AOI: 10.100.234513.0264
13th international conference on Microelectronics, Circuits and Systems, Micro2026

A Study on Machine Learning and Explainable AI in Diabetes Prediction and Management

Tanya Singh, Rakesh Kumar, Sunil K. Singh, Sudhakar Kumar

School of ICT, Gautam Buddha University
Gautam Buddha Nagar, Uttar Pradesh, India.
Department of CSE, Chandigarh College of Engineering and Technology, Chandigarh, India
Corresponding author: tanyasingh9270@gmail.com; ORCID (R. Kumar): 0000-0002-5370-3071;
ORCID (S. K. Singh): 0000-0003-4876-7190; ORCID (S. Kumar): 0000-0001-7928-4234

ABSTRACT

Diabetes mellitus (DM) is a widespread metabolic illness. The International Diabetes Federation has predicted that the disease will affect 700 million individuals globally by 2045. The clinical burden of the disease is growing at an unsustainable rate with multiple AI-based predictive models demonstrating clinical promise, but few are actually being implemented in clinical practice. Machine learning has changed the landscape of diabetes prediction, and it is now possible to analyse heterogeneous clinical datasets at a scale that was previously not possible. This review summarizes data on datasets, algorithms, training protocols, and measures of evaluation to characterize the structural constraints to clinical translation. A Gradient-boosted model, Extreme Gradient Boosting (XGBoost) and an Adaptive Boosting model (AdaBoost), were found to work better on structured tabular data on the Pima Indians Diabetes Database (PIDD) and the Behavioural Risk Factor Surveillance System (BRFSS), whereas transformer architectures and Convolutional Neural Networks (CNNs) have become essential for Continuous Glucose Monitoring (CGM) time-series analysis. Although the Sylhet EMR dataset was included only for context purposes and not evaluated with respect to the cross-dataset architectural conclusions. Taken together, these results affirm that the bottleneck in algorithmic performance is no longer the dominant one. The other obstacles are structural, and federated learning, structural causal modelling, and model distillation must take the forefront in the future to decrease the burden of this widespread disease in the world.

Keywords: Diabetes, Machine Learning, Explainable AI, Diabetes Prediction, Diabetes Management.

I. INTRODUCTION

Diabetes mellitus no longer fits the profile of a regionally bounded condition; its reach has widened considerably over the past four decades, and the pressure it places on hospitals, insurers, and national health budgets is now difficult to overstate. According to the International Diabetes Federation (IDF), the general prevalence of diabetes among adults aged 18-79 years was 4.7% in 1980 and 9.3% in 2019 and is estimated to reach 10.9% and 700 million diabetic people worldwide by 2045 [1]. The cases are distributed very unevenly within that overall figure. The greatest current raw caseloads are in China, India and the United States, but the fastest growth rates are occurring in other regions: Europe is set to increase by 15% and Africa by 143% in undiagnosed cases [1]. This disparity has practical implications because the regions with the most significant acceleration in the epidemic dissemination have the least amount of specialist clinicians. Unless there are scalable, automated screening solutions that can operate within strict resource limits, large volumes of patients will be silently advanced to end-stage complications incurring costs, both physiologically and economically, that health care systems in those areas cannot sustain.

A. The clinical burden: Complications and failure of conventional diagnostics

Chronic hyperglycaemia damages several organ systems in the body in a significant and permanent manner, which leads to a stress on the importance of early intervention in the disease process. DM is a disease that develops mainly due to two most significant metabolic catalysts: insulin resistance and destruction of pancreatic beta-cells. Uncontrolled, such a metabolic dysfunction will result in long-term complications and comorbidities like retinopathy, neuropathy, cardiovascular disease (CVD), and renal failure [2], [3]. Type 1 Diabetes Mellitus (T1DM) is among the most prevalent types of diabetes that affect 5-10% of the diabetic population and is growing internationally at a rate of 3% per year, often leading to a drastic decrease in life expectancy. Clinicians are only left with reactive diagnostic skills under the current diagnostic methods. Type 2 Diabetes Mellitus (T2DM) and Gestational Diabetes Mellitus (GDM) involve a long silent phase where physiologic damage is accrued in the long term before symptoms manifest themselves [4], [5]. Traditional techniques like HbA1c or fasting glucose level tests often miss subtle, early-stage pre-diabetic signatures, and this results in a multi-billion-dollar healthcare liability that can only be overcome by proactive, probabilistic monitoring. It is thus necessary that the field adopt predictive biomarker signatures that can detect the risk several years before clinical thresholds are violated.

B. The Artificial Intelligence Revolution: Technology Potential vs. Implementation Lapses

ML and DL have initiated a paradigm shift of high-precision predictive insight systems and systems of

ISBN: 978-81-689089-0-1

clinical decisions [6], [3], [4]. Wearable technology and Internet of Things (IoT)-aware frameworks contribute to this revolution and enable the real-time monitoring of glucose and autonomous insulin adjustment [7], [8], [9], [10], [11]. Recent preprocessing literature highlights that imputation strategy choice significantly affects downstream model performance in diabetic datasets, [12]. Evidence demonstrates that early intervention decreases complications of diabetes [13]. In practice, however, adoption among clinicians has been much slower than what published accuracy figures might suggest, a gap that goes beyond technical limitations and has more to do with how healthcare institutions are structured and how regulatory approval actually works. In cases where the DL model only provides a prediction score without explanation, there is no basis or regulatory protection to act on it, and thus interpretability is not an optional enhancement but a strict requirement for bedside deployment [14], [17]. In combination with this, pure ML and pure DL both have blind spots that can be filled by hybrid designs, such as the feature-extraction depth of a CNN and the tabular robustness of a Random Forest. Similarly necessary types of architectural mixing also prove to be practically required at the edge, where RAM budgets and inference-time ceilings do not permit models where efficiency is traded off against accuracy [5], [9], [26].

C. Strategic Objective: Accuracy and Efficiency Benchmarking

The paper compares the clinical accuracy and computational efficiency of different ML/DL architectures to determine the most appropriate architectures, and evaluate the interpretability and deployment gaps that will prevent real-world clinical use. The study is organized in the form of two key pillars:

- Precision and consistency: Comparison of traditional ML, deep neural networks (CNNs, RNNs), and hybrid architectures across a wide range of clinical settings to test model resilience [3], [4].
- Computational Efficiency: Measuring the inference time and memory consumption (RAM) to evaluate the feasibility of models to run on healthcare edge computing devices [7], [8].

Five datasets are used to test the models on the complete range of the disease: the PIDD (Datasets 1 and 2), Behavioural Risk Factor Surveillance System (BRFSS) to survey populations (Datasets 3 and 4), Sylhet Diabetes Hospital (Bangladesh) data (Dataset 5). The Sylhet data is meaningful as it is one of the initial symptomatic prediction points, and it offers a good contrast to the general clinical measures [4].

Evaluation Key Performance Indicators (KPIs):

- F1-Score: It is the key metric chosen for its balance of precision and recall in cases of the extreme class imbalance of diabetic datasets (especially BRFSS), and thus reducing the risk of false negatives.
- The area under the receiver operating characteristic curve (AUC-ROC): AUC-ROC was selected because it captures classifier discrimination across the full decision threshold range, a practical necessity when class distributions are as skewed as they are in most diabetes screening datasets, where a single-threshold metric would obscure meaningful variation.
- Inference Time (IT): Latency of prediction is vital in monitoring applications that require real-time response.
- Memory Usage (RAM): Determines hardware specifications to be used.

II. METHODOLOGY

The coverage was restricted to 2022-2025, a range that was selected to cover work reflecting the current generation of architectures and benchmarks, and older papers were only included where they provided contextual information not repeated in more recent literature. A variety of databases were consulted: IEEE Xplore, PubMed, Scopus, and Google Scholar. Search terms were applied both independently and in combination using Boolean operators (e.g., “diabetes prediction” and “machine learning” OR “edge computing”). Queries performed on these databases yielded a total of 120 titles, and queries conducted on additional databases yielded an additional 10 records, producing a total of 130 records. Deduplication procedures resulted in a total of 110 records. A Title-and-Abstract screen eliminated another 65 studies which failed to satisfy the criteria of the scope, and 45 articles were included in the full-text evaluation. Out of them 15 were not included due to the lack of quantitative evaluation metrics, were not diabetes-centred, or were opinion pieces or editorials. The rest of the 30 studies were taken forward for full synthesis and classified by architectural family, type of data and modality of deployment - a framework employed in the comparative analysis in Sections III-VII. To qualify, a paper had to be peer-reviewed, in the English language, and contain empirical or review data on AI used to predict diabetes, managing, or deploying AI in clinical settings.

III. ALGORITHMIC EFFICACY: ENSEMBLES VS. DEEP LEARNING VS. AUTOMATED MACHINE LEARNING (AUTOML)

A. Predictive Modelling: ML, DL, and Hybrid Architectures

Surveying the historical trajectory of AI in diabetology, a notable trend emerges: a steady shift toward less and less interpretable, single-stage classifiers, in favour of multi-layered architectures, which sacrifice transparency to the advantage of representational capability. To serve our needs in this review the landscape can be broken down into three broad families:

- Traditional ML: Logistic Regression (LR), Random Forest (RF), Support Vector Machine (SVM), XGBoost, and Adaptive Boosting (AdaBoost). They

ISBN: 978-81-689089-0-1

are crucial because of their interpretability and efficiency in resource-constrained edge settings.

- Deep Learning (DL): CNN, Recurrent Neural Network (RNN), Long Short-Term Memory (LSTM), and Gated Recurrent Unit (GRU) are good at identifying the temporal interactions and non-linearities of multiclass sequential data.
- Hybrid/Ensemble: Stacking and bagging strategies leverage meta-learners that aggregate base model outputs, reducing overfitting and improving generalization on heterogeneous data [1].

Optuna-based hyperparameter tuning is worth mentioning as a repeated methodological decision. Tree-structured Parzen Estimator (TPE) guides sampling toward configurations that recorded preliminary promise, rather than exhaustively testing all alternatives within a fixed grid space [3]. This method is much faster in construction of advanced models as opposed to the traditional means which demand the testing of the models. In order to have organized and tabular data and discrete clinical as well as anthropometric data, gradient boosted trees and ensemble strategies have continuously shown better performance at reduced calculations [1], [4]. Comparative studies have established that RF and XGBoost offer the highest ratio of speed to accuracy in small tabular data, mostly because of their natural anti-overfitting properties. Pairing gradient boosting with neural network layers in a stacked ensemble arrangement has been shown to yield meaningfully better sensitivity to cardiovascular risk signals within diabetic cohorts than either component achieves independently. A transition predictive model based on gestational diabetes to T2DM exhibited good results when trained with AdaBoost [9], whilst a strong result of 81% was obtained when extreme gradient boosting was used with adaptive synthetic sampling on heavily imbalanced clinical data [6]. Conversely, high dimensional and non-linear data produced by physiological monitoring and imaging require DL architectures. The most popular among the architectures is CNNs, which are significantly better designed to screen diabetic retinopathy than the traditional ML in the extraction of complex pixel level features [7]. Similar methods have been used on deep artificial neural networks to predict downstream diabetic complications, such as nephropathy and neuropathy, by estimating the localized risk of interconnected and complex biomarkers [8]. To reduce the distance between traditional ML and complex DL, AutoML systems, including H2O AutoML and AutoGluon, fully automate the hyperparameter optimisation and model selection, producing stacked ensembles with high accuracies (e.g., 85.01%-86.5%) at a considerably lower technical barrier to clinical deployment [9], [10]. Taken as a whole, these results indicate that model selection should not be viewed as a purely technical choice; in fact, the dataset characteristics (e.g., those relating to missingness, skewness and causal relations) will strongly influence the final model performance in the same way as the algorithm will.

IV. DATA IMPERFECTIONS, IMBALANCE, AND CAUSAL PARADIGMS

A. Data Engineering and Preprocessing Protocols

Before any model sees a single record, the decisions made during data preparation like what to impute, how to handle outliers, how to balance classes, quietly determine much of what the downstream results can and cannot show. Even in clinical data, zero values in variables like BMI, Insulin, or Blood Pressure often represent missing systematic values, instead of an actual physiological value, although it may have numerical meaning.

B. Imputation and Class Balancing

This review made a comparison of Minimum Imputation and Median Imputation. The findings indicate that Minimum Imputation appears to more often give better performance of the model across the datasets considered. This result confirms the hypothesis that, in most clinical environments, missing values could be a sign that the physiological values of a patient fell within the normal range and were thus not recorded as an abnormal outlier.

To ensure that the bias in the model arising as the result of class imbalances was eliminated, two different protocols were implemented:

- Adaptive Synthetic Sampling (ADASYN): ADASYN can enhance the robustness of models in areas of low density by creating artificial samples around the boundaries of classes [6], [15].
- Clustering-based Under sampling: Under sampling used on large-scale BRFSS data. With the K-means algorithm, the instances in the majority classes were clustered into representative centroids, maintaining the important distributional patterns without excessive data volume consumption, thereby greatly reducing the amount of data that should be calculated [7], [15].

To deal with the extreme physiological outliers (e.g. biologically implausible values of glucose, or BMI) log-sinh and adaptive Box-Cox transformations have been proposed to normalize feature distributions before Lightweight Convolutional Pyramid Squeeze Attention (LwCPSA) networks are extracted [16]. Standard ML models are based on the assumption that data is independent, and designed to optimize correlation as opposed to causality, thereby reducing their resilience to domain shifts or unforeseen clinical interventions. In response to this, causal discovery algorithms (e.g., LiNGAM) are proposed in conjunction with causal inference models (e.g., DoWhy), which can estimate the actual Average Treatment Effect (ATE) of individual variables, as well as give biologically sound understanding of how diabetes risks occur [17]. The applicability of strong forecasting structures is also shown in predicting the prevalence of diabetes in the world, where Transformer architectures with Variational

ISBN: 978-81-689089-0-1

Autoencoders (VAEs) show greater resistance to noisy and non-available global health statistics compared to more classic AutoRegressive Integrated Moving Average (ARIMA) statistical models [18]. Higher data quality and better-modelled distributions of classes increase the limit to which a model can perform, but do not in themselves render that model applicable at the bedside. Clinicians must know not only that a risk score was produced, but why that score is produced at that level, that is what has underpinned a new body of interpretability studies reviewed in the following section.

V. EXPLAINABLE AI, COUNTERFACTUALS, AND LLM INTEGRATION

A. Bridging the Interpretability Gap

Whenever a model fails to explain what it produces, clinicians do not trust it and the regulators do not accept it. The response to this problem is a two-pronged research frontier [14]:

- Explainable AI (XAI): Frameworks like SHapley Additive exPlanations (SHAP) and Proximity-Constrained Counterfactuals, which supply clinicians with actionable explanations of whatever risk profile is generated, are required as an industry standard [15], [19].
- LLM-based Support: Systems such as the IMPACT framework and personalized chatbot interfaces render complex model outputs into natural-language treatment recommendations [20], [21].

By the mid-2020s, SHAP and Local Interpretable Model-agnostic Explanations (LIME) had become the dominant benchmarks for feature attribution in clinical ML literature [10], [14], and the bar has been exceeded. The static attribution covers the question, “which features most influenced the prediction?” a fine beginning but not enough to take action. Counterfactual Analysis (CA) further extends the question, formulating scenarios called minimal-edit that give answers to the question of what would need to be different, in order that this patient would be given a non-diabetic prediction. This is operationalised in the form of Diverse Counterfactual Explanations (DiCE) that uses human feedback constraints in order to maintain the generated options as clinically plausible [19]. A Non-dominated Sorting Genetic Algorithm II (NSGA-II) may be used, where a patient has many risks at once, to generalize the same reasoning to comorbidities, thus, a Pareto-optimal set of interventions, which combines to decrease both risk of diabetes and stroke may be generated. When integrated into LLM-based chatbots (e.g., BioMistral-7B), these outputs are then transformed into the plain-language instructions that patients and generalist clinicians can follow [20], [21]. Clinical utility only matters when inference can occur during the point of care. The hardware needed to support this kind of local processing must be able to execute non-trivial models without necessarily sending them to the cloud to be processed and this is precisely what edge-ai research is trying to offer.

VI. REAL-TIME EDGE COMPUTING AND IOT ECOSYSTEMS

A key direction in this space is continuous non-invasive monitoring via Edge-AI and wearables, for example, ECG-based glucose tracking enables real-time T1DM management by moving inference to the device itself, allowing intervention before deterioration occurs [25], [26]. Predictive modelling is no longer the evaluation of the past electronic health records (EHR) but active, real-time surveillance made possible by the wearable and the Internet of Healthcare Things (IoHT) [22]. The transfer of foreseeable algorithms into a full digital ecosystem of a Virtual Ward allows a patient to be stratified constantly even without being present in the laboratory [23], [24]. An example: blood glucose patterns can be predicted on the fly and insulin delivery corrections automatically modified and the services of a specialist are not required in the loop when Deep Reinforcement Learning (DRL) is used with Bidirectional LSTMs [25]. When memory-consuming models are deployed to embedded hardware, architectural compression necessitating the use of Edge-AI techniques is aggressive and would therefore remove cloud latency and guarantee data privacy. The researchers have built a highly-optimized 1-D CNN to detect non-invasive diabetes using an ECG spectrogram; after the algorithmic quantization step, the algorithm was deployed on an STM32F401 microcontroller with a minimum flash memory of 347 kB [26]. Equally, a Temporal Fusion Transformer (TFT) was implemented on a Bluetooth Low Energy System-on-Chip (SoC) to provide multi-horizon glucose predictions with a mean power consumption of under 2 mW [27]. With foundation models like CGMformer, which are pre-trained on large unsupervised corpora of continuous glucose monitoring, can be fine-tuned and perform localized predictions without looking beyond the streams of wearable data [28]. For these Edge-AI models, empirical validation is done using conforming benchmarks like the open D1NAMO multimodal dataset [29], and the REPLACE-BG clinical trial parameters [30], and the pipeline of integrated deployment across patient, edge and clinician layers is shown in the figure below (Fig. 1). Combining the findings in this section, one can posit that the engineering gap between what researchers can create in the laboratory and what can be run on a wrist-worn sensor that is actually decreasing at a rate even greater than what the literature had predicted even three

ISBN: 978-81-689089-0-1

years ago. Section VII compares the performance figures of the four previous areas.

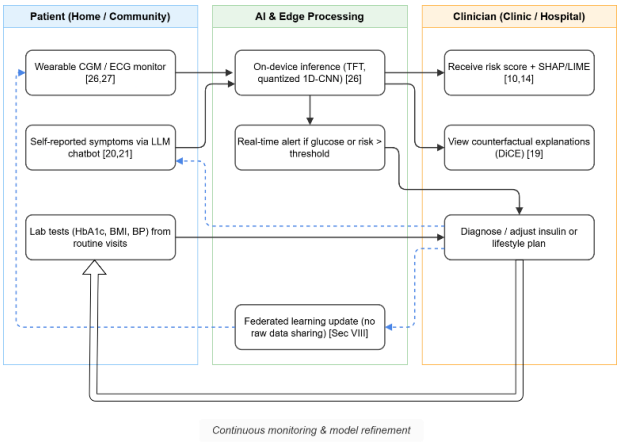

Fig. 1. Clinical deployment pipeline: End-to-end pipeline to deploy on patients, processors, and clinicians with explainable AI results, Edge-AI inference, and federated learning.

VII. COMPARATIVE ANALYSIS OF PREDICTIVE MODELS

A. Key Performance Indicators

The primary clinical diagnostic KPI that medical workers rely on most, presents its outcomes in the form of F1 Score. Accuracy measurement displays the percentage of both positives and negatives that model defined correctly. In the F1 score, false positive and false negative errors are balanced [13]. Rather than pointing to a single dominant approach, Table I exposes a more fundamental divide running through the model families themselves. On the one hand, there are AdaBoost and XGBoost, lightweight ensemble approaches. The lines that separate pre-diabetes and active diabetes are significantly smooth, that can scale to tabular data, while being confined to memory and latency resources that edge hardware imposes. Conversely, it is specifically on time-series and tasks based on ECG that the TFT and quantized 1D-CNN are justified because any classical approach will never be as time-intensive or even representative-rich as both require compression pipelines before they exit the server room. H2O AutoML occupies a curious in-between state: with an AUC of 0.909 (the highest-water mark of the table), but unable to distinguish prediabetes and active diabetes, H2O AutoML is a clinical disqualifier, which demonstrates why headline accuracy is an imperfect readiness metric. The practical implication is one that the whole of this review has been moving toward: no single architecture generalizes cleanly across all data conditions, and any selection that ignores deployment constraints is a category error.

TABLE I
COMPARATIVE ANALYSIS OF PREDICTIVE MODELS

Ref.	Model	Dataset	Key Metrics	Limitations
[5]	RF, DNN, LSTM, AdaBoost	PIMA, BRFSS	AdaBoost F1: 0.74	DNN computationally expensive; struggled with generalisation on imbalanced tabular data.
[8]	Hybrid Ensemble (XGBoost meta-learner)	Multiple CVD Datasets	Hybrid AUC: 0.82	High computational overhead restricts rapid deployment on resource-constrained devices.
[11]	Deep Neural Network (DNN)	Diabetes Complications Dataset	DNN AUC: 0.833	Treats all retinopathy cases homogeneously; no severity-staging stratification.
[13]	AutoML (H2O) Stacked Ensemble	NHANES 1999–2020	AUC: 0.909, Acc: 86.5%	Poor multiclass performance; cannot distinguish prediabetes from active diabetes.
[16]	LwCPSA (Lightweight CNN + PSA)	PIMA, LMCH	LwCPSA+GRO Acc: 99.65%	High feature variance required aggressive log-sinh and adaptive Box-Cox pre-transformations.
[17]	LiNGAM + DoWhy (Causal Inference)	Kaggle NIDDK	LogReg/DL Acc: 84.8%	Causal discovery depends on large observational datasets without guaranteed pre-treatment equivalence.
[19]	MLP + LIME/DiCE (Counterfactuals)	Diabetes Health Indicators	MLP Baseline Acc: 83.7%	Assumes feature independence; risks biologically unlikely counterfactual scenarios.
[20]	NSGA-II + Gemini Pro LLM	Stroke & Diabetes	Pareto multi-objective optimisation	Currently limited to a pre-defined dual-disease parameter space.
[26]	1D-CNN (Quantised)	D1NAMO (ECG Spectrograms)	1D-CNN Acc: 89.52%	Strict memory footprint [347 kB flash, 23 kB RAM] for STM32F401 Edge-AI deployment.
[27]	Temporal Fusion Transformer (TFT)	OhioT1DM, ShanghaiDM	RMSE: 14.7 mg/dL	Computationally intensive; requires significant SoC optimisation.

ISBN: 978-81-689089-0-1

B. Performance Synthesis by Dataset Category

The Logistic Regression models of the BRFSS data sets generated findings with classification measures significantly lower than the no-skill baseline - similar to a negative R 2 in regression terms - showing that the survey-level data of populations have non-linear trends that cannot be captured by linear decision boundaries. The failure of linear predictors to explain such variation proves the point that the lines that separate pre-diabetes and active diabetes are highly non-linear and that we should use the non-linear feature mapping provided by the DL models to do this. Without the Sylhet dataset, where perfect accuracy is due to structured binary symptoms and putative preprocessing leakage, rather than genuine model superiority, the rest of the datasets affirm the importance of dataset origin as no less important than the choice of algorithm. On externally validated datasets, the correct preprocessing pipeline is always found to be superior to algorithmic choice - a result that is corroborated across PIMA and BRFSS results.

The ideal performance on the Sylhet data is subject to question. The binary symptom surveys it uses form a remarkably clean feature space on which discriminative features such as polyuria and polydipsia can classically divide the classes, making it a property of the data and not the model. This outcome can be further exaggerated by possible data leakage due to preprocessing prior to splitting. It is not to be included in cross-dataset conclusions until externally validated. The converse is observed in case of BRFSS: the optimal F1-score achieved by any architecture with self-reported population surveys is below 0.45, with RNN and DNN architectures carefully selected to be able to handle nonlinear trends. It also measures the degree of noise and ambiguity of the data, which even the best models cannot fully overcome. PIMA is in-between them; the F1 values of 0.73-0.74 that AdaBoost and Random Forest can get at that point are practically a class-imbalance ceiling as opposed to model-expressiveness ceiling. Collectively, these findings support one important fact that one might be quick to dismiss in the benchmark literature is that the right preprocessing pipeline and the right data source can be more important than the right model.

TABLE II
PERFORMANCE SYNTHESIS BY DATASET CATEGORY

Dataset Type	Top Model	F1 / Acc
Clinical (PIMA-768)	AdaBoost	0.7438 (F1)
Clinical (PIMA-2000)	Random Forest	0.7273 (F1)
Survey (BRFSS Multiclass)	RNN	0.4414 (F1) / 0.7144 (Acc)
Survey (BRFSS Binary)	DNN	0.4500 (F1)
EMR (Sylhet Hospital) — Perfect accuracy; no outside validation; interpret with caution	Random Forest	1.0000 (Accuracy)

C. Computational Efficiency: Inference Time and Resource Consumption

Model depth and prediction speed must both be considered to achieve real-time clinical deployment. To be of practical clinical use the inference latency of one model should be sufficient to satisfy the demands of real-time monitoring use. The number of parameters (NoP) is a double-edged consideration: on one hand, the higher the NoP the greater the capacity to model complex BRFSS data, on the other hand, it results in a very real possibility of overfitting smaller clinical datasets like PIMA. Lightweight ML models (AdaBoost) often tend to perform better in the small sample regime than heavy DL models (RNNs), and their generalization and computational parsimony tend to be better. Table III gives the cost of depth in concrete terms. Bagging RNN was hundred times slower than Random Forest and consumed thirteen times more memory, yet yielded almost no benefit by the standards of tabular data, which Table II indicates is hardly much to take into account. AdaBoost presents its own riddle: although it has fewer than 1,000 parameters and a memory footprint of zero, it is in fact slower than Random Forest. In the case of a wearable device based on a coin-cell battery and a few kilobytes of flash memory, that sequential dependency is as disqualifying as the memory hunger of the RNN. It is only Random Forest of the three that are being benchmarked here that passes across to clear the speed and memory bars, without post-training quantization. The wide disparity between the performance requirements of the best-performing deep models and the hardware capabilities of the most constrained hardware is the central tension that is the subject of model distillation research, mentioned in the following section (VIII).

TABLE III
PERFORMANCE VS. RESOURCE TRADE-OFFS

Model	IT (s)	RAM	NoP
Random Forest (Mid)	0.42	36.0 kB	6,269
AdaBoost (Lightweight)	1.18	<1 kB	987
Bagging RNN (Heavy)	42.06	492.0 kB	113,785

VIII. DISCUSSION AND FUTURE DIRECTIONS

Combined over the entire sections, the evidence suggests few doubts that the current AI possesses the rudimentary forecasting capabilities required to really make a difference in the overall treatment of diabetes early diagnosis, prediction of complications, and autonomous control of metabolic levels are all meaningful and manageable now. What the evidence also demonstrates though is that the performance of algorithms would not be the bottleneck in isolation, trust is. A model whose reasoning cannot be interrogated by clinicians will not earn clinical trust and a model that has been trained on a North American or European population might yield actively harmful predictions when used in sub-Saharan Africa or South Asia. Excessive dependence upon

ISBN: 978-81-689089-0-1

synthetic augmentation techniques (e.g., SMOTE) and only retrospectively acquired training datasets create more problems than solutions with regard to generalizability. The risk associated with this issue is exemplified in the Sylhet dataset; perfect classification accuracy of 1.000 exists due to the clean binary separation of the symptom categories and probable pre-test/pre-train data leakage, and this has specifically been ruled out for use as a basis for reaching cross-datasets architectural conclusions in the review article. Future external prospectively validated cohorts of patients from the Sylhet cohort for instance are an open area of research at this time.

The following technically sophisticated directions are recommended for future research:

- External Validation through Federated Learning: It is no longer useful to validate solely on proprietary datasets. The following step involves the use of a decentralized federated training across geographically and demographically diverse Electronic Health Record (EHR) repositories, which adds phenotypic heterogeneity through design choice, without sharing patient records.
- Resource-Aware Distillation for Edge-AI: Transformer or LSTM architectures in their entirety are impractical to port onto wrist-worn or implantable devices because of the limitation of battery capacity and on-chip memory. A more promising approach that provides a principled pathway to sub-milliwatt, always on monitoring, without localising performance collapse that naive pruning normally produces, is to use a compact 1D CNN or spiking neural network to mimic the output of a large teacher model.
- Integration of Structural Causal Models (SCMs): Presently the techniques in the area of XAI (SHAP, LIME) are purely associational. Future counterfactual synthesis under frameworks like DiCE or NSGA-II must be carefully constrained by SCMs and DAGs to ensure that algorithms do not prescribe harmful lifestyle changes.

IX. CONCLUSION

This review demonstrates that no single architecture, dataset, or explainability tool will automatically make diabetes AI leave the research lab and enter into routine clinical care. What is desired and what is now showing itself is a conscious combination of complementary elements. Three practical conclusions follow. First, in the case of model selection depending on the dataset, no universal solution exists: Ensemble architectures such as Random Forest and AdaBoost appear to perform better than deep learning architectures on structured clinical datasets due to the fact that less computing resources are needed to achieve a good accuracy-efficiency trade off; however, when it comes to large survey datasets at the population level (e.g., BRFSS), RNNs are necessary to account for the non-linearities in the data that cannot be captured using linear/shallow learning models. Second, the XAI role should not be compromised, as with interpretability established as a prerequisite for clinical adoption, future pipelines must incorporate XAI layers, which will guarantee that AI-aided decisions are auditable and implementable by clinical professionals. Third, the way ahead is the Edge-AI transition, with the global diabetic population approaching 700 million people, non-invasive continuous patient healthcare at the edge is the only viable scheme of scalable healthcare infrastructure [22], [27].

On structured clinical tables, gradient-boosted ensembles explicitly can resolve the accuracy-efficiency trade-off sufficiently to deploy such ensembles; with wearable sensor streams, quantized time architectures are rapidly closing in. The interface between technical and institutional systems is the hardest barrier to overcome, ensuring that the outputs of XAI are trusted by the regulators, ensuring that federated protocols are adopted by hospitals across networks, and ensuring that distilled edge models are validated on the same prospective, multi-centre basis on which any other clinical diagnostic tool is validated. The dissonance between demonstrated ability and actual clinical use will be greater than it should be until the community changes its primary metric from held-out test-set accuracy to external validation. Sylhet EMR results are reported for descriptive purposes only and are explicitly excluded from all cross-dataset architectural recommendations due to suspected preprocessing leakage prior to train-test splitting. Also mentioned in this review is the danger of magnifying invalid benchmark scores; the Sylhet perfect-accuracy number is representative of the potential of structured clinical data to generate ceiling-level scores which fail to extrapolate to various real-life groups.

REFERENCES

[1] P. B. Khokhar, C. Gravino, and F. Palomba, “Advances in artificial intelligence for diabetes prediction: insights from a systematic literature review,” Artif. Intell. Med., vol. 164, p. 103132, Apr. 2025.
[2] R. Mittal et al., “Harnessing Machine Learning, a Subset of Artificial Intelligence, for Early Detection and Diagnosis of Type 1 Diabetes: A Systematic Review,” Int. J. Mol. Sci., vol. 26, no. 3935, Apr. 2025.
[3] Y. Li et al., “Machine learning-based risk predictive models for diabetic kidney disease in type 2 diabetes mellitus patients: a systematic review and meta-analysis,” Front. Endocrinol., vol. 16, p. 1495306, Mar. 2025.
[4] S. Corrao, M. Janic, V. Maggio, and M. Rizzo, “Machine learning and deep learning in diabetology: revolutionizing diabetes care,” Front. Clin. Diabetes Healthc., vol. 6, p. 1547689, May 2025.
[5] O. B. Ayoade, S. Shahrestani, and C. Ruan, “Machine Learning and Deep Learning Approaches for Predicting Diabetes Progression: A Comparative Analysis,” Electronics, vol. 14, no. 2583, Jun. 2025.
[6] G. A. Ansari, S. Shafi, M. D. Ansari, and A. Shadab, “Advanced supervised machine learning methods for precise diabetes mellitus prediction using feature selection,” Front. Med., vol. 12, p. 1620268, Sep. 2025.
[7] Y. Barzegar et al., “Machine Learning Pipeline for Early Diabetes Detection: A Comparative Study with

ISBN: 978-81-689089-0-1

Explainable AI," Future Internet, vol. 17, no. 513, Nov. 2025.
[8] S. Shah, M. Shukla, N. H. Dholakia, and H. Gupta, "Predicting cardiovascular risk with hybrid ensemble learning and explainable AI," Sci. Rep., vol. 15, no. 17927, 2025.
[9] J. Prashanthan and A. Prashanthan, "Predicting the future risk of developing type 2 diabetes in women with a history of gestational diabetes mellitus using machine learning and explainable artificial intelligence," Primary Care Diabetes, vol. 19, pp. 658–666, 2025.
[10] I. Tasin, T. U. Nabil, S. Islam, and R. Khan, "Diabetes prediction using machine learning and explainable AI techniques," Healthc. Technol. Lett., vol. 10, pp. 1–10, 2023.
[11] W. Gong et al., "Deep learning for enhanced prediction of diabetic retinopathy: a comparative study on the diabetes complications data set," Front. Med., vol. 12, p. 1591832, Jun. 2025.
[12] S. S. Bontha et al., "Predicting Risk and Complications of Diabetes Through Built-In Artificial Intelligence," Computers, vol. 14, no. 277, Jul. 2025.
[13] J. Liu, F. Ssewamala, R. An, and M. Ji, "Use of Automated Machine Learning to Detect Undiagnosed Diabetes in US Adults: Development and Validation Study," JMIR AI, vol. 4, p. e68260, 2025.
[14] R. Hasan, V. Dattana, S. Mahmood, and S. Hussain, "Towards Transparent Diabetes Prediction: Combining AutoML and Explainable AI for Improved Clinical Insights," Information, vol. 16, no. 7, Dec. 2024.
[15] P. Netayawijit, W. Chansanam, and K. Sorn-In, "Interpretable Machine Learning Framework for Diabetes Prediction: Integrating SMOTE Balancing with SHAP Explainability for Clinical Decision Support," Healthcare, vol. 13, no. 2588, Oct. 2025.
[16] S. S. Prakash and A. C. Subhajini, "An Efficient Prediction and Analysis of Diabetics Based on Hybrid Convolutional Pyramid Squeeze Attention Network with Explainable AI," Int. J. Pattern Recognit. Artif. Intell., vol. 39, no. 12, p. 2557012, 2025.
[17] M. J. Noh and Y. S. Kim, "Diabetes Prediction Through Linkage of Causal Discovery and Inference Model with Machine Learning Models," Biomedicines, vol. 13, no. 124, Jan. 2025.
[18] R. Esmaeilyfard and M. Bayati, "Enhancing AI-driven forecasting of diabetes burden: a comparative analysis of deep learning and statistical models," Sci. Rep., vol. 15, no. 29137, 2025.
[19] M. F. Kabir et al., "Proximity-Constrained Counterfactuals for Explainable Diabetes Risk Assessment," Appl. Comput. Intell. Soft Comput., vol. 2025, p. 3424976, 2025.
[20] P. K. Mohanty et al., "IMPACT: an interactive multi-disease prevention and counterfactual treatment system using explainable AI and a multimodal LLM," PeerJ Comput. Sci., vol. 11, p. e2839, Apr. 2025.
[21] E. Maimaitijiang, M. Aihaiti, and Y. Mamatjan, "An Explainable AI Framework for Online Diabetes Risk Prediction with a Personalized Chatbot Assistant," Electronics, vol. 14, no. 3738, Sep. 2025.
[22] R. A. Fraser et al., "Integration of artificial intelligence and wearable technology in the management of diabetes and prediabetes," npj Digit. Med., vol. 8, p. 687, 2025.
[23] M. Aljaafari et al., "Integrating Innovation in Healthcare: The Evolution of 'CURA's' AI-Driven Virtual Wards for Enhanced Diabetes and Kidney Disease Monitoring," IEEE Access, vol. 12, pp. 126389–126414, 2024.
[24] A. Naseem et al., "Novel Internet of Things based approach toward diabetes prediction using deep learning models," Front. Public Health, vol. 10, p. 914106, Aug. 2022.
[25] H. Farman et al., "IoT-Aware Real-Time Healthcare Diagnostic Framework for Diabetes Using Wearable Sensors Through Deep Reinforcement Learning," IEEE Internet Things J., vol. 12, no. 11, pp. 18183–18194, Jun. 2025.
[26] M. Gragnaniello et al., "Edge-AI Enabled Wearable Device for NonInvasive Type 1 Diabetes Detection Using ECG Signals," Bioengineering, vol. 12, no. 4, Dec. 2024.
[27] T. Zhu et al., "Population-Specific Glucose Prediction in Diabetes Care With Transformer-Based Deep Learning on the Edge," IEEE Trans. Biomed. Circuits Syst., vol. 18, no. 2, pp. 236–246, Apr. 2024.
[28] Y. Lu et al., "A pretrained transformer model for decoding individual glucose dynamics from continuous glucose monitoring data," Natl. Sci. Rev., vol. 12, p. nwaf039, Feb. 2025.
[29] F. Dubosson et al., "The open D1NAMO dataset: A multi-modal dataset for research on non-invasive type 1 diabetes management," Inform. Med. Unlocked, vol. 13, pp. 92–100, 2018.
[30] G. Aleppo et al., "REPLACE-BG: a randomized trial comparing continuous glucose monitoring with and without routine blood glucose monitoring in adults with well-controlled type 1 diabetes," Diabetes Care, vol. 40, pp. 538–545, 2017.

Cite this article as:

Tanya, Sudhakar and et. al., "A Study on Machine Learning and Explainable AI in Diabetes Prediction and Management", Proceedings of 13th international conference on Microelectronics, Circuits and Systems, Micro2026 (Vol-II)

Displayed as online on 24th July 2026.

Link: http://actsoft.org/science/micro2026-pro/411-micro2026.pdf

ISBN: 978-81-689089-0-1