 2 diabetes in the Chinese population. It achieved an AUC of 0.81, which outperformed traditional ML models. In Ref. 25, an ML model was developed to predict the probability of developing type 2 diabetes in the United Kingdom Biobank population. Their research used a dataset of over 500,000 individuals with various demographic, clinical and lifestyle factors. Several ML algorithms were then trained on this dataset to predict diabetes risk. The results showed that the ML model had high accuracy in predicting diabetes risk, ROC of 0.89. Similarly, Yan *et al.*[26] developed an ML model for predicting the risk of type 2 diabetes in a Chinese population. The ML algorithms employed include DT, LR and RF. The best AUC score of 0.83 was achieved by the RF algorithm, indicating good discrimination between individuals with and without diabetes. Similarly, in Ref. 27, an RF model was developed to predict the risk of diabetes using clinical and biochemical features. The model got an accuracy of 93.33% and an AUC of 0.992. In Ref. 28, the authors employed six ML algorithms to predict diabetes on a dataset of 403 instances and 11 attributes. The six ML algorithms achieved the following prediction accuracies, LR (69%), multi-layer perceptron (90.99%), SVM (92%), DT (96%), gradient boosting (97%) and RF, which had the best prediction accuracy of 98%. The authors of Ref. 29, used a dataset of 2056 de-identified patients which contained 50% type 2 diabetes patients and non-diabetic patients. They employed eight different ML algorithms and reported the superior performance of RF, which achieved an AUC score of 0.91. A fused ML model was developed in Ref. 30 to predict diabetes on a dataset of 520 examples. The fused model consists of ANN and SVM models whose predictions were fed into a fuzzy model which determines the final prediction. Using a 70:30 training/test split of the dataset, the authors reported a 94.87% prediction accuracy of the fused model which was better than those of the individual models with SVM achieving 89.1% and ANN achieving 92.31%. In Ref. 31, the authors used four tree-based classifiers (Gradient Boosting, Adaboost, DT and Extra Tress) to calculate the beginning of diabetes on the PIMA dataset[32] containing 768 instances. They found that the extra trees classifier achieved the best performance with an AUC of 96%. In this work, we employed a classification approach to predicting diabetes from personalized features. Using five different ML algorithms, we employed a binary classification approach and compared the performances of the algorithms on the diabetes dataset. This offers the advantage of providing direct prediction of the possibility of the disease and the opportunity to investigate the performances of the different algorithms before selecting an optimal one.

3. ML Models for Personalized Diabetic Risk Prediction

Generally, diabetes risk prediction models are valuable tools for recognizing personalities who have a great chance of developing diabetes before the beginning of symptoms. This is important because the earlier the identification, the faster the call for targeted interventions and lifestyle modification. Here, the diabetes risk

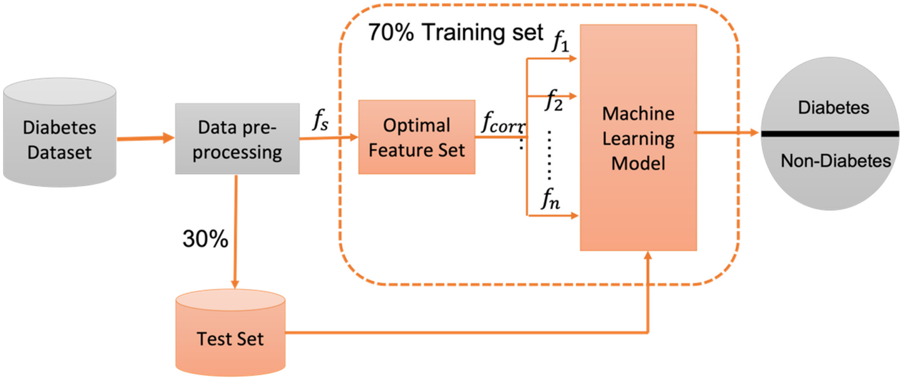

Fig. 1. Diabetic model for personalized risk prediction.

prediction models considered and used associated features in the dataset to classify the diabetic status of a patient. The procedure used in the model development is presented in Fig. 1 with an inherent description afterwards.

f_s is the recursive feature selection algorithm, f_{corr} defines Pearson's correlation coefficient, while $f_1, f_2, \ldots, f_n$ indicates the set of optimal features. Thus, from Fig. 1, the following processes are obtainable.

(1) **Data Collection:** Here a comprehensive dataset of anonymous records with a diverse range of information on behavioral lifestyles like Smoking, Glucose level, Blood Pressure, Income, Insulin, Cholesterol and demographical attributes like sex, and age were utilized to ensure reliability for analysis as the diabetics' dataset.

(2) **Preprocessing:** Initially, preliminary data pre-processing activities were conducted to handle missing values through mean imputation from the panda's library. Then, to address outliers, Tukey's method through the scipy library in python was adopted. Also, z-score normalization from sci-kit-learn was used to standardize the numerical variables. For further analysis, these procedures ensured the dataset was clean and well-prepared.

(3) **Feature Selection:** Using an automated feature algorithm, here, recursive feature elimination f_s, the most relevant features (predictors) from the dataset that have a strong association with the risk of diabetes were identified.

(4) **Correlation Analysis:** By using Pearson's correlation coefficient, f_{corr}, selected features were analysed to identify any strong correlations that may impact the accuracy of the model.

(5) **Model Training:** After preprocessing the data, the dataset is divided into a 70% training set and a 30% test set. The training set is then used to train the ML model using different algorithms, including DT, RF, XGBoost, SVM and NN. These algorithms enable the model to learn patterns and relationships within the data, allowing it to make predictions when presented with the test set.

Model Configurations: To enhance model performance and predictive accuracy, the model hyperparameters were fine-tuned. For NN, the optimal hyperparameter values are: Learning Rate: 0.015, Number of Epochs: 10 and Batch Size: 64. A feedforward NN architecture was employed with two hidden layers, dropout for regularization and batch normalization for improved convergence. The model was compiled with the Adam optimizer and binary cross-entropy loss for binary classification. Shifting the attention to XGBoost, the objective remains consistent: essential for maximizing its performance. Here, its optimal Learning Rate was 0.18, the Number of Estimators (Trees): was 80, and the Maximum Depth of Trees was adjusted randomly. DT parameters such as maximum depth, minimum samples split and minimum samples leaf are considered. The Number of Estimators (Trees): 100, Maximum Depth of Trees: None (or another reasonable value), Minimum Samples Split: 2 and Minimum Samples Leaf: 1 were the optimal fine-tuned hyperparameter values for RF. The SVM utilizes a C (Regularization Parameter) value of 1.0, and the radial basis function (RBF). The Gamma (Kernel Coefficient) was set to "auto" for RBF. All the models were trained on 70% of the dataset and tested with 30% of the dataset. The associated features delivered a comprehensive representation of the patterns and relationships in the data. With 56.5% of the attributes giving complementary information about the target variable, accuracy was enhanced when compared to using 100% features. However, the individual contributions of each feature to the classification decision are not reported in this instance. In Sec. 4, detailed experiments and algorithm evaluations are presented alongside exploratory information for consideration.

4. Experiments and Evaluations

From a dataset with 263,882 rows, and 23 columns with different lifestyles such as body-mass index (BMI), Age, Income, High blood pressure, education, exercise, smoker and more. The Python sci-kit-learn library was used to build five (5) distinct ML models from the top 15 pairs of associated features realized from 13 distinct features in the dataset. In Fig. 2, the feature importance scores are displayed in descending order. Preliminary investigation revealed that the independent importance of features does not correspond to the relationship existing between features, which impacts more on the performance of the predictive model. In Table 1, the top fifteen (15) paired associated features from the dataset are displayed with their respective scores. The table shows that high cholesterol and alcohol consumption are the most associated features with a correlation score of 0.99 while the general health (GenHealth) and BMI of a patient are the least associated pairs of features with a correlation score of 0.23.

To assess how well the model can predict the likelihood of diabetes founded on the given features, the trained model was appraised through standard performance metrics like accuracy, precision, recall and $F1$-score on the validation set. This step

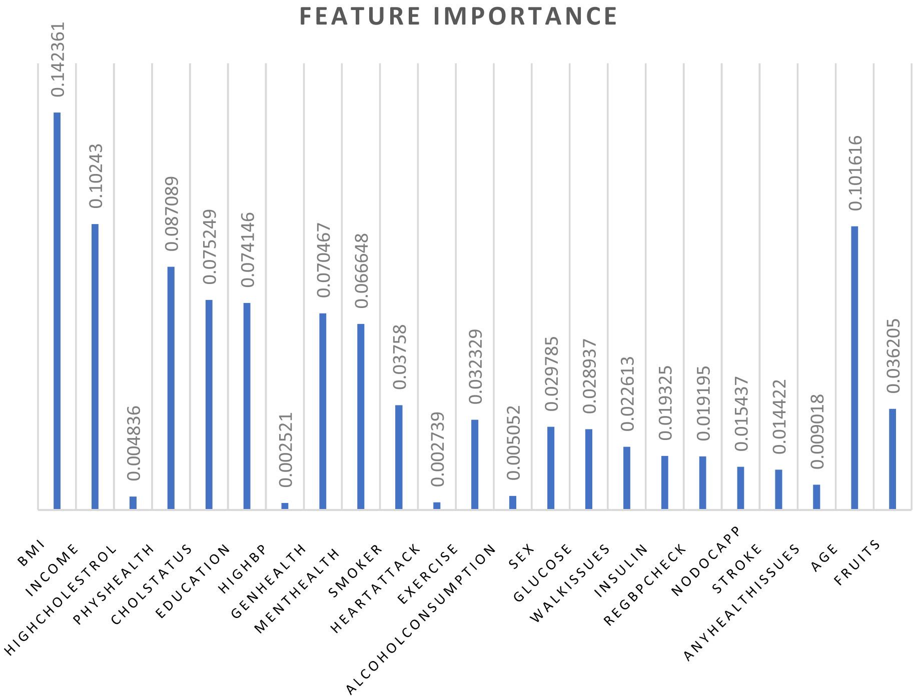

Fig. 2. Feature importance.

assesses how well the model can predict the risk based on the given features and the average result of each of the learning algorithms is presented in Table 2. As seen in Table 2, NN achieved the best average result of 93%, 91%, 86% and 87% recall, precision, accuracy and $F1$-score, respectively.

Table 1. Top 15 associated features contributing to diabetes from the dataset.

S/N	Paired Correlation Feature	Score
1.	High Cholestrol – Alcohol Consumption	0.99
2.	Heart Attack – High BP	0.87
3.	PhysHealth – GenHealth	0.52
4.	PhysHealth – Walking Issues	0.48
5.	Walking Issues – Gen Health	0.46
6.	Income – Education	0.45
7.	GenHealth – Income	0.37
8.	MentHealth – PhysHealth	0.35
9.	WalkIssues – Income	0.32
10.	MentHealth – GenHealth	0.30
11.	Education – GenHealth	0.28
12.	PhysHealth – Income	0.26
13.	Exercise – GenHealth	0.26
14.	Glucose – Fruits	0.25
15.	GenHealth – BMI	0.23

Table 2. ML model performance evaluation.

Model	Accuracy	Precision	Recall	$F1$-score
DT	0.82	0.80	0.77	0.74
NN	0.87	0.93	0.91	0.85
RF	0.84	0.86	0.87	0.83
SVM	0.76	0.73	0.68	0.69
XGboost	0.86	0.91	0.87	0.83

Overall, the accuracy of the average result of the associated features in the diabetic's classification achieved better results when compared to using all available features contributing to the targeted output. Hence, the predictive power of the model is enhanced just as the shared information between the selected features is highly relevant for the classification.

5. Discussion

First, from Table 2, understanding the performance metrics of these models is crucial for determining their effectiveness in identifying individuals at risk of diabetes. Now, let's delve into the specific implications for diabetic prediction: The NN model emerges as a standout performer, achieving the highest accuracy of 0.87. This signifies its ability to correctly predict both diabetic and non-diabetic cases, a critical aspect in diabetic prediction where misclassifications can have significant consequences. The NN model's precision of 0.93 emphasizes its accuracy in positive predictions, crucial for minimizing false positives in diabetic identification. Moreover, the NN model's recall of 0.91 indicates its proficiency in correctly identifying positive instances, a key factor in diabetic prediction where capturing all potential cases is imperative. The balanced $F1$-score of 0.85 for the NN model underscores its effectiveness, offering a well-rounded measure of precision and recall tailored to the specific needs of diabetic prediction. XGboost, with an accuracy of 0.86 and a precision of 0.91, proves to be a strong contender, especially in scenarios where a slightly lower recall (0.87) is acceptable. This model's overall $F1$-score of 0.83 suggests a balanced trade-off between precision and recall in the context of diabetic prediction. On the other hand, DT and RF models exhibit good accuracy but show

Table 3. ML model performance evaluation (associated features vs all features).

	Accuracy		Precision		Recall		$F1$-score	
Model	CF	All	CF	All	CF	All	CF	All
DT	0.82	0.67	0.80	0.71	0.77	0.74	0.74	0.65
NN	0.87	0.74	0.93	0.82	0.91	0.77	0.85	0.75
RF	0.84	0.69	0.86	0.78	0.87	0.74	0.83	0.72
SVM	0.76	0.66	0.73	0.69	0.68	0.66	0.69	0.64
XGboost	0.86	0.71	0.91	0.74	0.87	0.73	0.83	0.69

relatively lower precision and $F1$-scores. In the realm of diabetic prediction, where the focus is on identifying positive cases accurately, the NN and XGboost models outshine these alternatives. The SVM model, with its lower accuracy, precision, recall and $F1$-score, indicates limitations in its suitability for diabetic prediction in this instance. In a domain where robust predictive capabilities are crucial, the SVM model falls short compared to the NN and XGboost alternatives.

Subsequently from Table 3, in our investigation into diabetic risk prediction, a noteworthy observation emerged regarding the impact of feature pairs on model performance. Specifically, the analysis focused on distinguishing between CF and direct feature pairs (All) in the context of four ML models: DT, NN, RF and XGBoost.

Key Findings:

(1) **Consistent Performance Boost for CF:** Across multiple models, including NN, RF and XGBoost, a recurrent pattern of enhanced performance was observed when utilizing CF. This improvement manifested in higher accuracy, precision, recall and $F1$-score values.

(2) **Notable Differences in Precision and Recall:** The utilization of CF consistently resulted in substantially higher precision and recall values, indicating a superior balance between correctly predicted positive observations and overall positive cases. This finding implies that models trained on CF achieved a heightened ability to accurately identify positive instances.

(3) **Synergistic Information Contribution:** CF seemed to contribute synergistic information beneficial for the prediction task. The combined influence of these correlated features appeared to strengthen the models' predictive capabilities, showcasing a potential synergistic effect in the information they provided.

(4) **Potential Reduction in Redundancy:** The preference for CF might be associated with a reduction in redundancy within the data. This reduction likely facilitated models in capturing more distinctive and relevant information, contributing to improved generalization.

(5) **Model Sensitivity to Complex Relationships:** ML models employed, especially NN and ensemble methods like RF and XGBoost, demonstrated sensitivity to complex relationships between features. Correlated features, exhibiting intricate interactions, proved advantageous in capturing non-linear dependencies.

(6) **Potential Robustness to Noise:** Correlated features showed promise in offering a degree of robustness to noise within the data. In scenarios where noise or variability was present, the correlated relationships appeared to act as stabilizing factors, potentially enhancing the models' robustness.

In medical settings, minimizing false negatives (increasing recall) is often prioritized to avoid missing potential diabetic cases. Both the NN and XGboost models exhibit strong performance in diabetic prediction. The NN, in particular, emerges as the top-performing model across all evaluated metrics, showcasing its potential as a reliable tool in identifying individuals at risk of diabetes.

6. Conclusion and Future Work

While different factors contribute to the likelihood of diabetes in a person, the degree of contribution and impact do vary among individuals. To this end, this research identifies the mutually exclusive relationship between features while returning the top 15 correlated pairs from a dataset of 263,882 samples for prediction. These distinct features were used with five supervized machine-learning models. The results showed that these correlated pairs produce more accurate predictions of the likelihood of diabetes than the direct features with NN performing better than the other four ML models in all cases. This goes to show that the correlated features can indeed enhance ML-based diabetes prediction. The limitation of this work is engendered in the dynamic nature of the features, which requires constant tracking and monitoring if conducted in real-time. While the study effectively demonstrates the importance of considering feature correlations in predicting diabetes likelihood, the results may not apply to other datasets or populations. The dataset used in this study may not represent the diversity of demographics, lifestyles and healthcare systems found globally, limiting the external validity of the findings. Moreover, the study could benefit from a deeper exploration of the biological and clinical relevance of the identified feature pairs. Understanding the underlying mechanisms linking these features to diabetes could provide valuable insights for both prediction and intervention strategies. The future direction is to extend the associated features beyond a pair towards identifying personalized associated features that contribute massively to the diabetes profile of an individual. Further analyses, including considerations of model interpretability and validation on independent datasets, will be examined to enhance the models' applicability and trustworthiness in real-world diabetic prediction scenarios. Additionally, the findings can be validated using external datasets from different populations, domain-specific investigations to elucidate the biological and clinical significance of the identified feature pairs, enhance model interpretability through methods such as feature importance analysis, and explore the integration of additional data sources to improve predictive accuracy and broaden the scope of the analysis. This will also be more paramount if most of the features are trackable in real-time.

Acknowledgement

The corresponding author extends gratitude to the School of Computing and Mathematical Sciences, University of Greenwich, London for their invaluable support. Additionally, the authors acknowledge the Pattern Recognition and Artificial Intelligence Laboratory (PRAILAB) at the University of Ibadan, Nigeria, for their support during the formulation of the research idea.

ORCID

Ayodeji O. J. Ibitoye https://orcid.org/0000-0002-5631-8507
Joseph D. Akinyemi https://orcid.org/0000-0003-3121-4231
Olufade F. W. Onifade https://orcid.org/0000-0003-4965-5430

References

1. American Diabetes Association, “Standards of medical care in diabetes — 2021 abridged for primary care providers,” *Clin. Diabetes* **39**(1) (2021) 1–111.
2. A. Mishra and S. K. Mishra, Type 2 diabetes: An epidemic disease in India, *J. Nepal Med. Assoc.* **56**(210) (2017) 317–320.
3. https://www.mayoclinic.org/diseases-conditions/prediabetes/diagnosis-treatment/drc-20355284 Accessed on 2nd May, 2023.
4. I Federation, *International Diabetes 23 Federation, Idf Diabetes Atlas*, 10th edn., Brussels, Belgium: International diabetes federation, 2021.
5. T. Zhu, K. Li, P. Herrero and P. Georgiou, “Deep learning for diabetes: A systematic review,” *IEEE J. Biomed. Health Inf.* **25**(7) (2020) 2744–2757.
6. S. S. Nair and T. R. Mary, “Role of data mining in healthcare: A review”, *Int. J. Eng. Adv. Technol.* **8**(6S4) (2019) 394–399.
7. M. M. Christensen and R. L. Haupt, “Machine learning in diabetes research,” *Clin. Diabetes Endocrinol.* **5**(1) (2019) 1–8.
8. A. H. AlTabakhi, A. M. Hussain and H. M. Al-Angari, “Predictive modeling of diabetes-related complications: A systematic review of methods,” *J. Med. Syst.* **42**(7) (2018) 1–15.
9. A. Rawshani, A. Rawshani, S. Franzén, B. Eliasson, A. M. Svensson, M. Miftaraj and S. Gudbjörnsdottir, “Mortality and cardiovascular disease in type 1 and type 2 diabetes,” *N. Engl. J. Med.* **376**(15) (2018) 1407–1418.
10. R. Casanova, S. Saldana, S. L. Simpson, M. E. Lacy, A. R. Subauste, C. Blackshear and A. G. Bertoni, “Prediction of incident diabetes in the Jackson Heart Study using high-dimensional machine learning,” *PLoS One* **11**(10) (2016) e0163942.
11. A. Nayyar, L. Gadhavi and N. Zaman, Machine learning in healthcare: Review, opportunities and challenges. *Machine Learning and the Internet of Medical Things in Healthcare*, Academic Press, Chapter 2, 23–45. ISBN 9780128212295, https://doi.org/10.1016/B978-0-12-821229-5.00011-2.
12. S. M. S. Islam, A. Talukder, M. A. Awal, M. M. U. Siddiqui, M. M. Ahamad, B. Ahammed and R. Maddison, “Machine learning approaches for predicting hypertension and its associated factors using population-level data from three South Asian countries,” *Front. Cardiovasc. Med.* **9** (2022) 839379.
13. A. O. Ibitoye, R. F. Famutimi, D. O. Olanloye and E. Akioyamen, “User centric social opinion and clinical behavioural model for depression detection,” *Int. J. Intell. Inf. Syst.* **10** (2021) 69.
14. W. Yue, Z. Wang, H. Chen, A. Payne and X. Liu, “Machine learning with applications in breast cancer diagnosis and prognosis,” *Designs* **2**(2) (2018) 13.
15. A. O. Ibitoye and C. Nwosu, “A machine learning model for sobriety and relapse analysis IN drug rehabilitation,” *IJISCS, Int. J. Inf. Syst. Comput. Sci.* **5**(2) (2021) 93–99.
16. R. F. Famutimi, M. O. Oyelami, A. O. Ibitoye and O. M. Awoniran, “An empirical comparison of the performances of single structure columnar in-memory and disk-resident data storage techniques using healthcare big data,” *J. Big Data* **10**(1) (2023) 25.
17. H. Wu, J. Wu, Q. Yu, J. Yang and J. Wang, “Development of a machine learning model for predicting type 2 diabetes risk in the UK biobank population,” *Healthcare* **7**(3) (2019) 79, doi: 10.3390/healthcare7030079.

18. F. Atiquzzaman, S. A. S. M. Bari and M. A. Hossain, "A review of machine learning algorithms for diabetes prediction," *Diabetes Res. Clin. Pract.* **174** (2021) 108773, doi: 10.1016/j.diabres.2021.108773.
19. Y. Cui, W. Sun and S. Guo, "Application of support vector machine algorithm in the prediction of diabetes, *J. Med. Syst.* **45**(1) (2021) 6, doi: 10.1007/s10916-020-01713-7.
20. J. Yang, X. Liu and Y. Liu, "A machine learning approach for predicting the risk of type 2 diabetes in a Chinese population," *J. Med. Syst.* **44**(5) (2020) 1–9, doi: 10.1007/s10916-020-01589-9.
21. S. Ghosh and S. Pal, "Prediction of diabetes using logistic regression model," *Int. J. Sci. Res. Comput. Sci. Eng.* **6**(4) (2018) 12–15.
22. Y. Liu, M. Ksiazek and M. U. Akram, "A machine learning approach for predicting the risk of type 2 diabetes in a Chinese population," *PLoS One* **14**(7) (2019) e0219495, doi: 10.1371/journal.pone.0219495.
23. V. S. Mani, R. Venkatesan and V. Priya, "Prediction of diabetes using artificial neural network," *Int. J. Adv. Res. Comput. Sci. Softw. Eng.* **11**(5) (2021) 441–447.
24. R. Karim, M. R. Islam, M. A. R. Ahad and S. M. S. Islam, "Decision tree-based prediction of diabetic retinopathy utilizing demographic, anthropometric, and clinical data," *J. Med. Syst.* **42**(4) (2018) 71, doi: 10.1007/s10916-018-0902-2.
25. Y. Zhang, Y. Yeo and M. Loh, "Development of a machine learning model for predicting type 2 diabetes risk in the UK Biobank population," *Sci. Rep.* **11**(1) (2021) 1–11, doi: 10.1038/s41598-021-83252-4.
26. J. Yan, L. Li, Y. Wang, J. Li and H. Liu, "A gradient boosting model for diabetes prediction in a Chinese population," *PLoS One* **16**(3) (2021) e0248733, doi: 10.1371/journal.pone.0248733.
27. K. Sujatha and R. Jayaraman, "A random forest approach for diabetes diagnosis using clinical and biochemical features," *J. Ambient Intell. Humaniz. Comput.* **12**(6) (2021) 6175–6184, doi: 10.1007/s12652-020-03077-w.
28. S. S. Bhat, V. Selvam, G. A. Ansari, M. D. Ansari, Md H. Rahman, "Prevalence and early prediction of diabetes using machine learning in North Kashmir: A case study of District Bandipora," *Comput. Intell. Neurosci.* **2022** (2022) 2789760, doi: 10.1155/2022/2789760.
29. H. Lu *et al.*, "A patient network-based machine learning model for disease prediction: The case of type 2 diabetes mellitus," *Appl. Intell.* **52** (2022) 2411–2422, doi: 10.1007/s10489-021-02533-w.
30. U. Ahmed *et al.*, "Prediction of diabetes empowered with fused machine learning," *IEEE Access* **10** (2022) 8529–8538, doi: 10.1109/ACCESS.2022.3142097.
31. M. A. Hama-Saeed, Diabetes type 2 classification using machine learning algorithms with up-sampling technique," *Journal of Electrical Systems and Inf Technol* **10** (2023) 8, doi: 10.1186/s43067-023-00074-5.
32. PIMA Indians Diabetes Dataset, Available at: https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database?select=diabetes.csv, accessed on 26 May 2023.