Literature Review: Diabetes Prediction Using Machine Learning

Md Minhajul Abedin ID: 202483033 Email: mminhajula@mun.ca
Md Jawad Khan ID: 202381977 Email: mjawadk@mun.ca

1 Introduction

Diabetes, a chronic metabolic disorder, affects millions of people around the world and is one of the leading causes of mortality and morbidity. In addition to the consequences of poor control leading to complications such as heart disease or stroke, it can also result in kidney failure and neuropathy. Early detection and correct treatment are essential; however, traditional diagnostic methods often used rely mainly upon clinical symptoms plus biochemical tests of blood samples taken - these may not always be quick or reliable when the condition is in its early stages, with symptoms hardly noticeable as yet. Therefore, there is an urgent need for innovative methods to diagnose diabetes more accurately and efficiently.

Machine Learning (ML) has emerged as a promising tool in the healthcare sector, especially for diabetes prediction. By using a database of clinical and demographic information, ML algorithms can detect unknown patterns and give accurate predictions which are important for the early detection of diabetes. What's more, models such as Random Forests, K-Nearest Neighbors (KNN), Support Vector Machines (SVM), and ensemble learning have shown promise in improving prediction accuracy. These methods differ from traditional statistical approaches by considering complex, high-dimensional data: they are particularly well suited to healthcare applications where many different factors contribute to disease outcomes.

Recent studies using machine learning to forecast diabetes provide the focus of this literature review. It lists the most common algorithms and methodologies and discusses, respectively, their strengths, weaknesses and predictive performance. Furthermore, it also points out some of the major problems facing researchers today: limits on data sets and over representation in classes. The need for thorough preprocessing of data is also mentioned. By pulling together the conclusions of these studies, the study hopes to give useful insights on how machine learning might change diagnosis for diabetics and so improve clinical judgment. The ultimate goal is to help patients more.

2 Literature Review

M. Alehegn et al. [1] focused on utilizing machine learning techniques for early diabetes detection, increasing the accuracy of prediction. The authors used four well-known algorithms - K-Nearest Neighbors (KNN), Naïve Bayes, Random Forest, and J48 - to create an ensemble model, in an endeavor to overcome this limitation of individual classifier performance from which one can expect only so much change without considerable increases in risk. The findings show that if these classifiers are combined, it produces a more accurate model with better stability than simply using a single algorithm does. However, the study also notes that data prepossessing (such as noise removal and handling missing values) is essential for improving model accuracy. The study's methods align with previous research on machine learning algorithms applied to health data, such as Random Forest and Naïve Bayes; however, there is no evidence to suggest that combining models from a series of different classifiers effectively enhances prediction results. Also, the paper said that variable selection, such as glucose, blood pressure, and BMI, greatly influences the predictability of dependent variables. The study adds to the literature by demonstrating how ensemble models can enhance early diagnosis and highlights key areas for methodological advancement in diabetes prediction research, ultimately aiming to improve health outcomes through timely intervention.

By Md. Maniruzzaman et al. [2] aims to develop an effective machine learning system for predicting diabetes in individuals. The study is a diabetes dataset derived from the National Health and Nutrition Examination Survey (NHANES) that uses logistic regression for feature selection and compares its performance against four classifiers: Naïve Bayes, Decision Tree, Adaboost, and Random Forest. The methodology involved applying three cross-validation protocols (K2, K5, K10) and evaluating classifiers based on accuracy, sensitivity, and area under the curve (AUC). The key finding here is that the RF classifier, combined with feature selection of LR which is significantly more efficient than any of these other three. For K10 protocols, it gave an accuracy rate 94.25% and AUC came to 0.95. Strengths of this study are a large field study, with clear comparative analysis on many classifiers and robust evaluation performance indexes for the ML model. However, the paper's limitations include its reliance on a single dataset and possible confines of generalizability, and a lack of exploration into other advanced ML techniques like deep learning.

B. Sridhara Murthy et al. [3] mainly evaluates Logistic Regression and Decision Tree Classifier (DTC) for diabetes prediction based on Pima Indian dataset (768 records, 35% of them diabetic). With an 80:20 train-test split, it was found during testing that, compared to DTC which had a maximum accuracy of 78.57%, LR was superior at 82.46%. It had better recall (0.659 vs 0.61) permissive parameters (0.849 to 0.844) but suffered a worse precision at 0.76 and recall rates (specificity) were still lower than these of the DTC. The strengths of the development include a clear comparison of different algorithms and the use of multiple metrics (e.g., F1 score, specificity). However, limitations arise from the small size of these datasets and class imbalance; the exclusion of advanced models (such as Random Forest). In addition, generalizability is weak due to the lack of cross-validation or discussion on handle missing data in the Pima dataset – a problem which commonly occurs when working with humans. While the research showcases LR as of great predictive value for diabetes, future research needs to rigorously address data quality issues and use diverse algorithms so that results can be applied to actual clinical practice in different medical settings.

A. Mujumdara et al [4] proposed a predictive model built on a custom dataset (800 records, 10 variables, including socio-demographic factors such as occupation type) made in order to fill gaps in more traditional pieces of health care information like Pima. After completing preprocessing (missing value imputation, normalization) and using K-means clustering to identify possible relationships (e.g., glucose levels vs. age); it tried a total of 13 ML algorithms, among which Logistic Regression achieved 96% accuracy, a result greatly surpassing previous benchmarks (e.g., 76% with Pima). An additional gain in efficiency from taking a pipeline approach left their AdaBoost classifier with a margin of giving 98.8% accuracy, verifying the effectiveness of feature engineering and ensemble learning methods can have. A parade of strong points include methods to enrich lifestyle attribute sets, a thorough comparison between different algorithm families, and careful preparation for scale though shortcomings like sparse data size limit generalization, inability to touch on class unbalance mean that the study thus indicates the prospect that several complex pipeline assemblages and overall data inclusion hint for clinical predictions can be pursued, but should be studied on bigger samples as well as having more different areas.

I. Tasin et al. [5] proposed a diabetes prediction system using a merged dataset of 877 samples. The resulting whole-volume database combines the Pima Indian diabetic dataset and a private Bangladeshi database (RTML) taken from 203 female textile workers. The RTML dataset's missing insulin values were imputed through XGBoost regression. Class imbalance was treated by both SMOTE and ADASYN. Mutual information shows that glucose, BMI, age, and insulin are important predictors. With ADASYN, the XGBoost classifier attains an 81% accuracy rate and 0.81 F1-score, outperforming other models such as Random Forest (76%) and SVM (78%). Results from explainable AI tools (SHAP, LIME) show that glucose levels and number of pregnancies are the main decision drivers. The model was rolled out as a web app and an Android application for real-time predictions. Pros include the innovative integration of datasets, as well as interpretability, while cons cover small sample size and reliance on imputed insulin values. The work highlights the potential for using advanced ML techniques to combine domain-specific data for clinical applications.

M. K. Hasan et al. [6] propose a robust framework for diabetes prediction using the Pima Indian Diabetes dataset, featuring a preprocessing pipeline with outlier rejection (IQR), mean-based missing value imputation, standardization, correlation-based feature selection, and stratified K-fold cross-validation to address data quality and class imbalance. The framework evaluates classifiers like k-NN, Decision Trees, Random Forest, AdaBoost, Naive Bayes, XGBoost, and MLP, introducing a novel AUC-weighted soft voting ensemble, with the AdaBoost+XGBoost combination achieving an AUC of 0.950, sensitivity of 0.789, and specificity of 0.934, surpassing state-of-the-art methods (e.g., Maniruzzaman et al. (2018) AUC 0.930, Sisodia et al. (2018) AUC 0.819) by 2%. This approach mitigates limitations of prior work, such as inadequate preprocessing, by enhancing generalizability and reproducibility through publicly available code, demonstrating the synergy of ensembling and preprocessing for accurate, robust medical diagnostics.

K. Lu et al. [7] focuses on utilizing machine learning methods combined with data from the Canadian Primary Care Sentinel Surveillance Network (CPCSSN) so as to identify prediabetes in the Canadian population. To predict prediabetes. The authors assess seven ML models: logistic regression, Random Forest, XGBoost, Naive Bayes, KNN, SVM, and a Deep Neural Network (DNN). When trained with early stop regularization, the DNN achieved the highest recall rate at 60%. Channels such as age, BMI (body mass index), and taking hypertension medication were pinpointed through SHAP analysis; and yet again, literature normally studied that concerns diabetes risk factors is in line with these findings. The study underscores the difficulties of distinguishing individuals who are prediabetic from those who aren't. These people share biological characteristics, so model performance is lower than in studies that aimed to predict diabetes (e.g., Lai et al., 2019). Examples of ethical issues mentioned in this study include the lack of racial/ethnic data in datasets and the potential for ML applications to bring prejudices into healthcare. Further work should take in a wider space of variables (e.g., lifestyle factors) and make the model in general more versatile. The authors accentuate the need for early detection in order to stave off the transition to Type 2 diabetes. In this way, the study is a piece of research into health solutions involving ML that also make up for blind spots not covered by existing models.

M. Soni et al. [8] focuses on predicting diabetes on time for a more stringent concordance rate. It uses all kinds of machine learning models, this way most-likely and

TABLE 1: Summary of Diabetes Prediction Studies Using Machine Learning

Study (Author)	Algorithms Used	Main Focus	Key Features	Performance	Strengths
M. Alehegn et al. [1]	KNN, Naïve Bayes, RF, J48 (ensemble)	Ensemble models for diabetes prediction	Glucose, BP, BMI	Improved accuracy (no exact metric)	Ensemble benefits; data preprocessing
Maniruzzaman et al. [2]	LR, NB, DT, AdaBoost, RF	Evaluation of multiple ML classifiers	Glucose, BMI, age	94.25% acc., AUC 0.95 (RF)	Robust CV; comparative analysis
B. Sridhara Murthy et al. [3]	Logistic Regression, Decision Tree	Comparison on Pima Indian dataset	Glucose, BMI	LR: 82.46% accuracy	Multi-metric comparison
A. Mujumdara et al [4]	13 algorithms (incl. AdaBoost, LR)	Evaluation with feature engineering	Socio-demographics, glucose, age	AdaBoost: 98.8% accuracy	Feature engineering; pipeline
I. Tasin et al. [5]	XGBoost, RF, SVM	Integration with explainable AI	Glucose, BMI, age, insulin	XGBoost: 81% acc., F1 0.81	SHAP/LIME; deployed as app
M. K. Hasan et al. [6]	k-NN, DT, RF, AdaBoost, NB	Explainable AI with SHAP/LIME	Not specified	AUC 0.95	Public code; preprocessing
K. Lu et al. [7]	LR, RF, XGBoost, NB, KNN, SVM, DNN	Prediabetes identification	Age, BMI, hypertension meds	DNN: 60% recall	Ethical considerations; SHAP
M. Soni et al. [8]	KNN, LR, SVM, GB, RF	Type 2 diabetes prediction	Glucose, BMI, age	RF: 77% accuracy	Algorithm comparison
M. A. Sarwar et al [9]	KNN, NB, SVM, LR, DT, RF	Healthcare applications	Glucose, BMI, age	SVM/KNN: 77% accuracy	Practical healthcare focus
N. P. Tigga et al [10]	LR, KNN, SVM, NB, DT, RF	Ensemble for improved prediction	Age, family history, activity	RF: 94.1% acc., AUC 1.0	Hybrid dataset; strong CV

least-likely cases are known at once. The authors made use of classification and ensemble techniques, including K-Nearest Neighbor (KNN), Logistic Regression (LR), Support Vector Machine (SVM) and Gradient Boosting (GB). They applied these to the Pima Indian Diabetes Dataset, and file with 768 patients' data that covers a number of attribute values such as glucose levels, BMI and age. There were problems with the missing data and the classification of sets. The first problem was resolved by discarding the missing information, but this left us still in need of a method for dividing our dataset into 80% training set and 20% test sets. According to the table, the Random Forest model displays the highest accuracy with 77% while feature importances identified glucose levels as predictive variables most influencing diabetes. A key advantage of this study over other research in the field is that it compares different machine learning algorithms and provides solid evidence for the performance of Random Forest in Diabetes Prediction; some limitations also exist, however. The dataset used is unbalanced (500 non-diabetic cases versus 268 diabetic cases), so judgments must be made on this basis; generalizing a finding from one specific demographic (the Pima Indian population) to another one is fraught with difficulty. The investigators note that machine learning techniques can play a role in early prediction for Diabetes, although it remains to be confirmed in future research whether they have any advantage over other methods.

M. A. Sarwar et al [9] evaluate six machine learning algorithms (KNN, Naive Bayes, SVM, Logistic Regression, Decision Trees, Random Forests) on the Pima Indian Diabetes Dataset (768 records). There will be a more accurate model which better predicts diabetes from data such as theirs and hopefully can help early detection of this medical condition. The methodology included pre-processing (handling missing values and a 70-30 train-test split) and implementation of models in Python. Results showed that SVM and KNN both got the highest accuracy at 77%, while glucose levels, BMI, and age are important predictive features. There are Strengths, including a comparative analysis of multiple algorithms and practical healthcare, but the limitations arise from the small, imbalanced dataset as well as moderate accuracy; therefore, generalizability is restricted. The team concludes that machine learning has potential use for predictive diabetes, but it needs larger datasets and better quality of data before being applicable in clinical work.

N. P. Tigga, et al [10] tried to assess six machine learning algorithms—Logistic Regression, KNN, SVM, Naïve Bayes, Decision Trees, and Random Forest— for predicting diabetes risk in humans. A fully pre-processed data set of 952 people (from questionnaires) in combination with the Pima Indian dataset was chosen as the basis for this study. From data preprocessing to model assessment, a 10-fold cross-validation technique was applied. Based on the custom dataset, we found Random Forest to be the top performer, achieving 94.1% accuracy and 90% on Pima, an AUC of 1.0, showing excellent classification accuracy. Key drivers of prediction were found to be age, family history of diabetes, activity levels, and the use of medication. Most notably, this study offers insight into the performance of algorithms and details how they might be applied in practice to early detection for diabetes. And yet there were caveats. The data had an unbalanced distribution (fewer diabetic cases). Reporting on matters that relieve much too heavily with subjective self-reporting is subject to bias. The findings underscore Random Forest's reliability for diabetes prediction but suggest an increased need for large, more diverse datasets. It needs to be demonstrated how Random Forest, in this era of giant clinical data and continuous glucose monitoring, might still have worth in the clinical setting. Hence, future research might want to explore hybrid models or incorporate more risk factors in order to attain still better prediction capabilities.

3 Conclusion

In the reviewed studies, indicates that machine learning has great potential to markedly improve the accuracy and efficiency of predicting diabetes Ensemble models, like Random Forest and XGBoost, consistently give very good performance. In addition, advanced data preprocessing techniques and feature selection work effectively to improve model accuracy. Despite the progress of the past several years, problems such as dataset limitations, class imbalance, and ethical considerations still exist which are causing difficulties for development. Future research needs to integrate diverse datasets, explore hybrid models, and adopt explainable AI techniques for diabetes prediction systems so they become applicable and reliable at all times in the clinical setting. By continuous innovation and improvement, machine learning can play a crucial role in the early detection and management of diabetes with obvious benefits to patient outcomes.

References

[1] M. Alehegn and R. Joshi, "Analysis and prediction of diabetes diseases using machine learning algorithm: Ensemble approach," *Int. Res. J. Eng. Technol.*, vol. 4, no. 10, pp. 426-433, Oct. 2017.

[2] M. Maniruzzaman, M. J. Rahman, B. Ahammed, and M. M. Abedin, "Classification and Prediction of Diabetes Disease Using Machine Learning Paradigm," *Health Information Science and Systems*, vol. 8, no. 7, 2020.

[3] B. Sridhara Murthy and J. Srilatha, "Comparative Analysis on Diabetes Dataset Using Machine Learning Algorithms," *International Journal of Engineering Research & Technology (IJERT)*, vol. 9, no. 09, pp. 921-930, 2020.

[4] A. Mujumdara and V. Vb, "Diabetes Prediction Using Machine Learning Algorithms," *Procedia Computer Science*, vol. 165, pp. 292–299, 2019.

[5] I. Tasin, T. U. Nabil, S. Islam, and R. Khan, "Diabetes prediction using machine learning and explainable AI techniques," *Healthcare Technology Letters*, vol. 10, no. 1, pp. 1–10, 2023.

[6] M. K. Hasan, M. A. Alam, D. Das, E. Hossain, and M. Hasan, "Diabetes Prediction Using Ensembling of Different Machine Learning Classifiers," *IEEE Access*, vol. 8, pp. 76516–76529, 2020.

[7] K. Lu, P. Sheth, Z. L. Zhou, K. Kazari, A. Guergachi, K. Keshavjee, M. Noaeen, and Z. Shakeri, "Identifying prediabetes in Canadian populations using machine learning," *Healthcare Technology Letters*, vol. 10, no. 1, pp. 1-10, 2023.

[8] M. Soni and S. Varma, "Diabetes prediction using machine learning techniques," *International Journal of Engineering Research & Technology (IJERT)*, vol. 9, no. 9, pp. 921-926, Sep. 2020.

[9] M. A. Sarwar, N. Kamal, W. Hamid, and M. A. Shah, "Prediction of diabetes using machine learning algorithms in healthcare," *Proceedings of the 24th International Conference on Automation & Computing*, Newcastle University, Newcastle upon Tyne, UK, Sept. 2018.

[10] N. P. Tigga and S. Garg, "Prediction of Type 2 Diabetes using Machine Learning Classification Methods," *Proceedings of the International Conference on Computational Intelligence and Data Science (ICCIDS)*, Birla Institute of Technology, Mesra, India, Sept. 2019, *Procedia Computer Science*, vol. 167, pp. 706–716, 2020.