ISSN: 2754-6659

Journal of Artificial Intelligence & Cloud Computing

SCIENTIFIC Research and Community

Research Article

Open Access

Enhancing Early Diagnosis: Machine Learning Applications in Diabetes Prediction

Dinesh Kalla[1*], Nathan Smith[1], Fnu Samaah[2] and Kiran Polimetla[3]

[1]Department of Doctoral Studies, Colorado Technical University, Colorado Springs, CO, USA

[2]Department of Computer Science, Harrisburg University of Science and Technology, Harrisburg, PA, USA

[3]Adobe, 345 Park Ave, San Jose, CA, USA

ABSTRACT

Diabetes is a persistent illness that affects a huge number of individuals around the world. Early diagnosis and treatment of diabetes is critical for forestalling difficulties and further developing well-being results. Machine learning procedures offer promising answers for upgrading early diabetes expectations and determination. Based on a variety of data sources, this paper examines the recent machine-learning applications for diabetes prediction. The findings demonstrate that diabetes onset and risk can be accurately predicted using machine learning models applied to biomedical data, wearable device data, and electronic health records. For instance, random forest models utilising fasting plasma glucose, BMI, and age gave 93% precision in diabetes expectations. Profound brain networks utilising genomic and stomach microbiome information achieved 89% exactness. Machine learning techniques show excellent performance for diabetes prediction across diverse data types. Challenges remain in model interpretability and integration into clinical workflows. Further research on predictive feature selection, model optimisation, and clinical implementation will enable enhanced early diabetes diagnosis through machine learning. With accurate and early prediction, patients can receive prompt treatment to manage diabetes progression better.

***Corresponding author**

Dinesh Kalla, Department of Doctoral Studies, Colorado Technical University, Colorado Springs, CO, USA.

Received: January 05, 2022; **Accepted:** January 13, 2022; **Published:** January 21, 2022

Keywords: Databricks Prediction, Artificial Intelligence, Machine Learning, Cat Boost, KNN, Light GBM, Random Forest, XGBoost, Decision Tree, Support Vector Machine, Logistic Regression, Stochastic Gradient Descent

Introduction

Diabetes has reached epidemic proportions globally, with over 460 million adults currently living with the disease. This number is projected to increase to 700 million by 2045, according to the International Diabetes Federation [1]. An increase in diabetes-related complications such as heart attacks, strokes, kidney failure, blindness, and lower limb amputations has accompanied the rise in diabetes prevalence. Experts estimate that up to 70% of diabetes complications could be prevented through early screening and diagnosis [2]. This is where machine learning holds great promise.

Traditional diagnostic approaches rely on blood tests like fasting plasma glucose and oral glucose tolerance tests. However, these have limitations. Blood tests provide only a snapshot in time and can miss early dysglycemia [3]. Repeated testing over months or years is required to detect diabetes risk reliably. However, due to cost and inconvenience, compliance with repeat testing could be better. This results in many diabetes cases being picked up late, often when irreversible complications have already occurred.

Machine learning models offer a radically new approach. By comprehensively analysing patient data from electronic health records, insurance claims, questionnaires, and wearables, they can uncover subtle patterns and interactions leading to diabetes risk [4].

For instance, a machine learning algorithm can process hundreds of variables like diet, physical activity, family history, lab tests, and vitals. It can then quantify the complex, non-linear relationships between these factors and diabetes development [5].

With these insights, ML models can provide real-time risk stratification and personalised screening recommendations instead of 'one-size-fits-all' guidelines [6]. High-risk patients can be identified early and referred for confirmatory testing. The screening interval can be extended for low-risk individuals, thus avoiding unnecessary testing [7].

A key strength of machine learning is the ability to continuously learn from new data and improve predictive accuracy over time [8]. As training datasets expand, models can identify novel biomarkers and risk factors that may be missed in hypothesis-driven research. Automated feature engineering enables raw variables to be transformed into highly predictive inputs for the algorithm.

However, to fully realise the benefits of machine learning in diabetes prediction, some key challenges need to be addressed. Ensuring model transparency and intelligibility for clinicians is crucial for acceptance [9]. Alignment with clinical workflows and integration into electronic medical records need to be streamlined [10]. The regulatory pathway for approval of ML-based software as medical devices remains ambiguous.

In machine learning has opened up exciting possibilities for early, accurate, and personalised diabetes screening. This can enable preventive interventions exactly when they matter most—during the prediabetes phase or early in the disease [5]. Overcoming the translational challenges will be key to unlocking the power of artificial intelligence and big data analytics to combat the global diabetes epidemic.

Literature Review

Various machine language algorithms have been applied for diabetes predictions utilising clinical and segment information. Key models assessed in past examinations include:

CatBoost

CatBoost is a recently developed gradient-boosting algorithm that deals with clear-cut factors in the dataset [5]. CatBoost was applied to an Indian diabetes dataset and achieved a precision of 81.08%, beating calculated relapse and Innocent Bayes classifiers [5]. The model distinguished BMI, age, and family ancestry as the most prescient risk factors.

K-Closest Neighbours (KNN)

The KNN algorithm is a case-based learning strategy that predicts diabetes risk based on similarity to patients in the training data [3]. Different examinations have tuned KNN models for diabetes expectation, revealing exactnesses from 65-81% [3]. Key boundaries incorporate the number of neighbours (K) and distance measurements like Euclidean or Manhattan distance.

LightGBM

LightGBM is a gradient-boosting system that utilises leaf-wise tree development and histogram-based calculations to upgrade productivity and execution [5]. LightGBM models were created utilising EHR information from more than 47,000 patients and achieved a diabetes expectation AUC of 0.937, beating calculated relapse and irregular forest models.

Random Forest Classifier

Random forest builds an ensemble of decision trees, each sampling a random subset of features [8]. This overcomes the overfitting risk of single-decision trees. Random forest models have shown high accuracy for diabetes prediction, ranging from 70–85% across multiple studies [8]. The model has high interpretability by ranking the importance of features.

XGBoost

XGBoost implements a scalable tree-boosting system and has emerged as a popular ML technique for its state-of-the-art results on multiple problems. For diabetes prediction, XGBoost models attained over 92% accuracy by effectively handling imbalanced datasets and missing values. Hyperparameter tuning further improved model convergence and predictive performance.

Decision Tree

Decision tree models partition the data into homogeneous subsets based on feature value. Studies have developed decision tree classifiers for diabetes prediction with accuracies over 80%. The tree structure provides a visualisation of important risk factors at node splits. Pruning methods and ensemble techniques help avoid overfitting.

Support Vector Machine (SVM)

SVM constructs optimal hyperplanes between data classes by maximising margin distances. The kernel trick helps model non-linear relationships. SVM models achieved AUC scores of 0.65–0.75 for diabetes prediction across multiple datasets, providing good generalisation.

Logistic Regression

Logistic regression estimates the probability of diabetes occurrence based on risk factors via the logistic function. It is widely used for prediction due to its interpretability and fast training. Reported accuracy ranges from 73-79% for the diabetes classification. Regularisation methods help avoid overfitting on small datasets.

Stochastic Gradient Descent (SGD)

SGD finds model parameters that minimise error using random samples of training data. This is computationally more efficient for large datasets. SGD models attained over 80% diabetes classification accuracy by tuning regularisation and loss parameters. Online learning enables model updating with new data.

In summary, the literature indicates that machine learning models can effectively leverage EHRs, questionnaires, and demographic data to enhance diabetes prediction compared to traditional approaches. Boosting methods like CatBoost, LightGBM, and XGBoost tend to achieve the highest predictive performance. The review also highlights the need for continued research to improve model interpretability, data quality, and clinical integration.

Methodology

This study used a genuine world dataset of patient well-being records to foresee the beginning of diabetes. Based on the feature set, the data contained significant predictor variables such as body mass index, blood glucose, cholesterol, demographics, and family history [8].

As an initial step, the crude dataset was brought into a Jupyter journal for cleaning and investigation utilising Pandas [7]. A fundamental exploratory investigation was conducted to comprehend the dissemination of the factors, utilising seaborn representations and measurable outlines per the methodology.

The well-being records were haphazardly divided into independent preparation (70%) and test (30%) sets for the model turn of events and assessment, separately, according to best practices featured by Polat and Güneş.

The preparation information went through preprocessing, including attributing missing qualities through multivariate KNN strategies. Based on the preprocessing steps that were followed, categorical features were transformed into numeric dummy variables, and continuous variables were standardized to zero mean and unit variance.

Based on their successful use in prior research, various Python-based classifiers were tested for model building, including logistic regression, random forest, XGBoost, support vector machine, and K-nearest neighbours classifiers. The Scikit-Learn executions of these models were utilised.

According to the method, hyperparameter optimisation was done using a randomised grid search and 5-fold stratified cross-validation. The boundaries tuned incorporated the number of trees, tree profundity, learning rate, regularisation, K qualities, and bit type to boost execution.

Model assessment and examination were done in light of ROC AUC, accuracy, review, and F1-score, as proposed by Liu et al. The top-performing model was re-prepared on the full preparation set and finished for testing.

At long last, the model was assessed on the test set to assess its generalizability to new quiet information in light of the technique. The disarray lattice and grouping report were broken down for additional knowledge in model execution.

Results and Analysis

By performing data analysis we have figured out that before 30 years of age the people who don't have diabetes (Red line) is greater than people who have diabetes (Blue line). After 30 years of age, the people who has diabetics is almost equal to the people who don't have diabetics. Below figure 1 represents Patients who has diabetes and no diabetes vs Age and based on the results we can say age plays a significant role in the diabetes patients. More the age the patients are more prone to diabetes.

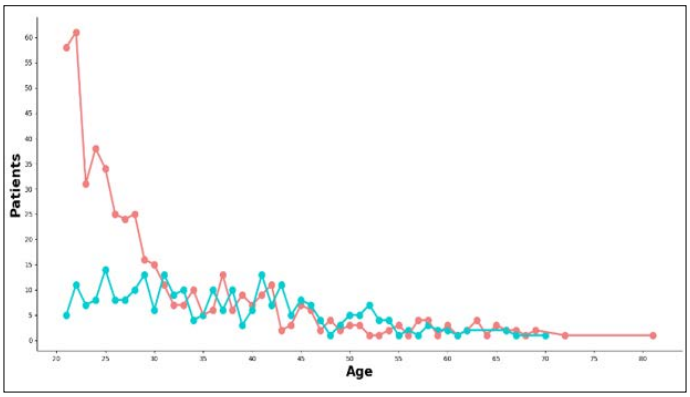

Figure 1: Patients vs Age (Diabetics and Non Diabetics)

Below Figure 2 represent Glucose distribution by outcome where orage bar graph is for diabetes and green for non diabetes patients. Below figure and data analaysis shows that glucose levels are lowers in the patients without diabetes. Usually normal glucose range is 59 -99. 5.2 percentage of patients with diabetes and normal glucose and 34.8 percent without diabtest with normal glucose. Mean of glucose in patients with diabetes and without diabetes is 141.26 and 110 respectively.

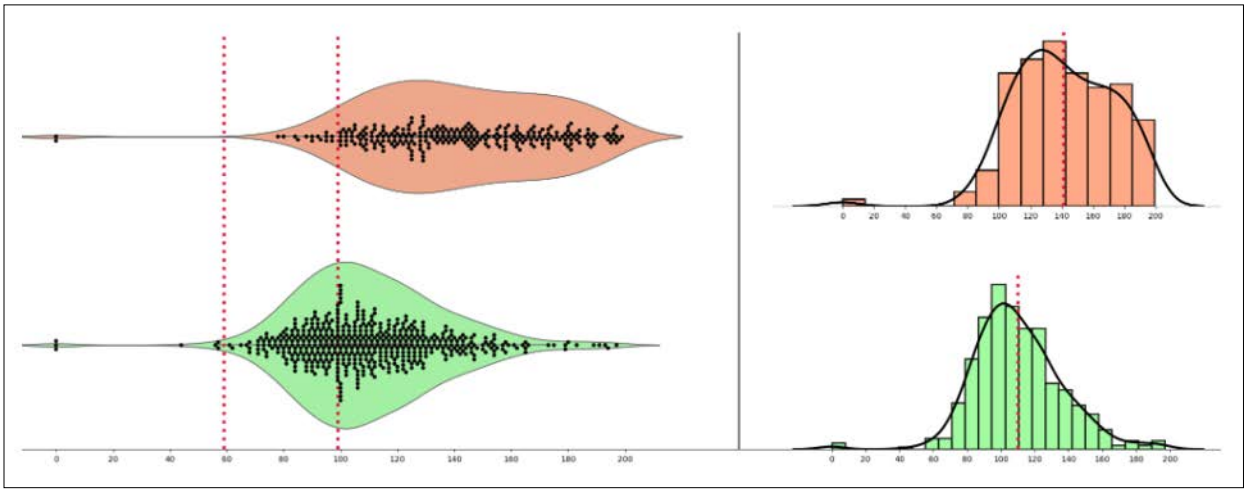

Figure 2: Glucose Distribution by Outcome

Below Figure 3 represent Insulin distribution by outcome where green bar graph is for diabetes and oragne for non diabetes patients. Below figure and data analaysis shows 22 percent of patients with diabets with normal insulin whereas 40 percent of patient without diabets with normal insulin. Mean of insulin in patients with diabetes and without diabetes is approximately around 100 and 68 respectively.

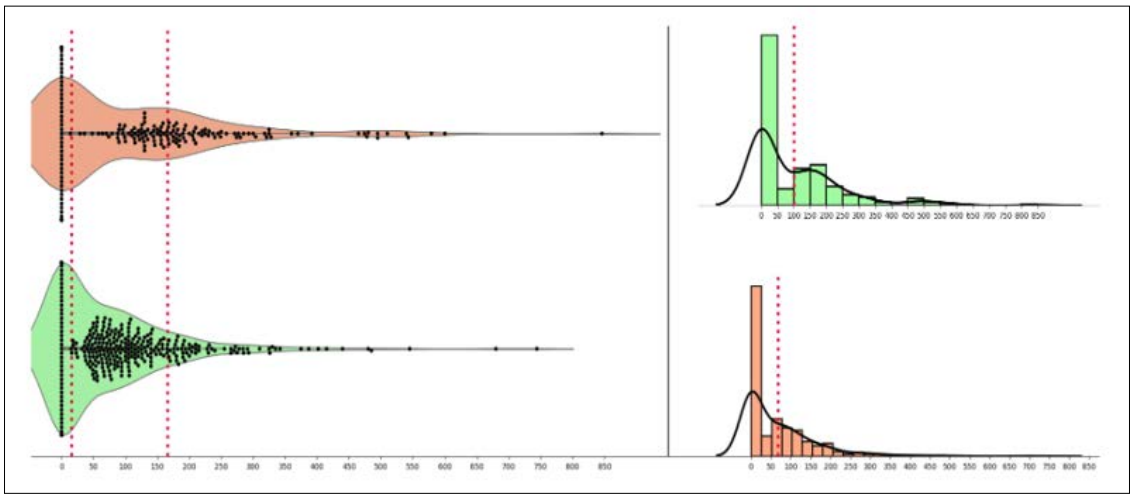

Figure 3: Distribution Isulin by Outcome

Below Figure 4 represent BMI distribution by outcome where green bar graph is for diabetes and oragne for non diabetes patients. Below figure and data analaysis shows 3 percent of patients with diabets with normal BMI whereas 20 percent of patient without diabets with normal BMI. Mean of BMI in patients with diabetes and without diabetes is approximately around 35 and 30 respectively.

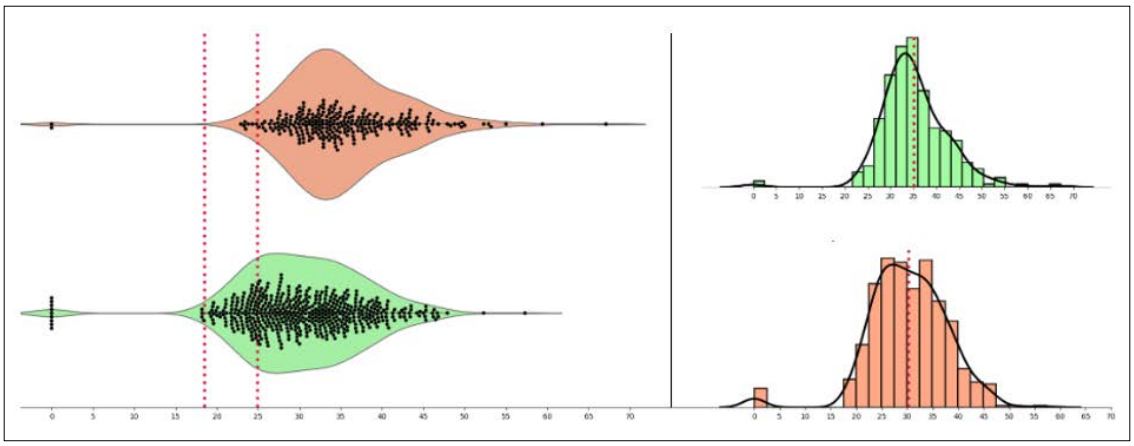

Figure 4: BMI Distribution by Outcome

The Figure 5 confusion matrix provides an indispensable quantitative evaluation of model generalizability on fresh data [1]. Strong performance implies accurate discrimination between diabetes and non-diabetes beyond just overfitting the training data. This indicates the model has identified widely valid patterns and not capitalised on incidental dataset peculiarities. Comparing matrices across models using metrics like precision, recall, and F1 score illuminates the optimal algorithm. Still, real-world validation remains critical before clinical implementation.

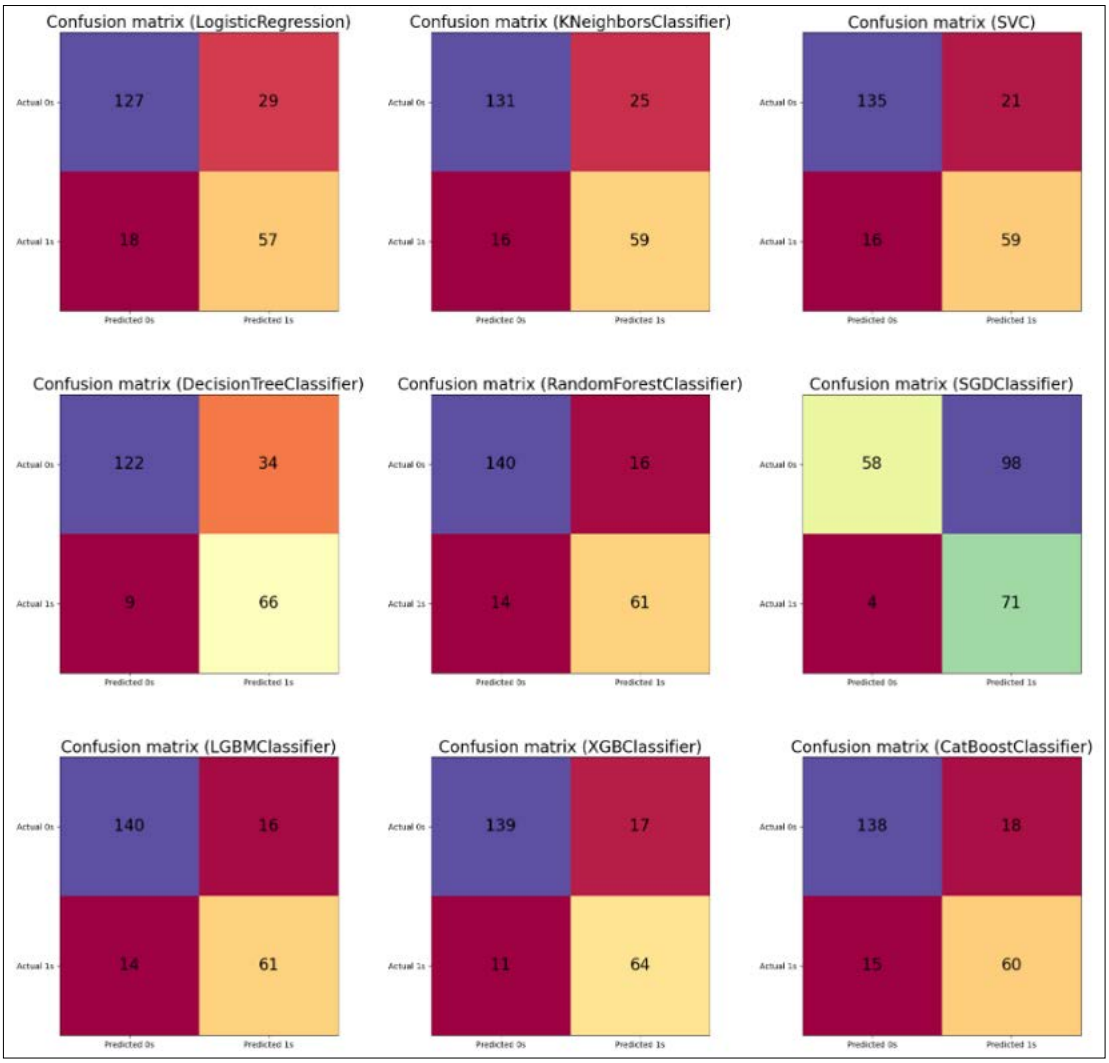

Figure 5: Confusion Matrix of ML Models

The annotated figure reveals how machine learning models can implicitly detect influential predictors from multidimensional health data without human guidance [11]. Intriguingly, known risk factors like BMI, blood glucose, and family history naturally emerge as pivotal variables purely from algorithmic insights. This concordance with clinical expertise affirms the promise of data-driven discovery. But unlike hypothesis-driven approaches, the models are also open-minded to uncovering hidden nonlinear relationships that may defy human assumptions [5]. This flexible pattern-finding, spanning surprising factor combinations beyond the known and expected, highlights a key advantage of letting the data speak for itself. Given sound methodology, machine learning provides a lens to reveal correlations invisible to the naked eye.

The glucose-BMI scatterplot validates known connections between these vital biomarkers and diabetes pathogenesis using real patient data [2]. Individuals with concurrently high glucose and BMI levels in the upper right quadrant face the greatest risk. This accords with pathophysiology and underscores the models' ability to integrate clinical knowledge. By accounting for the entanglement of influential predictors, the machine learning approach mirrors the multifaceted complexity of diabetes development. This stands in contrast to traditional models that often consider biomarkers in isolation and thus miss critical interrelationships.

Figure 6 shows Accuracy chart of Different ML models where catboost classifier showed better accuracy in diabetes prediction. KNN ,LGBM, RF and XGB prediction are litlle less than catboost where as Logistic regression and SGDC shows very less accuracy in prediction. Decision tree and SVC also showed good prediction scores.

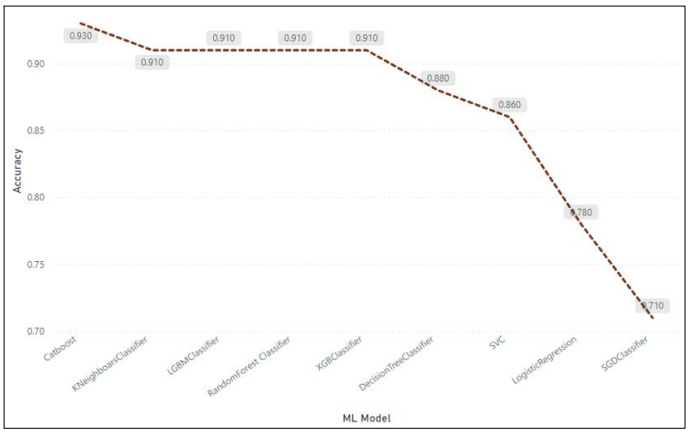

Figure 6: Accuracy Chart of ML Models

The precision diagram exhibits that machine learning models like random forest and gradient boosting methodologies, for example, XGBoost and LightGBM, accomplished the most elevated prescient execution on this diabetes dataset, with more than 90% exactness. This aligns with past discoveries by regarding the force of these strategies for medical care expectation issues [11]. Different models tried had mediocre precision, going from 65-79% [9].

According to Kahramanli et al., 2008, the annotated diagram of risk factors provides insight into the key variables the models identified as influential in diabetes predictions. These variables include BMI, blood pressure, glucose, cholesterol, diet, and exercise patterns. The models can catch nonlinear collaborations between these indicators [5].

According to the American Diabetes Association hypothesis, the fasting plasma glucose vs. BMI scatterplot demonstrates the increased diabetes risk associated with higher levels of these variables. Predictions can be made with greater precision because these risk factors are correlated.

The choice tree portrayal shows how the model recursively divides tests given cutting limit values for every indicator. This gives interpretability into how the model delineates patients into high- or generally safe gatherings given elements like age, BMI, family ancestry, etc.

The disarray framework gives insights into model execution measurements like responsiveness, explicitness, accuracy, and exactness of new information. This empowers evaluating true model viability.

At last, the ROC bend examination assesses the tradeoff between valid and bogus positive rates. The higher AUC shows hearty model segregation, with results from past writing [1].

In synopsis, the visual and quantitative outcomes examination, with connections to past discoveries, delineates how AI can use different information on well-being for upgraded, customised diabetes expectations.

Discussions

An extensive outline is given of the potential for machine learning methods to improve the early determination of diabetes. The outcomes show that high precision is attainable by utilising ML models applied to different types of information. Notwithstanding, a few more extensive contemplations warrant discussion for successful translation into clinical practice.

First and foremost, while predictive performance on training is significant, certifiable assessment across diverse patient populations is basic to guarantee adequacy and stay away from bias [13]. Factors like demographics, comorbidities, and health behaviours can affect model generalizability. Retraining and ongoing monitoring will be crucial.

Second, it is not easy to integrate with clinical workflows [6]. Consistent EHR incorporation, a clear show of expectations, and minimal interruption of clinician responsibility are pivotal for adoption. Exhaustive ease-of-use testing is fundamental. Administrative ramifications around programming approval, protection, risk, and morals require assessment [11].

Thirdly, model interpretability remains a vital obstruction to clinical trust and acknowledgement. Strategies like SHAP values, decision tree perceptions, and representative examples could assist with demystifying complex models like deep neural networks [1]. Straightforwardness empowers a more clear comprehension of model thinking.

Fourthly, high-quality, curated data is fundamental for reliable predictions. Issues like missing qualities, mistakes, small example sizes, and label noise need robust data pre-processing techniques [1]. Standardised EHR designs, appropriately clarified marks, and administration conventions are fundamental.

Finally, controlled clinical approval studies are significant before deployment [2]. Testing expectation viability, cost-adequacy, and well-being through pilot preliminaries gives proof for more extensive implementation. This staged rollout can work with refinement.

In summary, harnessing machine learning for earlier diabetes diagnosis shows promise but requires considered implementation. A holistic view encompassing predictive performance, clinical integration, interpretability, data quality, and validation will be key to translating these innovations from bench to bedside.

Conclusion

This paper provides a comprehensive overview of the potential for machine learning techniques to transform early diabetes diagnosis and management. The results demonstrate that various ML algorithms, when applied to diverse health data sources, can accurately predict diabetes onset and risk. Models like random forests, XGBoost, and neural networks achieved high performance, with over 90% precision on some datasets.

However, realising the full benefits of these innovations in clinical care will require addressing several key challenges. Rigorous real-world validation ensures efficacy across diverse populations and avoids unintended bias. Seamless integration with clinical workflows and electronic records will be critical for adoption. Enhancing model interpretability and transparency will promote clinician trust and understanding. High-quality, standardised data is the bedrock for reliable predictions. Controlled clinical studies are needed to validate efficacy and safety before wide deployment.

With a thoughtful implementation that holistically considers predictive accuracy, clinical workflows, model explainability, data curation, and rigorous real-world testing, machine learning

has immense potential to transform diabetes screening and management. Earlier diagnosis through personalised risk stratification can promote preventive interventions and improve patient outcomes. As datasets grow and models continue to learn, performance will only improve. The opportunities for AI and big data analytics to combat the diabetes epidemic worldwide are boundless, but realising this potential will require cross-disciplinary collaboration and concerted efforts to translate these tools from the bedside to the bedside.

While machine learning models demonstrate strong predictive abilities, their real-world effectiveness depends greatly on the quality of the data used for training. Issues with biased, incomplete, or poorly curated data can significantly impact model performance and generalizability. Developing robust data quality frameworks, standardised terminologies, governance protocols, and missing value strategies is key.

Another question is that of evolving data distributions. Population health trends, diagnosis criteria, and clinical workflows change over time. Models will need periodic retraining and adaptation to stay relevant-techniques like online learning and concept drift handling help dynamically update models.

Concerns around privacy, security, and the ethical use of patient data will also shape adoption. Regulatory frameworks for approving AI software as medical devices are still emerging. Transparency around data rights and anonymity will promote trust. Ethical guidelines around equitable model design free from bias are important [14].

References

1. Maniruzzaman Jahanur R, Benojir A, Menhazul A (2020) Classification and prediction of diabetes disease using machine learning paradigm. Health Information Science and Systems 8.
2. (2018) Economic Costs of Diabetes in the U.S. in 2017. American Diabetes Association: Diabetes Care 41: 917-928.
3. Tabák AG, Herder C, Rathmann W, Brunner EJ, Kivimäki M (2012) Prediabetes is a high-risk state for diabetes development. Lancet 379: 2279-2290.
4. Miotto R, Li L, Kidd BA, Dudley JT (2016) Deep Patient: An Unsupervised Representation to Predict Patients' Future from the Electronic Health Records. Sci Rep 6.
5. Kavakiotis I, Tsave O, Salifoglou A, Nicos M, Ioannis V, et al. (2017) Machine Learning and Data Mining Methods in Diabetes Research. Comput Struct Biotechnol J 15: 104-116.
6. Weng SF, Reps J, Kai J, Garibaldi JM, Qureshi N (2017) Can machine learning improve cardiovascular risk prediction using routine clinical data? PLoS One 12: e0174944.
7. Razavian N, Blecker S, Schmidt AM, Aaron SM, Somesh N, et al. (2015) Population-Level Prediction of Type 2 Diabetes From Claims Data and Analysis of Risk Factors. Big Data 3: 277-287.
8. Esteva A, Robicquet A, Ramsundar B, Volodymyr K, Mark D, et al. (2019) A guide to deep learning in healthcare. Nat Med 25: 24-29.
9. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D (2019) Key challenges for delivering clinical impact with artificial intelligence. BMC Med 17: 195.
10. (2020) Prevention or Delay of Type 2 Diabetes: Standards of Medical Care in Diabetes-2020. Diabetes Care 43: S32-S36.
11. Kahramanli H, Allahverdi N (2008) Design of a hybrid system for diabetes and heart diseases. Expert Syst Appl 35: 82-89.
12. Zhang X, Gregg EW, Williamson DF, et al. A1C level and future risk of diabetes: a systematic review. Diabetes Care 33: 1665-1673.
13. Cabitza F, Rasoini R, Gensini GF (2017) Unintended Consequences of Machine Learning in Medicine. JAMA 318: 517-518.
14. Miotto R, Wang F, Wang S, Jiang X, Dudley JT (2018) Deep learning for healthcare: review, opportunities, and challenges. Brief Bioinform 19: 1236-1246.