

Comparison of diabetes pre-risk prediction models

Kuo Han^{#a}, Qi Sun^{#a}, Jiahao He^{#a}, Xinran Zhang^{#a}, Xin Cheng^a, Siqi Wang^a,
Miao Yu^a, Linrong Yuan^a, Zaibo Pan^a, Yijie Li^b, He Ren^{*a}

^aFaculty of Medical Instrumentation, Shanghai University of Medicine and Health Sciences, Shanghai, China 201318; ^bCollege of Information Technology, Shanghai Ocean University, Shanghai, China.

[#]Disclosure: Kuo Han, Qi Sun, Jiahao He, Xinran Zhang are the co-first authors.

* Corresponding author: renh@sumhs.edu.cn

ABSTRACT

In the field of medicine, disease prevention is more important than treatment. Diabetes, as one of the diseases that are harmful and have a large number of patients, the prediction of diabetes using learning models is an essential part of diabetes prevention and treatment in the future medical field. In this study, compound feature selection was used to screen out the eight features with the most predictive ability, and diabetes was predicted by six machine learning models and two deep learning models, and the final results were obtained as follows: the XGBoost classification had the best prediction performance, with an accuracy of 99.1%, a precision of 99.0%, a recall rate of 99.2%, an F1 score of 99.1%, and an AUC value of up to 0.991; CatBoost and LightGBM models have the next best performance. This is consistent with our previous findings. This study further confirms the value and potential of the XGBoost model for diabetes prediction, identifying superior feature selection methods compared to previous studies, and improving the predictive performance of the model while reducing model complexity when dealing with more complex data.

Keywords: Diabetes, machine learning, deep learning, XGBoost, chi-square test, mutual information method

1. INTRODUCTION

Diabetes mellitus is one of the largest chronic metabolic diseases in the world, causing millions of deaths globally each year due to various health complications [1]. The magnitude of its harm and the number of people suffering from it have made it a serious public health problem. Prolonged hyperglycaemia can lead to severe damage to several organs, posing a threat to vital body organs such as the eyes, kidneys, nerves, heart and blood vessels, and can be life-threatening [2-4]. Currently an estimated 537 million adults worldwide have diabetes and there are 3.8 million deaths from diabetes complications each year [5], and studies of diabetes prevalence projections for 2030 and 2045 estimate that the number of people diagnosed with diabetes is projected to rise to 10.2 per cent (578 million) of the global adult population by 2030, and continue to rise by 2045 to 10.9% (700 million) by 2045 [6]. China, the most populous country, has had the highest number of diabetics in the world for more than 40 years. It is expected to reach 164 million by 2030 and 175 million by 2045 [7]. According to the International Diabetes Federation, diabetes is projected to become the seventh leading cause of death worldwide by 2030, and the epidemic is escalating, placing a huge economic burden on socio-healthcare systems [8]. Therefore, it is crucial to develop intelligent systems that help healthcare professionals in diagnosis and decision support [6].

2. LITERATURE REVIEW

In recent years, with the development of artificial intelligence, machine learning model and deep learning model have demonstrated their powerful predictive ability and parallel processing ability to deal with a large number of variables, and gradually entered all walks of life. In the medical field, machine learning model has produced many achievements in disease prediction, and many studies have proved that machine learning can be used as an effective way to predict diabetes. Dritsas et al.[8] developed five machine learning models to predict diabetes, and trained them based on NHANES data, and concluded that CATBoost performed best, with an accuracy rate of 82.1% and an AUC of 0.83. Xu Ying et al.[9] established the GAN optimization model for diabetes prediction based on PIMMA data set, and the accuracy of two classification was 96.27%, the accuracy of three classification was 99.31%, and the AUC reached 0.9702. Wee, Boon Feng et al. [10] also trained five learning models based on PIMMA data sets, and obtained that the

Logistic regression model has the highest accuracy, which is 81%. Wee, Boon Feng et al.[3] Based on the data set from Kaggle, a series of machine models and integration methods were used to predict diabetes, and the accuracy of CatBoost model was as high as 95.4, and the AUC score was 0.99. H. Ahmed et al.[2] trained and tested a variety of machine learning models, and got the conclusion that random forest and KNN performed best. Qin Yifan et al.[11] trained a variety of machine learning models through PIMMA data sets, and also got a good performance of random forests, with an accuracy of 80.869. Feng Xin et al.[12] trained the ensemble classifier of EnRfRsK through PIMMA data set, which included random forest, radial support vector machine and KNN, and the accuracy rate was 88.89%. Shaukat Zain et al.[13] proposed a new machine learning method, and the accuracy on PIDD data set reached 98.58%, and ROC was 0.965. Modak, Sandip Kumar et al.[14] also obtained the highest prediction accuracy of random forest classifier based on PIMMA, which was 97.75. While the accuracy of shallow neural network (SNN) on DRP 2020 data set of Dritsas Elias et al.[15] reached 99.23%. Based on the previous study of our group, this study adjusted the feature selection method and integrated various learning models to predict diabetes, so as to select the most efficient diabetes prediction model.

3. MATERIAL METHOD

3.1 Data set

The original data set used in this study has a total of 211,833 data samples, including 25 indicators, and the data have been desensitized. The dataset supporting the conclusions of this article is available in the DRYAD repository [unique persistent identifier and hyperlink to dataset in <https://doi.org/10.5061/dryad.ft8750v>].

Furthermore, to ascertain the model's reliability, this study employed a local dataset for training purposes. The dataset comprises physical examination records from Jiading District Central Hospital in Shanghai, spanning a three-year period. It has been meticulously desensitized and subjected to ethical scrutiny, encompassing a substantial 556,497 data samples across 116 distinct physical examination indicators.

3.2 Method

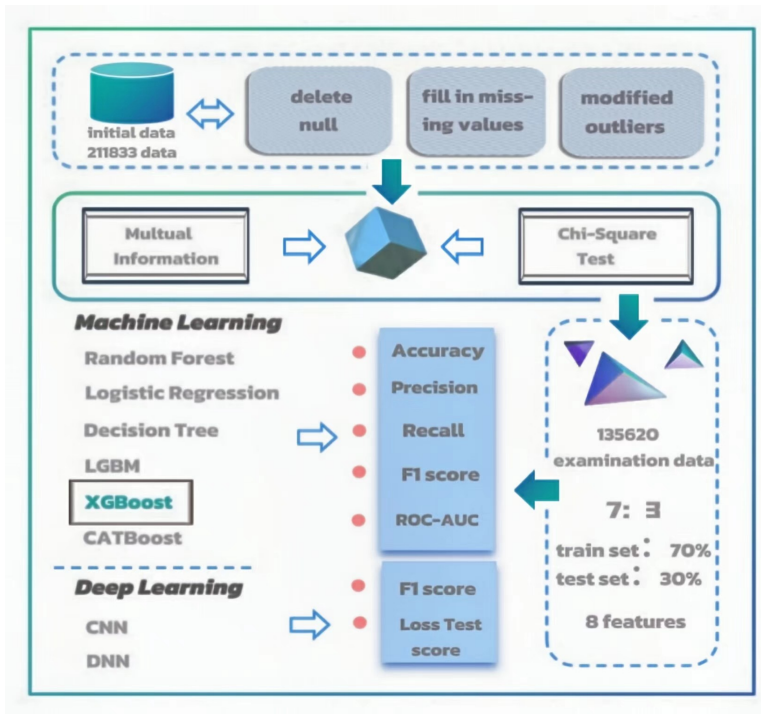


Figure 1. Research Process Framework

Figure 1 outlines the comprehensive methodology employed in this study, presented as a streamlined flowchart. The journey begins with rigorous data preprocessing, which encompasses the elimination of invalid data, the meticulous handling of abnormal values, and the strategic imputation of missing values to ensure data integrity. This foundation is

then strengthened through feature engineering, leveraging both the chi-square test and the mutual information method to meticulously select the most predictive features. Armed with the refined feature set, the study advances to the model training phase, where an array of six sophisticated machine learning models and a pair of deep learning models are deployed. The efficacy of these models is meticulously appraised through a suite of six pivotal metrics: accuracy, precision, recall, F1 score, ROC-AUC, and the loss function. This multifaceted evaluation offers a comprehensive assessment of their predictive prowess.

3.2.1 Data processing and feature engineering

In this study, we meticulously processed and curated features from a comprehensive dataset encompassing 211,833 data samples with 25 initial variables, culminating in the extraction of 135,620 refined training datasets characterized by 8 pivotal feature variables.

Our initial step involved rigorous data preprocessing, where we excluded samples with a missing data ratio exceeding 10%, yielding a refined dataset of 135,620 data points. Subsequently, we imputed the remaining missing values with the respective average, ensuring data integrity.

For feature selection, we initially eliminated 5 variables that were deemed irrelevant to our research objectives. We then leveraged the chi-square test [16-19], a statistical technique renowned for its efficacy in analyzing categorical data. The chi-square statistic escalates with sample size, making it an optimal choice for the preliminary feature screening phase of our study.

Subsequently, we employed the Mutual Information (MI) method [20-22], which assesses feature significance by quantifying the mutual information between each feature and the target variable. This metric serves as a robust measure of the interdependence between variables. Following the initial chi-square screening, which discarded features with minimal impact, we retained 16 features. Building upon this foundation, the MI method further refined our selection, identifying and retaining features with substantial influence on the target variable.

Ultimately, through a stringent correlation threshold of 0.7, we pruned highly correlated features, resulting in a final set of 8 features: Age, Height, Weight, SBP (Systolic Blood Pressure), DBP (Diastolic Blood Pressure), FPG (Fasting Plasma Glucose), ALT (Alanine Aminotransferase), and CCR (Creatinine Clearance Rate). The interrelations among these features are visually represented in a heat map, as depicted in Figure 2.

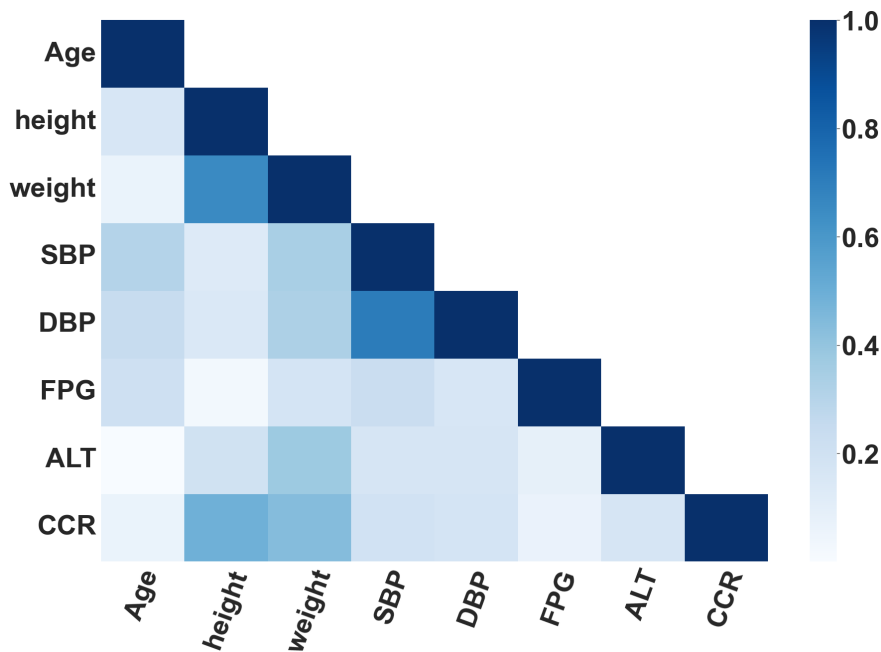


Figure 2. Characteristic correlation heat map

Guided by feature selection principles aimed at preventing overfitting, this study implemented a rigorous approach, setting a correlation threshold of 0.7 to eliminate features exhibiting strong inter-dependencies among the eight considered attributes. Subsequently, the refined feature set was visually represented through a distribution interval diagram, as depicted in Figure 3, offering a clear and insightful view of the final feature landscape.

Category imbalance [23], a prevalent challenge in machine learning, was encountered in the dataset derived from the aforementioned operations, where the count of normal samples significantly outweighed that of sick samples, with 133,685 normal instances compared to just 1,935 sick instances. To address this imbalance, SMOTE oversampling was strategically employed, resulting in balanced datasets of 133,685 instances for both normal and sick categories. This balanced representation facilitated more equitable model training and prediction, mitigating the effects of class imbalance.

In this study, the dataset is divided into training set and test set in 7:3 (70% training set and 30% test set), the test set is used to train the model and the test set is used to evaluate the model and to measure the model's ability to generalize over different instances.

The local dataset underwent identical processing, yielding a total of 22,284 samples characterized by 13 distinct features. These include pulse rate, blood pressure (systolic and diastolic), waist circumference, hemoglobin, white blood cell count, platelet count, glycated hemoglobin, serum alanine aminotransferase (ALT), serum glutamic oxaloacetic transaminase (AST), total bilirubin, serum creatinine, and urinary glucose levels. To address class imbalance, the Synthetic Minority Over-sampling Technique (SMOTE) was applied, resulting in an equitable distribution of 20,108 samples for both the control and case groups.

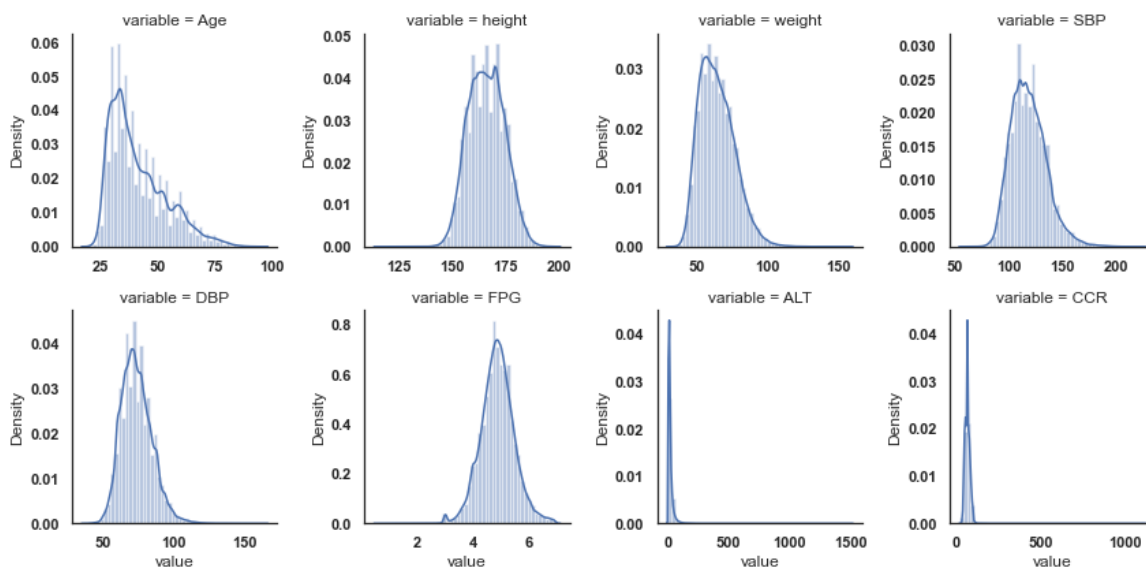


Figure 3. Feature distribution interval

3.2.2 Model construction and evaluation

In this comprehensive study, a comparative training approach was undertaken involving six machine learning models (Decision Tree, Logistic Regression, Random Forest, XGBoost, CatBoost, and LightGBM) and two deep learning models (Convolutional Neural Network, CNN, and Deep Neural Network, DNN). To assess the performance of the machine learning models, five pivotal classification evaluation metrics were employed: Accuracy, Precision, Recall, F1-score, and ROC-AUC. Meanwhile, the LOSS function served as the benchmark for evaluating the deep learning models. These diverse indicators collectively offer a nuanced and comprehensive analysis of the prediction capabilities of each model, illuminating their strengths and weaknesses from multiple angles.

XGBoost, an acronym for Extreme Gradient Boosting, is a sophisticated ensemble learning algorithm introduced by Tianqi Chen and Carlos Guestrin. This algorithm is distinguished by its meticulous optimization and exceptional scalability [24]. XGBoost's prowess lies in its innovative gradient boosting framework, its adept handling of the objective function's optimization, its nuanced approach to feature importance evaluation, and its adept management of

sparse data through efficient columnar and block-wise parallelization. These strengths solidify XGBoost's reputation as a leading solution for complex data challenges.

XGBoost builds upon the merits of traditional Gradient Boosting Decision Trees while significantly enhancing model training velocity and predictive accuracy through a suite of optimization techniques [25]. It refines the objective function optimization by employing a second-order Taylor expansion and capitalizes on the Column Block structure to streamline parallel data processing, thereby markedly improving computational efficiency. Additionally, XGBoost adeptly navigates model complexity with regularization parameters, including L1 and L2 regularization, effectively curbing overfitting. These robust features enable XGBoost to construct models that are both precise and resilient, adeptly handling intricate datasets with exceptional efficacy.

4. RESULT AND DISSCUSSION

4.1 Signature analysis

In this study, a dual-approach methodology was employed for feature selection. Initially, the chi-square test was applied to the pre-processed data to filter out features with negligible correlation. Subsequently, the mutual information (MI) technique was engaged to refine the selection, identifying features with significant correlation. The process culminated with the exclusion of features that exhibited a correlation coefficient exceeding the threshold of 0.7. To mitigate the effects of data imbalance, the dataset underwent oversampling, and the data was partitioned to facilitate model training.

Table 1. Single-factor and multifactor regression analysis

Variable	Univariable					Multivariable				
	coef	se	t	P> t	95%CI ¹	coef	se	t	P> t	95%CI ¹
Age	0.0010	2.6e-05	38.869	0.000	-0.030,-0.025	0.0951	0.097	0.985	0.327	-0.097,0.287
height	0.0003	3.87e-05	8.799	0.000	0.000,0.000	0.0037	0.095	0.039	0.969	-0.185,0.353
weight	0.0010	2.61e-05	39.309	0.000	0.001,0.001	0.1474	0.103	1.425	0.158	-0.058,0.353
SBP	0.0007	1.97e-05	34.010	0.000	0.001,0.001	0.0157	0.114	0.138	0.890	-0.210,0.242
DBP	0.0009	2.97e-05	29.821	0.000	0.001,0.001	-0.0753	0.095	-0.796	0.428	-0.263,0.113
FPG	0.0400	0.001	77.105	0.000	0.039,0.041	0.6697	0.101	6.623	0.000	0.469,0.871
ALT	0.0004	1.45e-05	27.652	0.000	0.000,0.000	0.1398	0.101	1.391	0.168	-0.060,0.340
CCR	0.0002	2.14e-05	8.416	0.000	0.000,0.000	-0.0185	0.107	-0.174	0.863	-0.231,0.194

CI=Confidence Interval, ²*p<0.05,**p<0.01,***p<0.001

Table 1 presents the results of both univariate and multivariate regression analyses for the identified characteristics. Notably, in the multivariate regression analysis, the P-value associated with fasting plasma glucose (FPG) is markedly below the threshold of 0.005, underscoring its substantial influence on diabetes. FPG, derived from direct measurement of plasma glucose levels, is grounded in extensive demographic data and serves as a critical and widely utilized metric for assessing glucose metabolism. The FPG measurement is a direct reflection of current blood glucose levels and is pivotal for diagnosing metabolic glucose disorders. It is important to note that the accuracy of FPG testing can be influenced by a spectrum of factors, including hepatic function, hormonal regulation, and neurological elements.

Furthermore, the P-values for fasting plasma glucose (FPG), body weight, age, and alanine aminotransferase (ALT) are notably lower, aligning with the feature importance rankings depicted in Figure 4. The t-values of these parameters in the

univariate regression analysis corroborate their substantial influence on diabetes risk, echoing findings from the study by Lara Lama et al. [26]. ALT, synthesized and secreted by hepatocytes, serves as a key biomarker for liver function, and its levels can modulate FPG concentrations. Obesity, a prevalent symptom across numerous pathologies, is also reflected in body weight, which is a component of the body mass index (BMI). An elevated body weight, indicative of obesity, exerts a significant effect on FPG levels. Excess adipose tissue can release inflammatory cytokines, impairing the functionality of pancreatic beta cells. Concurrently, elevated blood lipids can induce insulin resistance, precipitating a cascade of metabolic disorders such as hypertension and hyperlipidemia, which in turn can disrupt blood glucose regulation.

Figure 5 elucidates the diabetes machine learning investigation through the lens of SHAP (SHapley Additive exPlanations) values [27]. These SHAP values serve as a mechanism to interpret the predictive outcomes of the machine learning model, providing a quantifiable measure of each feature's influence on the model's diabetes diagnosis. Notably, fasting plasma glucose (FPG) exhibits a robust positive correlation with diabetes, signifying its substantial risk factor status. Concurrently, diastolic blood pressure (DBP) and alanine aminotransferase (ALT) are also identified as having significant correlations with elevated diabetes risk. DBP, a metric of arterial vascular compliance, is pivotal in the assessment of cardiovascular health, a domain closely intertwined with metabolic conditions that can influence blood glucose dynamics. This correlation aligns with the prior feature importance findings, substantiating the precision and broad applicability of the employed feature selection techniques and predictive models in diabetes risk assessment.

In this study, the selection process meticulously curated the features by eliminating those with low and high correlations. This strategic refinement has bolstered the predictive prowess of the model, simultaneously enhancing its robustness and its capacity for generalization across diverse datasets.

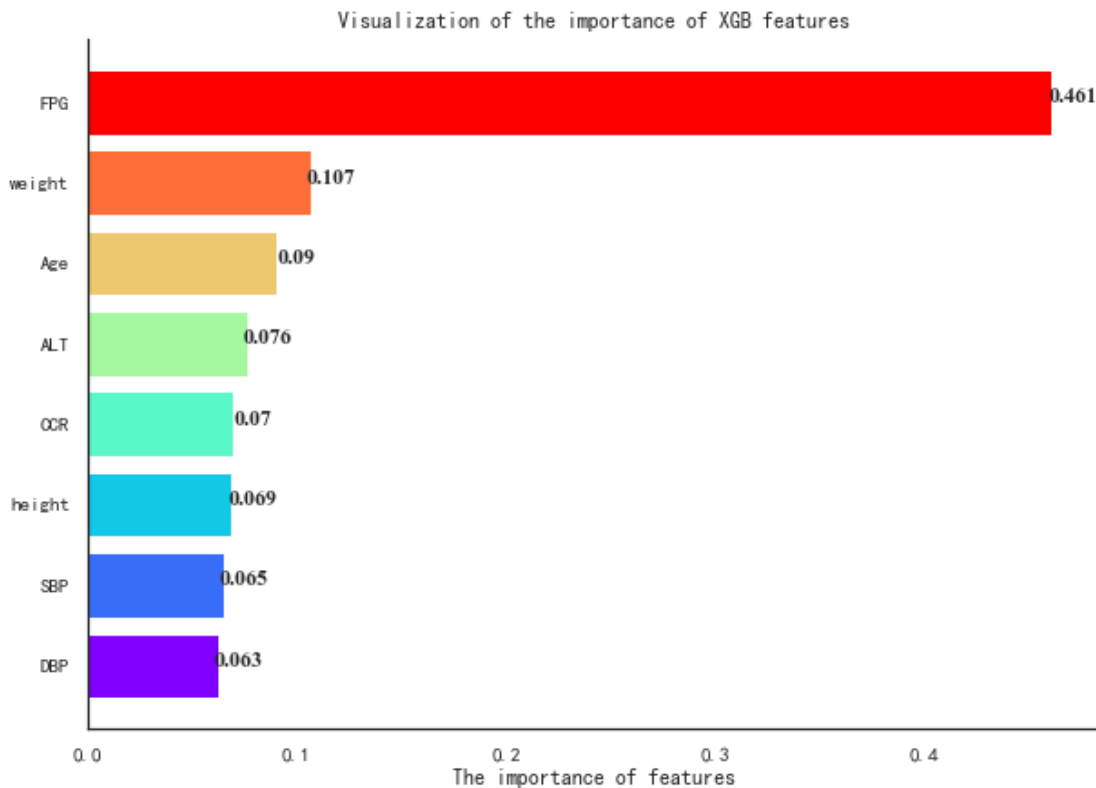


Figure 4. XGBoost feature importance ranking

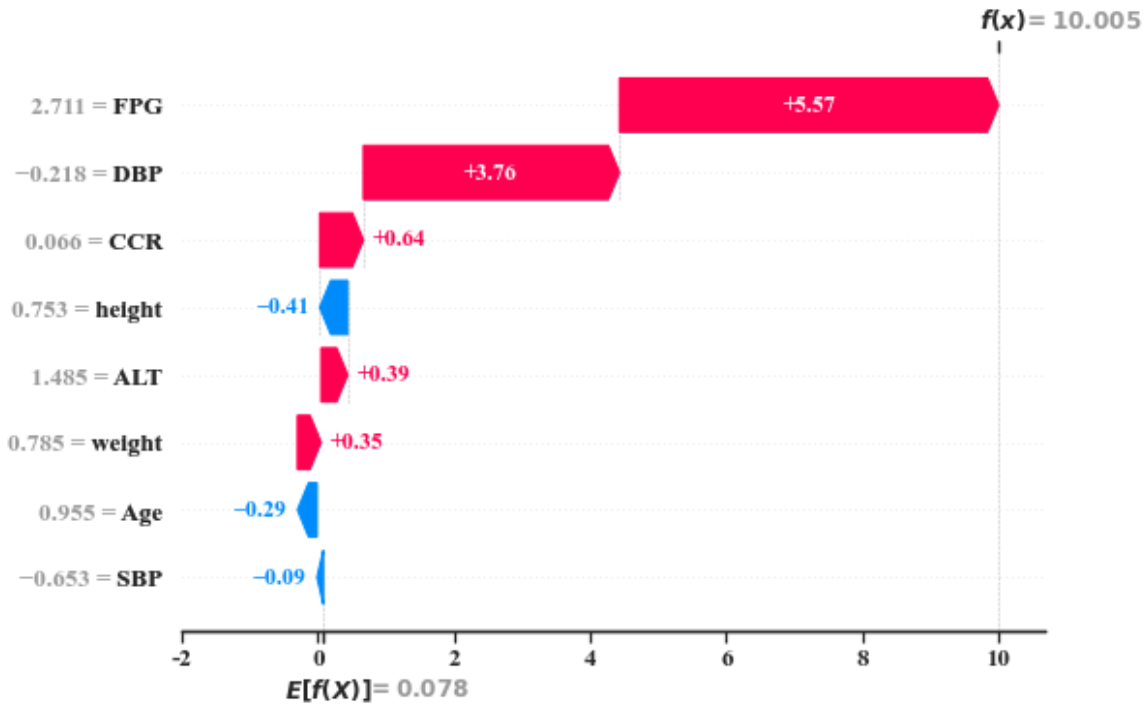


Figure 5. SHAP waterfall chart explaining XGBoost

4.2 Research results

This study conducts a comprehensive performance assessment of eight predictive models, scrutinizing five key classification metrics. The conclusive findings demonstrate that the XGBoost model outperforms its counterparts, exhibiting remarkable efficacy and resilience in the prediction of diabetes. A detailed account of the model's performance, alongside the evaluative metrics, is meticulously documented in Table 2.

Table 2. Multi-model comparison results on Dryad dataset

	Accuracy	Precision	Recall	F1-score	ROC-AUC	LOSS
XGBoost	99.1%	99.0%	99.2%	99.1%	0.991	—
LightGBM	99.1%	99.4%	98.8%	99.1%	0.991	—
CatBoost	99.0%	99.2%	98.8%	99.0%	0.990	—
RF	97.5%	96.1%	99.0%	97.5%	0.975	—
DT	91.3%	90.1%	92.9%	91.4%	0.913	—
LR	84.8%	83.7%	86.4%	85.0%	0.848	—
CNN	84.9%	—	—	—	—	0.373
DNN	86.7%	—	—	—	—	0.334

Table 2 offers a clear insight into the performance disparities among various models, with tree-based machine learning models, particularly XGBoost, LightGBM, and CatBoost, outshining the rest. The underwhelming performance of the deep learning model may stem from its relatively low complexity, failing to capture the intricacies of the data. Among the top performers, XGBoost emerges as the clear winner, boasting exceptional metrics and robustness. With an accuracy rate of 99.1%, precision of 99.0%, recall of 99.2%, F1 score of 99.1%, and an AUC value soaring at 0.991, XGBoost demonstrates unparalleled classification prowess. CatBoost and LightGBM follow closely, with LightGBM achieving the highest accuracy among the three. The classification capabilities of the machine learning models employed in this study, along with their impressive evaluation metrics, solidify the soundness of our feature selection method and underscore the reliability of XGBoost for diabetes prediction.

Table 3. Multi-model comparison results on local datasets

	Accuracy	Precision	Recall	F1-score	ROC-AUC	LOSS
XGBoost	96.9%	96.5%	97.3%	96.9%	0.969	—
LightGBM	96.8%	97.0%	96.7%	96.8%	0.968	—
CatBoost	96.6%	96.2%	96.9%	96.6%	0.966	—
RF	95.1%	94.6%	95.6%	95.1%	0.951	—
DT	88.1%	84.8%	92.9%	88.6%	0.881	—
LR	90.2%	89.8%	90.6%	89.8%	0.902	—
CNN	89.9%	—	—	—	—	0.310
DNN	86.2%	—	—	—	—	0.318

Table 3 presents the statistical outcomes of the validation experiments utilizing the Jiading Hospital dataset, revealing a consistency with the results depicted in Table 2. The trio of XGBoost, LightGBM, and CatBoost models continue to demonstrate superior performance, with the RF, LR, and DT models trailing in succession. Comparatively, the deep learning models have shown an improvement in performance over the prior results, yet on average, machine learning models still outperform them. Notably, the XGBoost model stands out with its robust and stable results, boasting an accuracy of 96.9%, a precision of 96.5%, a recall of 97.3%, an F1 score of 96.9%, and an AUC value of 0.969. These findings from this dataset reaffirm the prowess of the XGBoost model in the realm of diabetes prediction and classification.

5. CONCLUSION

This study conducted a comparative analysis of various classifiers, encompassing six machine learning and two deep learning models, to assess their efficacy in the early classification of diabetes. The findings underscore the XGBoost model's distinct edge across several pivotal classification metrics, eclipsing the benchmarks set by prior research. Furthermore, the study leverages both publicly accessible and local datasets for validation, where the XGBoost model consistently delivers exceptional performance. By diminishing reliance on conventional blood tests for diagnosis, this approach can preemptively pinpoint at-risk diabetes groups, markedly enhancing the precision of diabetes diagnosis. This advancement holds profound implications for the early intervention and management of the disease.

REFERENCES

- [1] U. Ahmed et al., "Prediction of Diabetes Empowered With Fused Machine Learning," IEEE Access, vol. 10, pp. 8529-8538, 2022, doi: 10.1109/ACCESS.2022.3142097.

- [2] E. Dritsas and M. Trigka, "Data-Driven Machine-Learning Methods for Diabetes Risk Prediction," *Sensors*, vol. 22, no. 14, p. 5304, 2022. [Online]. Available: <https://www.mdpi.com/1424-8220/22/14/5304>.
- [3] S. K. S. Modak and V. K. Jha, "Diabetes prediction model using machine learning techniques," *Multimedia Tools and Applications*, vol. 83, no. 13, pp. 38523-38549, 2024/04/01 2024, doi: 10.1007/s11042-023-16745-4.
- [4] C. Olisah, O. Adeleye, L. Smith, and M. Smith, "A Robust Machine Learning Framework for Diabetes Prediction," Cham, 2022: Springer International Publishing, in *Proceedings of the Future Technologies Conference (FTC) 2021, Volume 2*, pp. 775-792.
- [5] H. Ahmed, E. M. G. Younis, and A. A. Ali, "Predicting Diabetes using Distributed Machine Learning based on Apache Spark," in *2020 International Conference on Innovative Trends in Communication and Computer Engineering (ITCE)*, 8-9 Feb. 2020 2020, pp. 44-49, doi: 10.1109/ITCE48509.2020.9047795.
- [6] B. F. Wee, S. Sivakumar, K. H. Lim, W. K. Wong, and F. H. Juwono, "Diabetes detection based on machine learning and deep learning approaches," *Multimedia Tools and Applications*, vol. 83, no. 8, pp. 24153-24185, 2024/03/01 2024, doi: 10.1007/s11042-023-16407-5.
- [7] Y. Xu et al., "Performance of different machine learning algorithms in identifying undiagnosed diabetes based on nonlaboratory parameters and the influence of muscle strength: A cross-sectional study," *Journal of Diabetes Investigation*, vol. 15, no. 6, pp. 743-750, 2024, doi: <https://doi.org/10.1111/jdi.14166>.
- [8] Y. Qin et al., "Machine Learning Models for Data-Driven Prediction of Diabetes by Lifestyle Type," *International Journal of Environmental Research and Public Health*, vol. 19, no. 22, p. 15027, 2022. [Online]. Available: <https://www.mdpi.com/1660-4601/19/22/15027>.
- [9] X. Feng, Y. Cai, and R. Xin, "Optimizing diabetes classification with a machine learning-based framework," *BMC Bioinformatics*, vol. 24, no. 1, p. 428, 2023/11/13 2023, doi: 10.1186/s12859-023-05467-x.
- [10] Z. Shaukat et al., "Revolutionizing Diabetes Diagnosis: Machine Learning Techniques Unleashed," *Healthcare*, vol. 11, no. 21, p. 2864, 2023. [Online]. Available: <https://www.mdpi.com/2227-9032/11/21/2864>.
- [11] E. N. H. Kirğil, B. Erkal, and T. E. Ayyıldız, "Predicting Diabetes Using Machine Learning Techniques," in *2022 International Conference on Theoretical and Applied Computer Science and Engineering (ICTASCE)*, 29 Sept.-1 Oct. 2022 2022, pp. 137-141, doi: 10.1109/ICTACSE50438.2022.10009726.
- [12] B. Amma N.G, "En-RfRsk: An ensemble machine learning technique for prognostication of diabetes mellitus," *Egyptian Informatics Journal*, vol. 25, p. 100441, 2024/03/01/ 2024, doi: <https://doi.org/10.1016/j.eij.2024.100441>.
- [13] Y. Singh and M. Tiwari, "Revolutionizing Diabetes Disease Prediction Through Novel Machine Learning Techniques," *Nano*, vol. 19, no. 04, p. 2350056, 2024/04/01 2023, doi: 10.1142/S179329202350056X.
- [14] S. S. Bhat, V. Selvam, G. A. Ansari, and M. D. Ansari, "Analysis of Diabetes mellitus using Machine Learning Techniques," in *2022 5th International Conference on Multimedia, Signal Processing and Communication Technologies (IMPACT)*, 26-27 Nov. 2022 2022, pp. 1-5, doi: 10.1109/IMPACT55510.2022.10029058.
- [15] Q. A. Al-Haija, M. Smadi, and O. M. Al-Bataineh, "Early Stage Diabetes Risk Prediction via Machine Learning," Cham, 2022: Springer International Publishing, in *Proceedings of the 13th International Conference on Soft Computing and Pattern Recognition (SoCPaR 2021)*, pp. 451-461.
- [16] U. Das, A. Y. Srizon, M. A. Islam, D. S. Tonmoy, and M. A. M. Hasan, "Prognostic Biomarkers Identification for Diabetes Prediction by Utilizing Machine Learning Classifiers," in *2020 2nd International Conference on Sustainable Technologies for Industry 4.0 (STI)*, 19-20 Dec. 2020 2020, pp. 1-6, doi: 10.1109/STI50764.2020.9350498.
- [17] V. Rupapara, F. Rustam, A. Ishaq, E. Lee, and I. Ashraf, "Chi-Square and PCA Based Feature Selection for Diabetes Detection with Ensemble Classifier," *Intelligent Automation & Soft Computing*, vol. 36, no. 2, pp. 1931--1949, 2023. [Online]. Available: <http://www.techscience.com/iasc/v36n2/51102>.
- [18] R. Huang and H. Cui, "Consistency of chi-squared test with varying number of classes," *Journal of Systems Science and Complexity*, vol. 28, no. 2, pp. 439-450, 2015/04/01 2015, doi: 10.1007/s11424-015-3051-2.
- [19] C. Shen, S. Panda, and J. T. Vogelstein, "The Chi-Square Test of Distance Correlation," (in eng), *J Comput Graph Stat*, vol. 31, no. 1, pp. 254-262, 2022, doi: 10.1080/10618600.2021.1938585.
- [20] N. Hoque, D. K. Bhattacharyya, and J. K. Kalita, "MIFS-ND: A mutual information-based feature selection method," *Expert Systems with Applications*, vol. 41, no. 14, pp. 6371-6385, 2014/10/15/ 2014, doi: <https://doi.org/10.1016/j.eswa.2014.04.019>.
- [21] G. Chen, W. Qiu, S. Xia, and L. Wang, "Investigating Key Genes in Type 2 Diabetes Mellitus via Combining mAP-KL and Mutual Information Network," *Current Bioinformatics*, vol. 12, no. 5, pp. 416-422, 2017, doi: <http://dx.doi.org/10.2174/1574893611666160916171028>.

- [22] W. Gao, L. Hu, and P. Zhang, "Feature Selection by Maximizing Part Mutual Information," presented at the SPML '18: 2018 International Conference on Signal Processing and Machine Learning, Shanghai China, 2018-11, 2018. [Online]. Available: <http://dx.doi.org/10.1145/3297067.3297068>.
- [23] A. Roy, V. Iosifidis, and E. Ntoutsi, "Multi-fairness Under Class-Imbalance," Cham, 2022: Springer Nature Switzerland, in *Discovery Science*, pp. 286-301.
- [24] A. Paleczek, D. Grochala, and A. Rydosz, "Artificial Breath Classification Using XGBoost Algorithm for Diabetes Detection," *Sensors*, vol. 21, no. 12, p. 4187, 2021. [Online]. Available: <https://www.mdpi.com/1424-8220/21/12/4187>.
- [25] A. Prabha, J. Yadav, A. Rani, and V. Singh, "Design of intelligent diabetes mellitus detection system using hybrid feature selection based XGBoost classifier," *Computers in Biology and Medicine*, vol. 136, p. 104664, 2021/09/01/2021, doi: <https://doi.org/10.1016/j.combiomed.2021.104664>.
- [26] L. Lama et al., "Machine learning for prediction of diabetes risk in middle-aged Swedish people," *Heliyon*, vol. 7, no. 7, p. e07419, 2021/07/01/2021, doi: <https://doi.org/10.1016/j.heliyon.2021.e07419>.
- [27] E. Shakeri, T. Crump, E. Weis, R. Souza, and B. Far, "Using SHAP Analysis to Detect Areas Contributing to Diabetic Retinopathy Detection," in *2022 IEEE 23rd International Conference on Information Reuse and Integration for Data Science (IRI)*, 9-11 Aug. 2022 2022, pp. 166-171, doi: [10.1109/IRI54793.2022.00046](https://doi.org/10.1109/IRI54793.2022.00046).