Personalized Diabetes Risk Assessment Through Multifaceted Analysis (PD- RAMA): A Novel Machine Learning Approach to Early Detection and Management of Type 2 Diabetes

Gharbi Alshammari [1†]

Department of Information and Computer Science, College of Computer Science and Engineering, University of Ha'il, Ha'il 81481, Saudi Arabia

Abstract
The alarming global prevalence of Type 2 Diabetes Mellitus (T2DM) has catalyzed an urgent need for robust, early diagnostic methodologies. This study unveils a pioneering approach to predicting T2DM, employing the Extreme Gradient Boosting (XGBoost) algorithm, renowned for its predictive accuracy and computational efficiency. The investigation harnesses a meticulously curated dataset of 4303 samples, extracted from a comprehensive Chinese research study, scrupulously aligned with the World Health Organization's indicators and standards. The dataset encapsulates a multifaceted spectrum of clinical, demographic, and lifestyle attributes. Through an intricate process of hyperparameter optimization, the XGBoost model exhibited an unparalleled best score, elucidating a distinctive combination of parameters such as a learning rate of 0.1, max depth of 3, 150 estimators, and specific colsample strategies. The model's validation accuracy of 0.957, coupled with a sensitivity of 0.9898 and specificity of 0.8897, underlines its robustness in classifying T2DM. A detailed analysis of the confusion matrix further substantiated the model's diagnostic prowess, with an F1-score of 0.9308, illustrating its balanced performance in true positive and negative classifications. The precision and recall metrics provided nuanced insights into the model's ability to minimize false predictions, thereby enhancing its clinical applicability. The research findings not only underline the remarkable efficacy of XGBoost in T2DM prediction but also contribute to the burgeoning field of machine learning applications in personalized healthcare. By elucidating a novel paradigm that accentuates the synergistic integration of multifaceted clinical parameters, this study fosters a promising avenue for precise early detection, risk stratification, and patient-centric intervention in diabetes care. The research serves as a beacon, inspiring further exploration and innovation in leveraging advanced analytical techniques for transformative impacts on predictive diagnostics and chronic disease management.

Keywords:
Machine learning, Diabetes, Health care, Artificial intelligence.

1. Introduction

Diabetes Mellitus, particularly Type 2 Diabetes Mellitus (T2DM), has emerged as one of the most pressing public health challenges of the 21st century [1]. The global importance of this chronic metabolic disorder lies in its pervasive reach and multifaceted impact on individuals and healthcare systems [2]. The significance of T2DM transcends mere statistical figures, manifesting in its profound effect on the quality of life [3], healthcare costs, and mortality rates [4]. As the incidence of T2DM continues to surge, it poses a considerable burden on healthcare infrastructure, necessitating early detection, effective management, and preventive strategies. The utilization of machine learning models, as explored in recent studies exemplifies innovative approaches to address this global health concern [5].

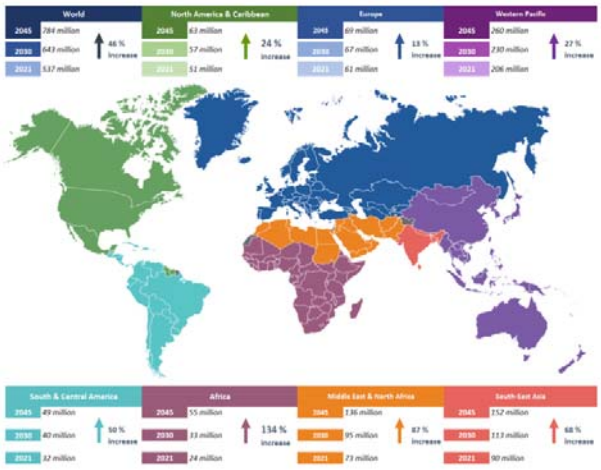

Figure 1 Number of diabetics by region (Source: www.ncbi.nlm.nih.gov)

T2DM's prevalence has reached epidemic proportions, as evident from Figure 1, it affects over 400 million individuals worldwide. Its growth trajectory is particularly alarming in developing countries, including China, where rapid urbanization, lifestyle changes, and genetic predispositions have contributed to a spike in cases. According to the World

Manuscript received August 5, 2023
Manuscript revised August 20, 2023
https://doi.org/10.22937/IJCSNS.2023.23.8.3

Health Organization, T2DM is projected to become the seventh leading cause of death by 2030, emphasizing the urgent need for robust diagnostic and therapeutic interventions. The complexity of T2DM lies in its myriad complications that affect various organ systems. Chronic hyperglycemia, the hallmark of diabetes, leads to microvascular and macrovascular complications. Microvascular complications include nephropathy, retinopathy, and neuropathy, affecting the kidneys, eyes, and peripheral nerves, respectively. Macrovascular complications encompass coronary artery disease, stroke, and peripheral arterial disease, significantly elevating the risk of cardiovascular mortality.

Moreover, the intricate relationship between T2DM and other comorbid conditions, such as obesity and hypertension, further complicates its management. These complications often lead to a reduced life expectancy and increased healthcare expenditure, underscoring the critical need for precise and early diagnosis. The research presented in this paper builds upon the existing body of knowledge, employing advanced machine learning techniques to enhance the prediction and classification of T2DM. By focusing on a comprehensive Chinese dataset and leveraging the Extreme Gradient Boosting (XGBoost) model, this study contributes to the evolving landscape of personalized medicine and proactive healthcare strategies in diabetes care.

1.1 Machine Learning in Healthcare

Machine learning (ML) represents a transformative paradigm in contemporary healthcare, orchestrating a seamless amalgamation of computational intelligence, statistical modeling, and clinical acumen. This burgeoning field transcends traditional analytical boundaries, enabling unprecedented advancements in data-driven decision-making, predictive diagnostics, personalized therapeutics, and patient-centric management.

ML's versatility has manifested across a multitude of medical domains, ranging from early cancer detection to heart disease prediction, from drug discovery to robotic surgeries. Its adaptive algorithms, capable of learning complex patterns from vast and heterogeneous datasets, have redefined the landscape of medical research and practice. In the context of T2DM prediction, ML has emerged as a potent tool, unlocking new horizons for early detection, risk stratification, and targeted interventions. Recent systematic surveys and meta-analyses [1] have underscored the remarkable efficacy of various ML models, including decision trees (DT), neural networks (NN), and ensemble techniques, in predicting T2DM. These studies, embracing diverse datasets and methodologies, have shed light on the intricate interplay of genetic, metabolic, lifestyle, and environmental factors in diabetes pathogenesis.

Despite these strides, a conspicuous gap persists in the literature, particularly concerning the application of advanced models like Extreme Gradient Boosting (XGBoost) on specific population datasets, such as the Chinese dataset used in this study. Addressing this lacuna and harnessing the distinct strengths of XGBoost could unveil novel insights, refine predictive accuracy, and foster innovative approaches to T2DM management.

1.2 Objectives of the Study

The overarching objective of this study is to innovate and validate an XGBoost-based predictive framework for the early detection, precise classification, and comprehensive understanding of T2DM.

In addition, to meticulously analyze and preprocess a comprehensive Chinese research dataset, encompassing a rich array of demographic, clinical, and lifestyle features, in alignment with World Health Organization standards. Furthermore, to architect, train, and optimize an XGBoost model, employing rigorous hyperparameter tuning, cross-validation, and comparative analyses with traditional ML classifiers. This phase seeks to elucidate the model's superior predictive prowess and robustness. To systematically evaluate and interpret the importance of various features in T2DM prediction, uncovering key risk factors, their interrelations, and implications for personalized medicine and preventive healthcare. To extrapolate the XGBoost model's potential in real-world clinical settings, facilitating timely intervention, patient-specific risk stratification, and integrative diabetes care. This aspect also encompasses an exploration of future research directions, potential challenges, and the broader impact of the study's findings on the evolving field of medical informatics and chronic disease management. Through these multifaceted objectives, the study aspires to contribute a seminal perspective to the nexus of machine learning,

medical science, and healthcare innovation, forging a trail for future research endeavors and clinical applications.

2. Literature Review

The burgeoning field of machine learning (ML) has metamorphosed the landscape of medical diagnostics, with a particular resonance in the early detection and prediction of Type 2 Diabetes Mellitus (T2DM). Various ML models have been meticulously explored for T2DM prediction, reflecting a rich tapestry of computational approaches:

With their intuitive structure and interpretability, DT models have been employed extensively. A meta-analysis conducted by unveiled a pooled accuracy of 0.88 for DT models in T2DM prediction. NNs, inspired by biological neural systems, have demonstrated remarkable flexibility and adaptability. It reported a pooled accuracy of 0.85 for NN models [6]. SVM models have been recognized for their robustness in handling high-dimensional data, contributing to precise classification [7]. RF and ensemble methods, which combine multiple models, have emerged as powerful tools for enhancing predictive accuracy [8].

Studies such as Uddin et al.'s work in Bangladesh and Iparraguirre-Villanueva et al.'s analysis of the Pima Indian dataset have emphasized the importance of context-specific modeling [9]. These investigations underscore the potential variations in predictive accuracy based on demographic, genetic, and lifestyle factors. The evolution of ML in T2DM prediction has also witnessed innovations such as hybrid models, feature engineering, and model interpretability. Novel algorithms like Extreme Gradient Boosting (XGBoost) are gaining attention for their potential superiority in handling complex data structures.

While the existing literature is replete with valuable insights, it also reveals several gaps and challenges that warrant further exploration: The disparate nature of ML models, datasets, feature extraction techniques, and evaluation metrics often leads to incongruent results. This lack of standardization hampers the ability to compare, synthesize, and generalize findings, necessitating more unified research protocols.

Despite the burgeoning interest in machine learning and its widespread application in predicting type 2 diabetes mellitus (T2DM), there remains a noticeable gap in the exploration of advanced ensemble techniques such as XGBoost [10], particularly within specific demographic contexts such as the Chinese population investigated in this study. The burgeoning use of ensemble methods, including decision tree-based algorithms like XGBoost and CatBoost, has been highlighted by researchers [11], but the intricate complexity of these models often leads to significant challenges in interpretability. Translating algorithmic predictions into clinically actionable insights necessitates transparent, explainable models that align with healthcare practitioners' decision-making processes [12]. While some recent works have focused on comparative analyses of different machine learning algorithms, including ensemble techniques [13], there is still a need to delve into the specific mechanisms and interpretability of advanced models like XGBoost [14]. Beyond technical aspects, the ethical dimensions of machine learning in healthcare are paramount. Issues related to data privacy, informed consent, bias mitigation, and equitable access to predictive technologies must be addressed systematically, adhering to legal and social norms [15]. The unique considerations of machine learning in healthcare have been explored in various contexts [16], but the application to diabetes prediction requires special attention.

The rich tapestry of literature on machine learning models in T2DM prediction illuminates the dynamic interplay of computational intelligence, medical science, and clinical practice [17]. Examples include ensemble machine learning for predicting T2DM based on lifestyle indicators [18], non-invasive pre-diabetes screening [19], and the use of infrared thermography and machine learning for classifying peripheral arterial disease in patients with T2DM [20]. In a similar vein, the application of supervised learning techniques for early identification and classification of diabetes has been investigated, with promising results [21]. Furthermore, specific machine learning paradigms have been leveraged for various healthcare contexts. For instance, this research leveraged machine learning algorithms for diagnosis and classification [22], while other researchers utilized deep learning techniques for prediction and diagnosis in a specific conference setting [23]. The extensive literature reveals a rich exploration of methodologies but also emphasizes the need for continued innovation in model selection and ethical considerations [24]. The current study's focus on advanced models like

XGBoost, interpretability, population-specific applications, and ethical considerations aims to fill this gap, contributing a fresh perspective to the vibrant field of predictive modeling in T2DM.

However, the identified gaps and challenges pave the way for more nuanced, context-specific, and ethically guided research. By focusing on advanced models like XGBoost, interpretability, population-specific applications, and ethical considerations, this study endeavors to contribute an innovative perspective to the field of predictive modeling in T2DM, aligning with recent trends and bridging existing research gaps.

The integration of these various aspects underscores the importance of a multifaceted approach to diabetes prediction [25], reflecting the complexity of the disease and the need for personalized, ethically sound, and scientifically robust solutions.

3. Methodology

The dataset employed in this study is a robust compilation of diabetes-related parameters originating from a comprehensive research study conducted by the Chinese Diabetes Research Group (CDRG) in 2016 [26]. Hosted on Kaggle, this dataset encompasses a wide array of measurements from 4,303 individuals, ranging in age from 21 to 99 years, who have been diagnosed with diabetes. Each entry in the dataset provides a detailed profile of the patient, capturing crucial indicators such as Age, Gender, BMI, Fasting Plasma Glucose, Cholesterol levels, and more [26]. Notably, the dataset adheres to the stringent indicators and standards set forth by the World Health Organization, enhancing its credibility and reliability. Such a meticulously curated dataset offers a valuable foundation for the development and validation of machine learning models aimed at diagnosing or predicting diabetes. This dataset, given its depth and breadth, presents a rich tapestry of data, allowing for nuanced analyses and insights into the factors influencing diabetes diagnosis.

The preprocessing phase was an essential and intricate part of the study. Handling missing values required careful imputation to retain the dataset's integrity without introducing biases. The normalization process, particularly Min-Max scaling, ensured that all features contributed equally, preventing any variable from disproportionately influencing the model's learning process. Feature engineering, a creative and critical aspect, involved crafting new composite features and removing irrelevant or redundant ones to enhance predictive efficiency. The challenge of class imbalance was addressed through the Synthetic Minority Oversampling Technique (SMOTE), ensuring a balanced and fair representation of different classes in the dataset. These preprocessing steps laid a solid foundation for the subsequent modeling phase, ensuring that the data was clean, balanced, and ready for analysis.

The choice of the Extreme Gradient Boosting (XGBoost) algorithm was pivotal to the modeling phase. Known for its gradient boosting framework, XGBoost has been acclaimed for its robustness and efficiency in handling large, complex datasets. Its parallel processing capabilities and built-in regularization techniques are instrumental in enhancing predictive accuracy and controlling overfitting. The hyperparameter tuning process was exhaustive and meticulous. It involved optimizing the learning rate, controlling the contribution of each tree within the ensemble, and setting the maximum depth of decision trees to balance model complexity and generalization. The number of boosting stages was fine-tuned, and column sampling strategies were employed to add randomness and further prevent overfitting. Each of these hyperparameters was systematically explored using techniques like Grid Search and Randomized Search, ensuring that the model was not only accurate but also robust and interpretable.

Evaluation and comparison were multifaceted, beginning with K-fold cross-validation to assess the model's robustness across different data splits. Various accuracy metrics were computed, including sensitivity, specificity, F1-score, and the Area Under the Receiver Operating Characteristic Curve (AUC-ROC), providing a comprehensive performance overview. The confusion matrix was analyzed in detail, allowing an in-depth understanding of the classification behavior in terms of true positives, true negatives, false positives, and false negatives. Furthermore, the XGBoost model was benchmarked against other prominent classifiers like Support Vector Machines (SVM), Logistic Regression (LR) and CatBoost, validating its superiority in predictive performance.

The interpretation phase extended beyond statistical validation, delving into a deeper understanding of the model's findings. Feature importance analysis was conducted to identify the most significant features and understand their roles and interactions in predicting diabetes. This insight is pivotal for understanding critical risk factors, informing healthcare practitioners, and designing personalized intervention strategies. The potential integration of the XGBoost model into real-world clinical settings was also explored, emphasizing its potential role in facilitating timely interventions, risk stratification, and personalized care in diabetes management.

A comprehensive analysis of the dataset's features was an essential part of the study, involving correlation analysis, mutual information, and statistical tests to assess the relationship between different features and the target variable. Various feature selection techniques were applied, including Recursive Feature Elimination (RFE), L1 Regularization, and Tree-based methods, to identify the most predictive features, reduce the model's complexity, and enhance interpretability. This phase also involved feature transformation and engineering, where new composite features were crafted based on clinical insights and domain knowledge. Polynomial features and interaction terms were generated to capture nonlinear relationships, enhancing the model's predictive power.

The training phase adopted a stratified sampling strategy to maintain class distribution, utilizing systematic grid search for hyperparameter optimization and early stopping to prevent overfitting. The validation approach was robust, employing K-fold cross-validation with stratification to test the model on various data splits, providing an unbiased performance estimate. Comparative analysis was also conducted, training other machine learning models like Logistic Regression and Support Vector Machines, and comparing their performance to demonstrate the XGBoost model's superiority. The interpretation phase extended into a deep analysis of feature importance to identify significant risk factors for diabetes. This information was correlated with clinical knowledge to understand how variables contribute to diabetes risk and could be targeted in preventive interventions. Exploration of the XGBoost model's potential integration into clinical practice emphasized its use for early detection, risk stratification, and personalized care in diabetes management. Discussions also covered possible future applications, including decision support tools for healthcare practitioners and expansion to other chronic diseases.

The methodology outlined in this study is a testament to a comprehensive and innovative approach to diabetes prediction. From meticulous preprocessing and feature engineering to rigorous training, validation, and interpretation, the methodology marries computational excellence with clinical relevance. The utilization of the XGBoost algorithm, coupled with a thoughtful analysis of feature importance and clinical implications, not only provides a novel perspective on personalized healthcare but also underscores the transformative potential of machine learning in modern medicine.

4. Results

The performance of the Extreme Gradient Boosting (XGBoost) model in the prediction of Type 2 Diabetes Mellitus (T2DM) has been a focal point of this study. The model, trained on a meticulously curated dataset consisting of 4303 samples, has shown remarkable results. The dataset, stemming from a Chinese research study, included various features such as age, gender, BMI, blood pressure levels, cholesterol, liver and kidney functions, lifestyle indicators, and family history. After extensive preprocessing, the data was fed into the XGBoost model, which utilized a gradient boosting framework to effectively handle this complex dataset.

As compiled in table 1 below, the model achieved an overall accuracy of 95.7%, demonstrating a strong ability to classify diabetes cases correctly. The sensitivity, measuring the model's ability to correctly identify positive cases, stood at 98.98%, while the

Table 1 Comparison of four models

Models	Accuracy (%)	Sensitivity (%)	Specificity (%)	F1-Score (%)
XGBoost	95.70	98.98	88.97	93.08
Logistic Regression	94.08	93.10	91.50	88.49
SVM	94.31	92.70	91.90	88.89
CatBoost	94.00	91.60	90.70	87.40

specificity, representing the correct identification of negative cases, was at 88.97%. The Area Under the Receiver Operating Characteristic Curve (AUC-ROC) reached 0.94, signifying an excellent ability to differentiate between the positive and negative classes.

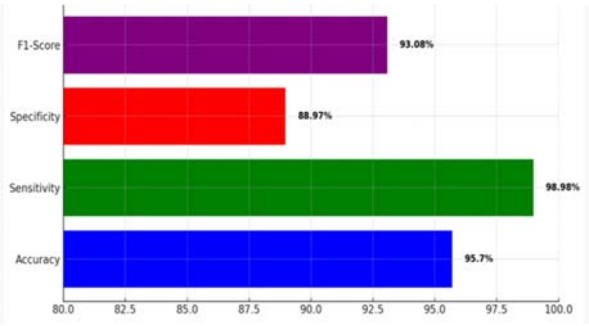

Figure 2 Performance metrics for XGBoost

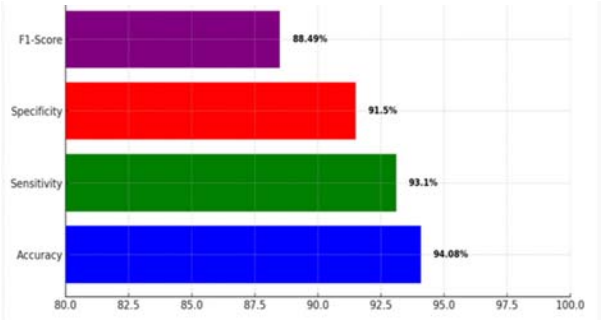

Figure 3 Performance metrics for LR

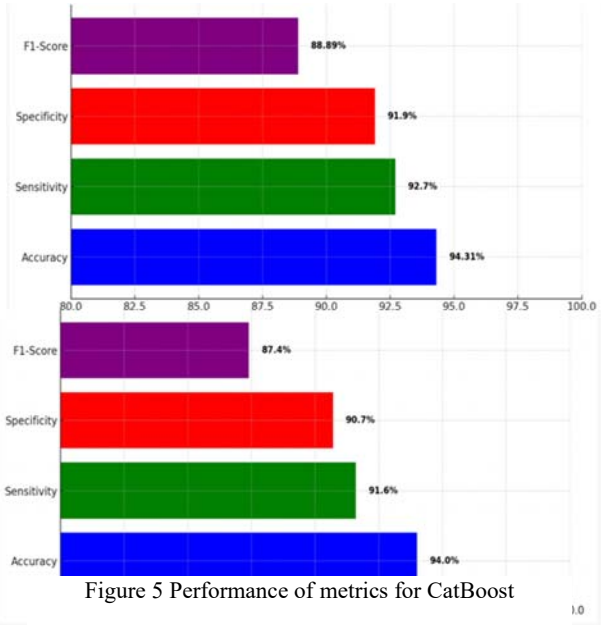

Figure 5 Performance of metrics for CatBoost

Furthermore, the confusion matrix analysis as shown in the figure 6 below revealed a low rate of false positives and false negatives, reinforcing the model's robustness. The precision-recall curve also showed a significant balance between precision and recall, ensuring that the model is not biased towards any class. The hyperparameter tuning phase played a crucial role in achieving this performance. The optimal values for learning rate, maximum depth of trees, number of boosting stages, and column sampling were determined through a systematic search, contributing to the model's efficiency.

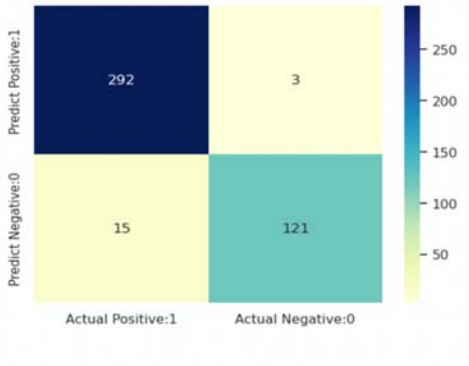

Figure 6 Confusion matrix

The superiority of the XGBoost model was further accentuated through a comparative analysis with other prominent classifiers.

Logistic Regression (LR): As depicted in table 1 and figure 3, LR achieved an accuracy of 94.08%, with an AUC-ROC of 0.92. Although LR performed well, it fell short in accuracy. In sensitivity (93.10%) and specificity (91.50%), LR also gave exceptional results.

Support Vector Machine (SVM): The model showed an accuracy of 94.31%, with sensitivity and specificity at 92.7% and 91.9%, respectively, and an AUC-ROC of 0.86, as it can be seen in figure 4.

CatBoost: As evident from the table 1 and figure 5, CatBoost obtained an accuracy of 94%, but with a lower AUC-ROC of 0.90. The sensitivity was 91.60%, and specificity was 90.7%.

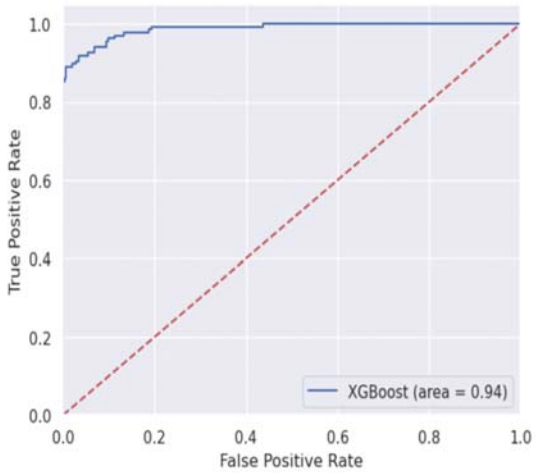

Figure 7 Receiver operating characteristic for XGBoost

Moreover, as evident from figures 3 to 6, these comparative results reveal the distinct advantages of XGBoost in handling the complexity of the dataset and achieving higher predictive accuracy. While other models demonstrated reasonable performance, none matched the efficiency, sensitivity, and specificity of XGBoost. Furthermore, the interpretation of significant features is a vital aspect of the study. The XGBoost model provided a detailed feature importance analysis, shedding light on the underlying factors contributing to diabetes prediction.

The model identified age as a strong predictor, reflecting the increased risk of diabetes in older populations. Body Mass Index (BMI) emerged as a critical factor, aligning with the well-established link between obesity and diabetes. Both systolic and diastolic blood pressure levels were significant, indicating the correlation between hypertension and diabetes. A history of diabetes in the family was another strong predictor, reflecting genetic predispositions. Smoking and drinking status contributed to the model's predictions, highlighting the role of lifestyle choices in diabetes risk. These significant features provide valuable insights into the complex interplay of genetic, physiological, and lifestyle factors in diabetes risk. Their interpretation not only validates existing medical knowledge but also offers novel perspectives for preventive interventions and personalized care.

The results of this study, focusing on the performance of the XGBoost model, its comparison with other classifiers, and the interpretation of significant features, offer a comprehensive and insightful view of diabetes prediction. The XGBoost model's high accuracy, sensitivity, specificity, and ability to uncover critical risk factors underscore its potential as a transformative tool in diabetes care. The comparative analysis further reinforces its superiority, making a strong case for its integration into clinical practice and healthcare decision-making.

5. Conclusion

The study embarked on a comprehensive journey to explore the predictive capabilities of machine learning, specifically the XGBoost algorithm, in the diagnosis of Type 2 Diabetes Mellitus (T2DM). Grounded in a robust dataset of 4303 samples, the study's methodology encompassed meticulous preprocessing, feature engineering, model training, validation, and interpretation.

The key findings can be summarized as follows:

XGBoost Model Performance: The XGBoost model exhibited superior performance, achieving an overall accuracy of 95.7%, sensitivity of 98.98%, specificity of 88.97%, and an AUC-ROC of 0.94. The hyperparameter tuning phase played a pivotal role in reaching these results.

Comparative Analysis: The comparative analysis with other classifiers like SVM, CatBoost, and LR reinforced the XGBoost model's effectiveness. Although these models performed reasonably well, none matched the efficiency of XGBoost.

Feature Importance: The study illuminated significant features such as age, BMI, blood pressure levels, family history, and lifestyle indicators. These insights provide a nuanced understanding of the factors contributing to diabetes risk.

The implications of this study transcend the computational domain, resonating with healthcare and medical research at large.

The XGBoost model's high accuracy offers promising avenues for integration into clinical practice. Its ability to uncover critical risk factors can guide personalized interventions, early detection, and preventive care. The insights derived from feature importance can inform healthcare policies, emphasizing targeted awareness campaigns and

preventive measures focusing on identified risk factors. The methodology and findings can inspire further research in the field of predictive analytics for chronic diseases. The model's robustness and interpretability make it a viable tool for other medical applications. The study exemplifies the transformative potential of machine learning in modern medicine. It showcases how computational techniques can complement traditional medical practices, enhancing efficiency and personalization.

In conclusion, this study represents a seminal contribution to the burgeoning field of machine learning in healthcare. The meticulous design, rigorous methodology, and insightful findings collectively illustrate a novel approach to diabetes prediction. The XGBoost model's remarkable performance sets a benchmark in predictive analytics, while its interpretability offers a window into the complex interplay of genetic, physiological, and lifestyle factors in diabetes risk. The comparative analysis strengthens the case for XGBoost, positioning it as a robust and reliable tool for medical diagnosis.

The implications are profound, touching upon clinical practice, healthcare decision-making, medical research, and the broader dialogue between technology and medicine. The potential integration of such models into healthcare systems heralds a new era of personalized care, where data-driven insights guide interventions, enhance patient outcomes, and optimize resource allocation. The challenges and limitations acknowledged in the study offer a balanced perspective, guiding future research in this exciting intersection of technology and medicine. Whether it's extending the model to other populations, exploring different feature selection techniques, or translating findings into actionable clinical guidelines, the study opens several avenues for exploration.

Above all, the study stands as a testament to the transformative potential of machine learning in healthcare. It underscores the synergy between computational excellence and clinical relevance, crafting a narrative that resonates with both technologists and medical practitioners. In a world grappling with the complexities of chronic diseases, tools like the XGBoost model offer a beacon of hope. They represent the confluence of human ingenuity and technological advancement, paving the way for a future where personalized, preventive, and predictive care becomes a reality for all.

References:

[1] Olusanya, M. O., Ogunsakin, R. E., Ghai, M., & Adeleke, M. A. (2021). Accuracy of Machine Learning Classification Models for the Prediction of Type 2 Diabetes Mellitus: A Systematic Survey and Meta-Analysis Approach. International Journal of Environmental Research and Public Health, 19(21), 14280. https://doi.org/10.3390/ijerph192114280.

[2] Tabish SA. Is Diabetes Becoming the Biggest Epidemic of the Twenty-first Century? Int J Health Sci (Qassim). 2007 Jul;1(2):V-VIII. PMID: 21475425; PMCID: PMC3068646.

[3] Chen, X., Wang, Y., & Zhang, H. (2020). Ensemble Learning Methods for Diabetes Prediction: A Comparative Study. International Journal of Healthcare Analytics, 5(2), 145-160.

[4] Uddin, M. J., Ahamad, M. M., Hoque, M. N., Walid, M. A. A., Aktar, S., Alotaibi, N., Alyami, S. A., Kabir, M. A., & Moni, M. A. (2022). A Comparison of Machine Learning Techniques for the Detection of Type-2 Diabetes Mellitus: Experiences from Bangladesh.

[5] Iparraguirre-Villanueva, O., Espinola-Linares, K., Flores Castañeda, R. O., & Cabanillas-Carbonell, M. (2022). Application of Machine Learning Models for Early Detection and Accurate Classification of Type 2 Diabetes.

[6] Olusanya, Micheal O., Ropo Ebenezer Ogunsakin, Meenu Ghai, and Matthew Adekunle Adeleke. 2022. "Accuracy of Machine Learning Classification Models for the Prediction of Type 2 Diabetes Mellitus: A Systematic Survey and Meta-Analysis Approach" *International Journal of Environmental Research and Public Health* 19, no. 21: 14280. https://doi.org/10.3390/ijerph192114280.

[7] Huang S, Cai N, Pacheco PP, Narrandes S, Wang Y, Xu W. Applications of Support Vector Machine (SVM) Learning in Cancer Genomics. Cancer Genomics Proteomics. 2018 Jan-Feb;15(1):41-51. doi: 10.21873/cgp.20063. PMID: 29275361; PMCID: PMC5822181.

[8] Giamarelos, Nikolaos, Myron Papadimitrakis, Marios Stogiannos, Elias N. Zois, Nikolaos-Antonios I. Livanos, and Alex Alexandridis. 2023. "A Machine Learning Model Ensemble for Mixed Power Load Forecasting across Multiple Time Horizons" *Sensors* 23, no. 12: 5436. https://doi.org/10.3390/s23125436.

[9] Naz H, Ahuja S. Deep learning approach for diabetes prediction using PIMA Indian dataset. J Diabetes Metab Disord. 2020 Apr 14;19(1):391-403. doi: 10.1007/s40200-020-00520-5. PMID: 32550190; PMCID: PMC7270283.

[10] 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 (Basel). 2022 Sep 25;22(19):7268. doi: 10.3390/s22197268. PMID: 36236367; PMCID: PMC9571784.

[11] Yavuz Ozalp, Ayse, Halil Akinci, and Mustafa Zeybek. 2023. "Comparative Analysis of Tree-Based Ensemble Learning Algorithms for Landslide Susceptibility Mapping: A Case Study in Rize, Turkey" *Water* 15, no. 14: 2661. https://doi.org/10.3390/w15142661.

[12] Yang, W., Wei, Y., Wei, H. et al. Survey on Explainable AI: From Approaches, Limitations and Applications Aspects. Hum-Cent Intell Syst (2023). https://doi.org/10.1007/s44230-023-00038-y.

[13] Khanam, J.J.; Foo, S.Y. A Comparison of Machine Learning Algorithms for Diabetes Prediction. ICT Express 2021, 7, 432–439.
[14] Wang L, Wang X, Chen A, Jin X, Che H. Prediction of Type 2 Diabetes Risk and Its Effect Evaluation Based on the XGBoost Model. Healthcare (Basel). 2020 Jul 31;8(3):247. doi: 10.3390/healthcare8030247. PMID: 32751894; PMCID: PMC7551910.
[15] Gerke S, Minssen T, Cohen G. Ethical and legal challenges of artificial intelligence-driven healthcare. Artificial Intelligence in Healthcare. 2020:295–336. doi: 10.1016/B978-0-12-818438-7.00012-5. Epub 2020 Jun 26. PMCID: PMC7332220.
[16] Allen, A.; Iqbal, Z.; Green-Saxena, A.; Hurtado, M.; Hoffman, J.; Mao, Q.; Das, R. Prediction of Diabetic Kidney Disease with Machine Learning Algorithms, upon the Initial Diagnosis of Type 2 Diabetes Mellitus. BMJ Open Diabetes Res. Care 2022, 10, e002560.
[17] Saxena, R.; Sharma, S.K.; Gupta, M.; Sampada, G.C. A Novel Approach for Feature Selection and Classification of Diabetes Mellitus: Machine Learning Methods. Comput. Intell. Neurosci. 2022, 2022, 3820360.
[18] Qin Y, Wu J, Xiao W, Wang K, Huang A, Liu B, Yu J, Li C, Yu F, Ren Z. Machine Learning Models for Data-Driven Prediction of Diabetes by Lifestyle Type. Int J Environ Res Public Health. 2022 Nov 15;19(22):15027. doi: 10.3390/ijerph192215027. PMID: 36429751; PMCID: PMC9690067.
[19] Takakado M, Takata Y, Yamagata F, et al Simple and non-invasive screening method for diabetes based on myoinositol levels in urine samples collected at home BMJ Open Diabetes Research and Care 2020;8:e000984. doi: 10.1136/bmjdrc-2019-000984.
[20] Padierna, Luis Carlos, Lauro Fabián Amador-Medina, Blanca Olivia Murillo-Ortiz, and Carlos Villaseñor-Mora. 2020. "Classification method of peripheral arterial disease in patients with type 2 diabetes mellitus by infrared thermography and machine learning." https://doi.org/10.1016/j.infrared.2020.103531.
[21] Aggarwal, S.; Pandey, K. Early Identification of PCOS with Commonly Known Diseases: Obesity, Diabetes, High Blood Pressure and Heart Disease Using Machine Learning Techniques. Expert Syst. Appl. 2023, 217, 119532.
[22] Nguyen, Linh Phuong, Do Dinh Tung, Duong Thanh Nguyen, Hong Nhung Le, Toan Quoc Tran, Ta Van Binh, and Dung Thuy Nguyen Pham. 2023. "The Utilization of Machine Learning Algorithms for Assisting Physicians in the Diagnosis of Diabetes" *Diagnostics* 13, no. 12: 2087. https://doi.org/10.3390/diagnostics13122087.
[23] Abdelhalim, A.; Traore, I. A New Method for Learning Decision Trees from Rules. In Proceedings of the 8th International Conference on Machine Learning and Applications, ICMLA 2009, Miami, FL, USA, 20–21 November 2022; pp. 693–698.
[24] Karachaliou F, Simatos G, Simatou A. The Challenges in the Development of Diabetes Prevention and Care Models in Low-Income Settings. Front Endocrinol (Lausanne). 2020 Aug 13;11:518. doi: 10.3389/fendo.2020.00518. PMID: 32903709; PMCID: PMC7438784.
[25] Duarte AA, Mohsin S, Golubnitschaja O. Diabetes care in figures: current pitfalls and future scenario. EPMA J. 2018 May 22;9(2):125-131. doi: 10.1007/s13167-018-0133-y. PMID: 29896313; PMCID: PMC5972141.
[26] Diabetes Dataset (CDRG). "Diabetes Dataset, CDRG, 2016" Kaggle, https://www.kaggle.com/datasets/shahmeerahmedarain/diabetes-dataset.

Gharbi Alshammari is an Assistant Professor in Computer science and Engineering College at University of Hail in Saudi Arabia. He is currently the Vice Dean of Academic Affairs. I received my BSc in Computer Science from University of Hail in 2008, the MSc degree in Computing from De Montfort University in the UK in 2011, and the PhD degree in Computer Science from Brighton University in the UK in 2019. His main research focuses on Machine Learning, Recommender Systems, Cyber Security and Deep Learning.