A Hybrid ANN–XGBoost Framework for Early Diabetes Risk Prediction Under Class Imbalance: External Clinical Validation in an African Healthcare Setting.

Abstract :

Diabetes is one of the leading chronic diseases worldwide and remains a major public health challenge, particularly in low- and middle-income countries where limited diagnostic resources often delay early detection and treatment. Although machine learning has shown considerable promise for diabetes prediction, most existing studies rely on publicly available datasets, compare only a limited number of algorithms, and rarely validate their models using real-world African clinical data. This study proposes a Hybrid ANN–XGBoost framework for early diabetes risk prediction under class-imbalanced conditions. The proposed architecture combines the nonlinear feature representation capability of Artificial Neural Networks (ANNs) with the gradient boosting classification mechanism of XGBoost. The framework was evaluated on a large-scale dataset containing 100,000 patient records and compared with eight baseline models, including Logistic Regression, Support Vector Machines, Random Forest, Gradient Boosting,

Key words:-
Diabetes prediction; Hybrid ANN-XGBoost; Machine learning; External clinical validation; Class-imbalanced learning; African healthcare.

LightGBM, CatBoost, standalone ANN, and standalone XGBoost.External clinical validation was subsequently performed using 2,182 patient records collected at the Endocrinology and Metabolic Diseases Center of CNHU-HKM in Cotonou, Benin. Despite the substantial distribution shift between the development and clinical datasets, the proposed framework achieved an accuracy of **95.37%**, a precision of **99.30%**, a recall of **95.46%**, an F2-score of **96.21%**, and an AUC of **98.23%**, demonstrating excellent predictive performance and strong generalization under real-world clinical conditions.

The framework was further compared with conventional diagnostic criteria based on HbA1c and fasting blood glucose, highlighting its ability to exploit multiple clinical variables for complementary risk assessment. Permutation Feature Importance identified HbA1c and blood glucose level as the two most influential predictors, supporting both the interpretability and clinical relevance of the proposed framework.

Overall, the proposed Hybrid ANN-XGBoost framework provises a robust and clinically relevant solution for early diabetes risk prediction and represents a promising decision-support tool for ressource-constrained African healthcare settings.

1. Introduction: -

Diabetes is one of the most prevalent chronic diseases worldwide and remains a major public health challenge, particularly in low- and middle-income countries. According to the World Health Organization (WHO), diabetes is characterized by chronic hyperglycemia resulting from defects in insulin secretion, insulin action, or both, and is associated with severe long-term complications affecting the cardiovascular,

renal, and nervous systems [1]. In many African countries, limited access to diagnostic facilities, specialized healthcare professionals, and routine screening programs frequently leads to delayed diagnosis, increasing morbidity, mortality, and healthcare costs.

Early identification of individuals at risk is therefore essential for improving patient outcomes and reducing diabetes-related complications. In recent years, machine learning (ML) has emerged as a promising approach for diabetes prediction by exploiting demographic, behavioral, and clinical variables. Conventional supervised learning algorithms, including Logistic Regression, Support Vector Machines (SVM), Random Forest, and Artificial Neural Networks (ANN), have demonstrated encouraging predictive performance for diabetes classification [3,5,6]. More recently, gradient boosting algorithms, particularly XGBoost, LightGBM, and CatBoost, have achieved excellent performance on structured healthcare datasets owing to their ability to model complex nonlinear relationships while maintaining strong robustness against heterogeneous clinical data [7,10].

Despite these advances, several important limitations remain. First, most published studies rely on publicly available datasets, particularly the Pima Indians Diabetes Dataset, which contains relatively few observations and does not adequately represent the demographic, behavioral, and clinical diversity of African populations [4,11]. Consequently, the external validity and real-world applicability of many proposed models remain insufficiently investigated.

Second, comparative evaluations often involve only a limited number of machine learning algorithms, making it difficult to establish a comprehensive assessment across conventional statistical methods, ensemble learning techniques, and recent boosting algorithms [12,13]. Moreover, many studies primarily emphasize predictive accuracy while paying comparatively little attention to methodological innovation, external clinical validation, or the robustness of the proposed models under real-world clinical conditions [9,14,17]. As emphasized by Kelly et al. [18], high predictive performance on benchmark datasets alone does not guarantee successful implementation in routine clinical practice. These limitations considerably reduce the practical value of artificial intelligence systems intended for clinical decision support, particularly in resource-constrained healthcare environments [2,18].

Another important challenge concerns the distribution shift between development datasets and real clinical populations. Models trained on population-based datasets frequently experience performance degradation when applied to specialized healthcare centers, where disease prevalence, patient characteristics, and clinical practices differ substantially. Although this issue directly affects model generalization and clinical reliability, it has received relatively limited attention in the diabetes prediction literature [19,20].

To address these limitations, this study proposes a **Hybrid ANN–XGBoost framework** for early diabetes risk prediction under class-imbalanced conditions. The proposed architecture exploits the complementary strengths of Artificial Neural Networks and XGBoost. The ANN first learns compact nonlinear latent representations from routinely collected clinical variables, which are then concatenated with the original features before classification by XGBoost. This design aims to enrich the feature space while preserving the original clinical information, allowing XGBoost to construct more discriminative decision boundaries than when operating solely on the raw clinical variables. By combining representation learning with gradient boosting classification, the proposed framework seeks to improve diabetes detection while reducing false-negative predictions, a critical objective in clinical screening.

Unlike previous studies, the proposed framework is evaluated through a comprehensive comparison with multiple baseline models, including Logistic Regression, Support Vector Machines, Artificial Neural Networks, Random Forest, Gradient Boosting, LightGBM, CatBoost, and standalone XGBoost. The

evaluation follows a two-stage validation strategy. First, experiments are conducted on a large-scale publicly available dataset containing 100,000 patient records to benchmark predictive performance under highly class-imbalanced conditions. Second, external clinical validation is performed using an independent dataset comprising 2,182 patient records collected at the Endocrinology and Metabolic Diseases Center of CNHU-HKM in Cotonou, Benin. This second evaluation enables assessment of the proposed framework under real-world clinical conditions while explicitly investigating its robustness to distribution shift between development and validation datasets. In addition, the proposed framework is compared with conventional clinical diagnostic criteria based on HbA1c and fasting blood glucose thresholds to evaluate its potential complementary value beyond standard rule-based clinical assessment.

The main contributions of this work are summarized as follows:

1. **A Hybrid ANN-XGBoost framework** that combines ANN-based latent feature representation learning with gradient boosting classification to improve diabetes risk prediction under class-imbalanced conditions.
2. **A comprehensive comparative evaluation** involving Logistic Regression, Support Vector Machines, Artificial Neural Networks, Random Forest, Gradient Boosting, LightGBM, CatBoost, standalone XGBoost, and the proposed Hybrid ANN–XGBoost framework.
3. **An analysis of distribution shift** between a population-based dataset and an independent hospital dataset, highlighting its impact on model generalization and clinical applicability.
4. **External clinical validation** using 2,182 real patient records collected at CNHU-HKM, providing evidence of the framework's robustness in an African healthcare setting.
5. **A comparison with conventional clinical diagnostic criteria** based on HbA1c and fasting blood glucose thresholds, demonstrating the complementary value of the proposed framework for early diabetes risk assessment.

The remainder of this paper is organized as follows. Section 2 reviews the related work on machine learning-based diabetes prediction. Section 3 presents the materials and methods, including the datasets, data preprocessing, machine learning models, and the proposed Hybrid ANN–XGBoost framework. Section 4 presents the experimental results, including comparative performance, external clinical validation, distribution shift analysis, comparison with conventional clinical diagnostic criteria, and model interpretability. Section 5 discusses the main findings, their clinical relevance, the limitations of the study, and directions for future research. Finally, Section 6 concludes the paper.

2. Related Work: -

Machine learning has become one of the most widely adopted approaches for diabetes prediction owing to the increasing availability of electronic health records and routinely collected clinical data. Most existing studies formulate diabetes prediction as a supervised binary classification problem using demographic, clinical, and physiological variables to identify individuals at risk as early as possible.

Early research primarily relied on conventional supervised learning algorithms. Dagliati et al. [6] developed predictive models using routinely collected clinical variables, including age, body mass index (BMI), glycated hemoglobin (HbA1c), hypertension, and smoking status. Similarly, Kavakiotis et al. [4] conducted a comprehensive review of machine learning techniques for diabetes prediction and highlighted the effectiveness of Logistic Regression, Support Vector Machines (SVM), K-Nearest Neighbors, and Random Forest [22]. Their analysis showed that nonlinear and ensemble-based methods generally outperform linear classifiers because they better capture the complex relationships among physiological variables.

Subsequent studies confirmed the advantages of ensemble learning for diabetes prediction. Hasan et al. [12] demonstrated that combining multiple classifiers improves predictive robustness, while Ooi Ting Kee et al. [9] reported that ensemble methods and neural networks consistently outperform conventional machine learning algorithms. Likewise, Mujumdar et al. [14] and Soni and Varma [15] identified Random Forest and Gradient Boosting among the most effective classifiers for diabetes prediction.

More recently, gradient boosting algorithms, particularly XGBoost, LightGBM, and CatBoost, have become leading approaches for structured healthcare data because of their ability to efficiently model nonlinear feature interactions while incorporating regularization mechanisms that reduce overfitting and improve generalization [7,10]. Among these algorithms, XGBoost has consistently demonstrated excellent predictive performance across a wide range of medical classification problems involving heterogeneous tabular datasets, making it one of the strongest standalone classifiers for diabetes prediction.

In parallel, Artificial Neural Networks (ANNs) have attracted considerable attention because of their ability to automatically learn complex nonlinear feature representations directly from clinical variables. Rather than relying exclusively on the original feature space, ANNs transform the input variables into latent representations capable of capturing hidden dependencies that may not be explicitly observable in the original data. Nevertheless, when used as standalone classifiers, neural networks often require careful hyperparameter tuning, may suffer from reduced interpretability, and can exhibit unstable performance on relatively small or heterogeneous clinical datasets [17].

These observations indicate that ANNs and gradient boosting algorithms possess complementary rather than competing characteristics. ANNs are particularly effective at learning informative nonlinear representations of clinical variables, whereas XGBoost provides robust and highly accurate classification on structured tabular data. Recent studies have therefore explored hybrid learning strategies that combine neural representation learning with boosting-based classifiers to improve predictive robustness and generalization in medical prediction tasks [24,25]. Instead of replacing conventional boosting models, these hybrid approaches enrich the original feature space with latent representations learned by the neural network before performing the final classification, thereby exploiting the strengths of both paradigms.

Although hybrid learning frameworks have shown promising results in several medical prediction tasks, their application to diabetes prediction remains relatively limited. Few studies investigate whether ANN-derived latent representations provide measurable improvements over standalone XGBoost, especially when evaluated under class-imbalanced conditions and validated on independent clinical datasets [24,25].

Several research gaps remain in this literature: limited use of representative African clinical data (most rely on the small, non-representative Pima Indians dataset) [4,13], a narrow set of algorithms compared per study [11,14], limited attention to distribution shift between development and clinical populations [18,20], scarce comparison with established biomarker-based diagnostic criteria such as HbA1c and fasting blood glucose, and few rigorous external validations using real-world African clinical data [2,9,16,18].

Motivated by these limitations, this study addresses them through the Hybrid ANN–XGBoost framework and evaluation protocol detailed in Section 1 and developed in Section 3.

3. Materials and Methods: -

This study proposes a **Hybrid ANN–XGBoost framework** for early diabetes risk prediction that integrates large-scale model benchmarking, nonlinear feature representation learning, comprehensive comparative evaluation, and external clinical validation within a unified methodological workflow. The

proposed methodology is specifically designed to address the challenges of class-imbalanced learning, model generalization, distribution shift, and clinically relevant performance evaluation.

As illustrated in **Figure 1**, the proposed workflow consists of five successive stages: **(1)** data collection, **(2)** data preprocessing and class imbalance handling, **(3)** comparative evaluation of baseline machine learning models, **(4)** development of the proposed Hybrid ANN–XGBoost framework through ANN-based latent feature extraction followed by XGBoost classification, and **(5)** external clinical validation using an independent real-world dataset collected in an African healthcare setting.

The following subsections describe each stage of the proposed methodology in detail, from data acquisition and preprocessing to hybrid model construction, experimental evaluation, and external clinical validation

3.1 Data Sources

Two complementary datasets were used to evaluate both the predictive performance and the generalization capability of the proposed Hybrid ANN-XGBoost framework.

The first dataset consists of 100,000 anonymized patient records obtained from the publicly available Diabetes Prediction Dataset hosted on Kaggle [23]. It includes eight demographic and clinical variables commonly associated with diabetes risk: age, gender, hypertension, heart disease, smoking history, body mass index (BMI), glycated hemoglobin (HbA1c), and blood glucose level. This dataset was used for model development, hyperparameter tuning, and comparative benchmarking of all evaluated machine learning models. The second dataset comprises 2,182 real clinical records collected at the Endocrinology and Metabolic Diseases Center of CNHU-HKM in Cotonou, Benin. These data were not used during model training or hyperparameter tuning but were reserved exclusively for external clinical validation, enabling an independent assessment of the proposed framework under real-world clinical conditions in an African healthcare setting.

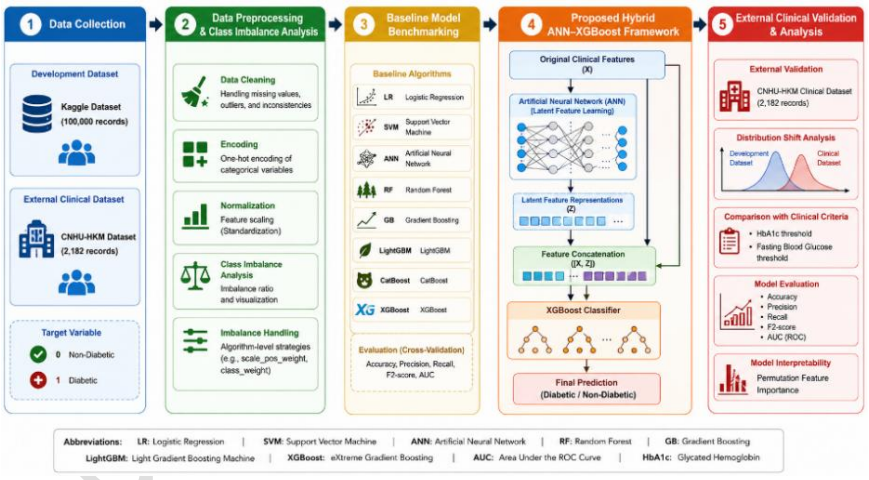

Figure 1. Overall methodological workflow of the proposed Hybrid ANN–XGBoost framework.

3.2 Data Preprocessing and Class Imbalance Handling

Prior to model training, standard preprocessing procedures were applied to ensure data consistency and compatibility with the evaluated machine learning algorithms. Categorical variables were encoded, numerical features were normalized using statistics computed exclusively from the training set to avoid data leakage, and missing values were handled according to established machine learning practices.

A major challenge addressed in this study is the pronounced class imbalance observed in both datasets. As illustrated in **Figure 2**, diabetic cases represent only **8.8%** of the large-scale Kaggle dataset, reflecting a population-based screening scenario, whereas they account for **88.9%** of the CNHU-HKM clinical dataset, which represents a specialized diabetes care center. This substantial difference illustrates the distribution shift between the development and external validation datasets [19].

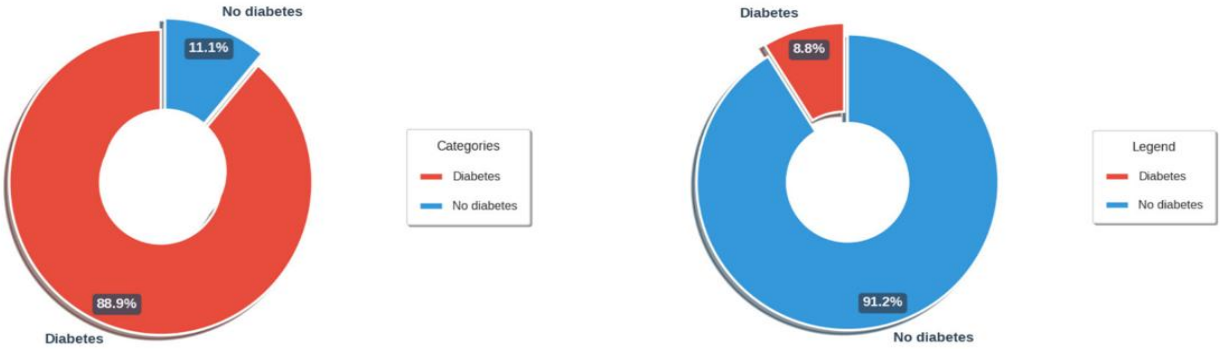

Figure 2. Class imbalance in experimental and clinical datasets

Class imbalance is a well-known challenge in medical machine learning because it may bias classifiers toward the majority class, leading to apparently high overall accuracy while failing to correctly identify minority-class instances [20]. In diabetes screening, false-negative predictions are clinically more critical than false positives because missed cases may delay diagnosis and treatment.

For this reason, no data-level resampling technique (e.g., SMOTE, random oversampling, or undersampling) and no explicit class weighting were applied. All models were trained using the original class distribution, with stratified splitting employed to preserve the relative class proportions across the training, validation, and test sets. Consequently, model performance was primarily evaluated using **recall** and the **F2-score**, while accuracy was reported only as a complementary metric.

3.3 Model Selection

Several supervised machine learning algorithms commonly used in medical prediction were selected.

These were chosen to provide a comprehensive comparison of different learning paradigms. The evaluated models include Logistic Regression (LR), Support Vector Machines (SVM), Artificial Neural Networks (ANN), Random Forest (RF), Gradient Boosting (GB), LightGBM, CatBoost, and XGBoost.

Logistic Regression and SVM were included as representative baseline classifiers because of their widespread use in clinical decision support and medical prediction tasks. Ensemble learning methods, including Random Forest and Gradient Boosting, were selected for their ability to model nonlinear feature interactions and their demonstrated effectiveness in diabetes prediction [6,12,14,22].

More recent boosting algorithms, namely LightGBM, CatBoost, and XGBoost, were also evaluated because of their strong performance on structured tabular healthcare data, their robustness to heterogeneous clinical features, and their built-in regularization mechanisms [7,8,10]. Among these models, XGBoost has consistently demonstrated excellent predictive performance and therefore serves as the principal reference model against which the proposed hybrid framework is evaluated.

Artificial Neural Networks were included because of their ability to learn nonlinear latent feature representations directly from clinical variables. Unlike tree-based boosting methods, which construct decision boundaries directly from the original input variables, ANNs progressively transform the input

space into higher-level representations capable of capturing complex interactions among clinical predictors.

These complementary characteristics motivated the development of the proposed **Hybrid ANN-XGBoost framework**. Rather than replacing XGBoost, the ANN is employed as a feature representation learner that extracts latent representations from the original clinical variables. These learned representations are subsequently concatenated with the original feature set before being supplied to XGBoost for the final classification. This design enables XGBoost to exploit both the original clinical information and the additional nonlinear representations learned by the ANN, thereby potentially improving discrimination between diabetic and non-diabetic patients while preserving the robustness and generalization capability of gradient boosting algorithms [24,25].

The comparative evaluation of the baseline models establishes robust reference performance across conventional statistical models, ensemble learning methods, boosting algorithms, and deep learning approaches. More importantly, the inclusion of both standalone ANN and standalone XGBoost enables an ablation-style evaluation of the proposed framework, allowing the additional contribution of ANN-based latent feature learning to be assessed directly by comparison with each individual component.

3.4 Evaluation Strategy and Performance Metrics

Given the pronounced class imbalance in both datasets, model evaluation was designed to prioritize clinically meaningful performance rather than overall accuracy. In diabetes screening, false-negative predictions are particularly critical because undetected diabetic patients may experience delayed diagnosis and treatment. This consideration is consistent with established recommendations for learning from imbalanced datasets, where recall-oriented metrics provide a more informative assessment than accuracy alone [20].

Accordingly, multiple complementary evaluation metrics were considered, including **accuracy, precision, recall (sensitivity), F1-score, F2-score, confusion matrices,** and the **Area Under the Receiver Operating Characteristic Curve (AUC)**. Among these metrics, **recall** was regarded as the most clinically relevant because it directly measures the proportion of diabetic patients correctly identified by the model.

To further emphasize this clinical objective, the **Fβ-score** was adopted as the primary evaluation metric. Unlike the F1-score, the Fβ-score allows greater emphasis to be placed on recall when $\beta > 1$. In this study, $\beta = 2$ was selected to give greater importance to recall than precision, reflecting the clinical objective of reducing false-negative predictions. The F2-score is defined as:

$$F_\beta = (1 + \beta^2) \frac{\text{Precision} \times \text{Recall}}{(\beta^2 \times \text{Precision}) + \text{Recall}}$$

Confusion matrices were analyzed to provide detailed insight into the types of classification errors, while Receiver Operating Characteristic (ROC) curves and the corresponding Area Under the Curve (AUC) were used to evaluate the discriminative ability of each model across all possible decision thresholds [27].

To ensure a **fair and reproducible comparison**, all models were trained and evaluated using identical training, validation, and test partitions. Hyperparameter optimization was performed exclusively on the training and validation sets, whereas the final performance metrics were computed only on the independent test set. No information from the test data was used during model development or hyperparameter tuning. The same evaluation protocol was applied consistently to all baseline models and to the proposed Hybrid ANN-framework.

Hyperparameters were selected through iterative manual tuning on the training and validation sets, guided primarily by recall and F2-score. The selected configurations were subsequently fixed for final evaluation on the independent test set. Binary classification decisions were made using the standard probability threshold of 0.5, without additional threshold optimization or probability calibration.

3.5 Proposed Hybrid ANN-XGBoost Architecture

Motivated by the complementary strengths of Artificial Neural Networks (ANNs) and gradient boosting algorithms, this study proposes a Hybrid ANN-XGBoost framework for early diabetes risk prediction. While ANNs are effective at learning complex nonlinear feature representations from clinical variables, XGBoost is recognized for its robustness, built-in regularization mechanisms, and excellent performance on structured tabular healthcare data [7,8,10]. Recent advances in medical artificial intelligence have highlighted the potential of hybrid learning architectures that combine representation learning with ensemble-based classifiers to improve predictive performance and generalization [24,25]. Building on these complementary characteristics, the proposed framework integrates ANN-based feature representation learning with XGBoost classification within a unified predictive architecture.

The proposed framework consists of two successive stages. First, an Artificial Neural Network receives the eight clinical variables considered in this study: age, gender, hypertension, heart disease, smoking history, body mass index (BMI), glycated hemoglobin (HbA1c), and blood glucose level.

The ANN comprises an input layer followed by two fully connected hidden layers containing **100** and **50** neurons, respectively. This relatively shallow architecture was deliberately adopted to provide sufficient representation learning capacity while limiting model complexity and reducing the risk of overfitting on structured clinical data. Rectified Linear Unit (ReLU) activation functions were selected because of their computational efficiency and their ability to alleviate the vanishing-gradient problem during training. To further improve generalization, a dropout layer (rate = 0.2) was inserted after each hidden layer.

The network was trained using the Adam optimizer with a learning rate of **0.001**, a batch size of **32**, and binary cross-entropy as the loss function. Training was performed for a maximum of **500 epochs**, while early stopping with a patience of **20 epochs** was employed to automatically terminate the optimization process when the validation loss no longer improved, thereby reducing overfitting and improving reproducibility.

After training, the activations of the final hidden layer are extracted as latent feature representations of the original clinical variables. Rather than relying exclusively on the original feature space, these learned representations capture complex nonlinear interactions among predictors that may not be explicitly modeled by conventional machine learning algorithms.

The extracted latent representations are subsequently concatenated with the original clinical variables, generating an enriched feature space that serves as the input to the XGBoost classifier. This enriched representation enables XGBoost to simultaneously exploit the original clinical information and the additional nonlinear representations learned by the ANN, thereby facilitating the construction of more discriminative decision boundaries.

The XGBoost classifier was optimized using the development dataset. The final model employed **500 decision trees**, a **maximum tree depth of 6**, a **learning rate of 0.05**, a **subsample ratio of 0.8**, a **column sampling ratio of 0.8**, and **L2 regularization** to improve generalization and reduce overfitting. These hyperparameters were selected through iterative experimental tuning on the training and validation sets, with recall and the F2-score used as the primary selection criteria and were subsequently fixed for the final evaluation.

Table 1 summarizes the principal hyperparameters used for training the proposed Hybrid ANN-XGBoost framework. Reporting these parameters improves the reproducibility of the proposed methodology and facilitates future comparative studies.

Table 1. Hyperparameter configuration of the proposed Hybrid ANN–XGBoost framework.

Component	Hyperparameter	Value
ANN	Hidden layers	2
	Neurons per hidden layer	100, 50
	Activation function	ReLU
	Dropout rate	0.2
	Optimizer	Adam
	Learning rate	0.001
	Batch size	32
	Loss function	Binary Cross-Entropy
	Maximum epochs	500
	Early stopping	Patience = 20
XGBoost	Number of trees	500
	Maximum tree depth	6
	Learning rate	0.05
	Subsample ratio	0.8
	Column sampling ratio	0.8
	Regularization	L2

Unlike a standalone ANN, which simultaneously performs feature learning and classification, the proposed framework uses the ANN exclusively as a nonlinear feature extractor. Likewise, unlike standalone XGBoost, which operates directly on the original clinical variables, the proposed architecture enriches the feature space before classification. This design allows XGBoost to benefit from informative latent representations while preserving its well-established robustness and predictive performance on structured tabular healthcare data.

To quantify the contribution of this design, the proposed framework is experimentally compared with both standalone ANN and standalone XGBoost under the same experimental protocol. This comparison enables the additional value of ANN-based latent feature representation learning to be directly assessed.

From a clinical perspective, the proposed architecture is designed to improve the identification of diabetic patients while minimizing false-negative predictions, which is a key requirement for early diabetes screening. **Figure 3** illustrates the overall Hybrid ANN–XGBoost framework.

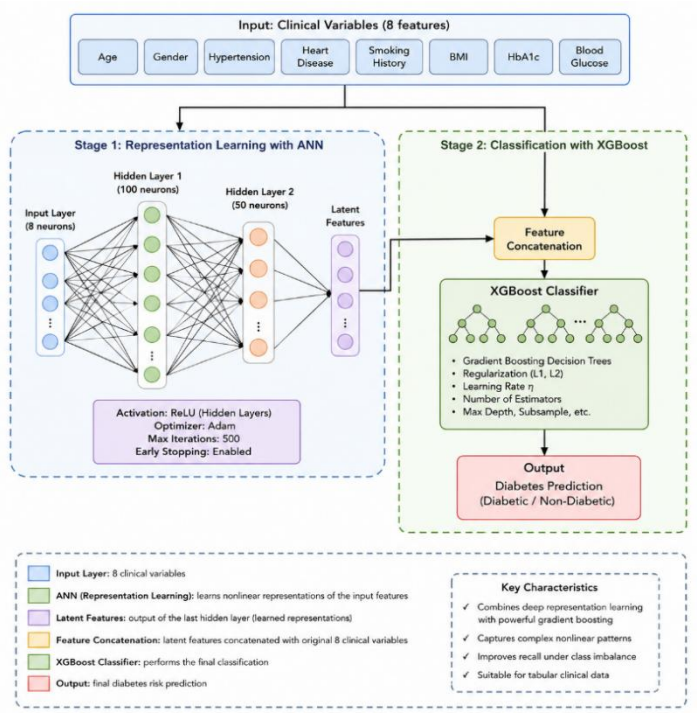

Figure 3. Hybrid ANN–XGBoost Architecture

The ANN extracts latent feature representations that are concatenated with the original clinical variables before XGBoost classification.

4. Results:-

This section presents the experimental evaluation of the proposed **Hybrid ANN–XGBoost framework** and the selected baseline machine learning models. The evaluation is conducted in two successive stages. First, all models are benchmarked on a large-scale public dataset to assess their predictive performance under class-imbalanced conditions. Second, the proposed framework is externally validated using an independent clinical dataset collected at the Endocrinology and Metabolic Diseases Center of CNHU-HKM in Cotonou, Benin.

Beyond predictive performance, the experiments investigate the contribution of the proposed hybrid architecture through comparison with standalone ANN and standalone XGBoost models, evaluate its robustness under distribution shift, and assess its clinical relevance by comparing its predictions with conventional diagnostic criteria based on HbA1c and fasting blood glucose. Finally, model interpretability is examined through Permutation Feature Importance and ROC analysis.

4.1 Model Comparison on the Large-Scale Dataset

The first series of experiments was conducted on the large-scale dataset containing **100,000 patient records** to compare the proposed **Hybrid ANN-XGBoost framework** with several widely used machine learning algorithms under highly class-imbalanced conditions.

Table 2 summarizes the performance of all evaluated models in terms of **precision, recall, F2-score (β = 2),** and **accuracy**. Because the primary objective of diabetes screening is to minimize false-negative predictions, particular emphasis was placed on **recall** and the **F2-score**, which assign greater importance to correctly identifying diabetic patients than to overall classification accuracy.

Table 2. Performance comparison of baseline models and the proposed Hybrid ANN–XGBoost framework.

Model	Precision	Recall	F2-score (β=2)	Accuracy

Logistic Regression	0.8421	0.6013	0.6405	0.9382
SVM	0.8614	0.6158	0.6572	0.9425
ANN	0.8897	0.6425	0.6834	0.9513
Random Forest	0.9476	0.6342	0.6791	0.9567
Gradient Boosting	0.9870	0.6795	0.7240	0.9705
LightGBM	0.9612	0.6728	0.7185	0.9691
CatBoost	0.9587	0.6754	0.7213	0.9698
XGBoost	0.9474	0.6792	0.7200	0.9684
Hybrid ANN–XGBoost	**0.9011**	**0.7086**	**0.7402**	**0.9674**

The proposed **Hybrid ANN–XGBoost framework** achieved the highest **recall (0.7086)** and the highest **F2-score (0.7402)** among all evaluated models. Compared with standalone **XGBoost**, the hybrid approach increased recall from **0.6792** to **0.7086** (**+2.9 percentage points**) and improved the F2-score from **0.7200** to **0.7402** (**+2.0 percentage points**), while maintaining a comparable overall accuracy.

4.2 Clinical Validation on Real-World Data

The second phase of experimentation focused on external clinical validation using real-world data collected at the Endocrinology and Metabolic Diseases Center of CNHU-HKM in Cotonou, Benin. This independent dataset comprises **2,182 patient records** and was used exclusively to evaluate the generalization capability of the proposed Hybrid ANN–XGBoost framework.

Figure 4 presents the confusion matrix obtained on the external clinical dataset.

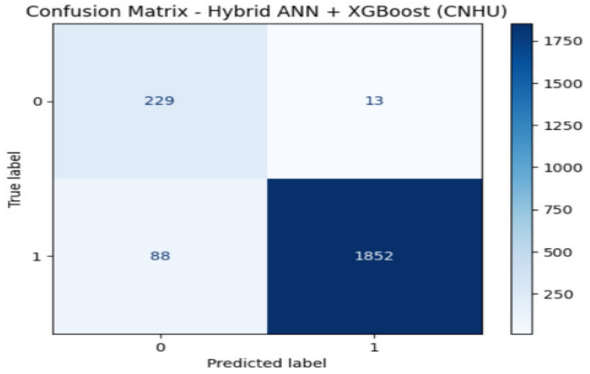

Figure 4. Confusion matrix obtained with the Hybrid ANN–XGBoost model on the CNHU clinical test set

The proposed framework correctly classified **1,852 diabetic patients (true positives)** and **229 non-diabetic patients (true negatives)**, while producing only **88 false negatives** and **13 false positives**.

Overall, the Hybrid ANN–XGBoost framework achieved an **accuracy of 95.37%**, a **precision of 99.30%**, a **recall of 95.46%**, an **F2-score of 96.21%**, and an **AUC of 98.23%**.

4.3 Distribution Shift Analysis

An important contribution of this study is the analysis of the **distribution shift** between the development and external validation datasets. The large-scale Kaggle dataset serves as a public benchmark, where diabetic patients account for only **8.8%** of the observations. In contrast, the CNHU-HKM dataset reflects routine clinical practice in a specialized endocrinology center, with a diabetes prevalence of **88.9%**.

Despite this pronounced distribution shift, the proposed Hybrid ANN-XGBoost framework maintained the excellent predictive performance reported in Section 4.2 during external clinical validation.

4.4 Feature Importance Analysis Using Permutation Importance

To improve model interpretability, **Permutation Feature Importance (PFI)** was performed using the **F2-score** as the evaluation metric, consistent with the recall-oriented objective of this study. PFI estimates the contribution of each predictor by measuring the decrease in model performance after randomly permuting its values, thereby providing a model-agnostic measure of feature importance [26].

Figure 5 presents the permutation feature importance obtained with the proposed Hybrid ANN–XGBoost framework on the external CNHU-HKM clinical dataset.

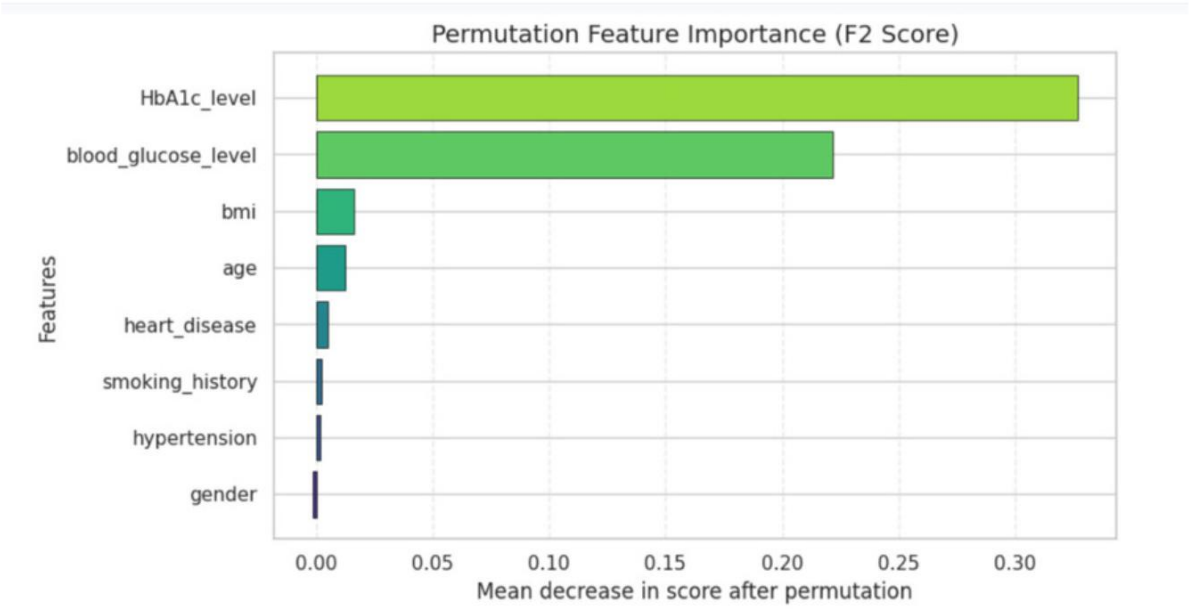

Figure 5. Permutation feature importance (F2-score) obtained with the Hybrid ANN–XGBoost model on the CNHU clinical dataset.

Permutation importance was computed exclusively on the independent clinical test set (**n = 2,182**) to avoid optimistic bias and to ensure that feature importance reflected the model's behavior on previously unseen clinical data.

The results reveal a clear predominance of glycemic biomarkers in the model's decision process. **HbA1c** was identified as the most influential predictor, with an average decrease of approximately **0.32** in the F2-score after permutation. **Blood glucose level** ranked second, producing an average decrease of approximately **0.22**. Together, these two variables contributed most strongly to the predictive performance of the proposed framework.

The remaining variables, including **BMI, age, heart disease, smoking history, hypertension,** and **gender**, showed comparatively smaller contributions. Among them, **BMI** retained a modest influence, whereas the other predictors individually produced only limited decreases in the F2-score once the dominant glycemic biomarkers were incorporated into the model.

4.5 ROC Curve and AUC Analysis

To further assess the discriminative ability of the proposed framework, **Receiver Operating Characteristic (ROC)** analysis was performed on the independent CNHU-HKM clinical dataset (**n = 2,182**). Unlike threshold-dependent metrics such as precision, recall, and the F2-score, the ROC curve evaluates model performance across all possible classification thresholds, thereby providing a threshold-independent assessment of classification performance [27].

The ROC curve obtained for the proposed Hybrid ANN–XGBoost framework is shown in Figure 6. The model achieved an Area Under the Curve (AUC) of 0.9823.

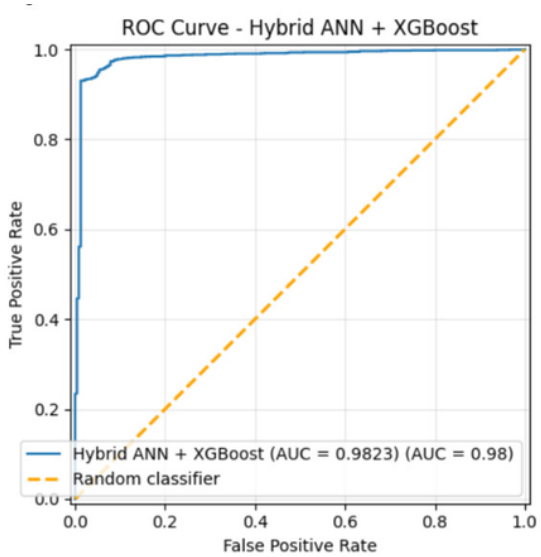

Figure 6. Receiver Operating Characteristic (ROC) curve of the Hybrid ANN–XGBoost model on the CNHU clinical test set.

5. Discussion:-

5.1 Model Performance and Contribution of the Hybrid Architecture

All evaluated models achieved accuracy values exceeding **93%**, confirming that routinely collected clinical variables provide sufficient information for diabetes prediction. However, because of the pronounced class imbalance, **accuracy alone was not sufficiently discriminative** for comparing model performance, highlighting the importance of recall-oriented evaluation metrics.

Linear classifiers, including **Logistic Regression** and **SVM**, produced the lowest recall values, indicating a limited ability to correctly identify diabetic patients. Ensemble learning methods consistently outperformed conventional classifiers, confirming the advantages of nonlinear learning strategies for structured medical data [9,12,14]. Among the standalone boosting algorithms, **Gradient Boosting, LightGBM, CatBoost,** and **XGBoost** achieved comparable performance, with XGBoost providing one of the best overall trade-offs between precision, recall, and F2-score.

The standalone **ANN** achieved a recall of **0.6425** and an F2-score of **0.6834**, outperforming the conventional linear models but remaining below the boosting-based methods. This result indicates that, although ANNs effectively learn nonlinear feature representations, their classification performance on structured tabular clinical data is less competitive when used as standalone classifiers.

These improvements indicate that the latent representations learned by the ANN provide complementary information that enhances the discriminative capability of XGBoost.

The comparison between **standalone ANN**, **standalone XGBoost**, and the proposed **Hybrid ANN-XGBoost framework** provides an ablation-style evaluation of the contribution of ANN-based feature representation learning. The ANN alone is effective at extracting nonlinear relationships but exhibits lower classification performance than boosting methods, while XGBoost alone performs robust classification using only the original clinical variables. Their combination enables XGBoost to exploit both the original feature space and the ANN-derived latent representations, resulting in improved sensitivity for diabetes detection.

It is noteworthy that the hybrid framework does not achieve the highest accuracy among all evaluated models. Instead, its principal advantage lies in improving **recall** and the **F2-score**, which were intentionally selected as the primary evaluation criteria because of their greater clinical relevance in

diabetes screening. From aclinical perspective, reducing false-negative predictions is generally more important than maximizing overall accuracy, since missed diabetic cases may delay diagnosis and treatment.

Overall, these results demonstrate that the proposed **Hybrid ANN-XGBoost framework** provides the most favorable trade-off between sensitivity and predictive performance among the evaluated approaches, thereby justifying its selection for external clinical validation.

5.2 Clinical Validation and Clinical Implications

Unlike the population-based development dataset, the CNHU-HKM dataset reflects routine clinical practice in a specialized endocrinology center and exhibits a markedly different class distribution, with a predominance of diabetic patients. Consequently, this independent evaluation provides a realistic assessment of the robustness and clinical applicability of the proposed framework under real-world conditions.

These results indicate excellent predictive performance and strong generalization to independent clinical data [18,19], despite the substantial difference in class prevalence and clinical setting between the development and external validation datasets.

The exceptionally high precision (**99.30%**) indicates that nearly every patient predicted as diabetic was correctly classified, thereby minimizing unnecessary follow-up investigations. More importantly, the recall (**95.46%**) confirms that most diabetic patients were successfully identified, a key requirement for screening applications where missed diagnoses may delay treatment and increase the risk of long-term complications.

The **F2-score (96.21%)**, which places greater emphasis on recall, further demonstrates that the proposed framework achieves an appropriate balance between sensitivity and precision. In addition, the **AUC of 98.23%** confirms the excellent discriminative ability of the model across a wide range of decision thresholds.

To further assess its clinical relevance, the proposed framework was compared qualitatively with conventional diagnostic criteria based on the **American Diabetes Association (ADA)** thresholds, namely **HbA1c ≥ 6.5%** and **fasting blood glucose ≥ 126 mg/dL**. While these rule-based criteria remain the clinical reference for diabetes diagnosis, they rely primarily on individual glycemic biomarkers considered independently. In contrast, the proposed Hybrid ANN-XGBoost framework simultaneously integrates multiple demographic and clinical variables and models their nonlinear interactions to estimate an individualized risk profile. Rather than replacing established diagnostic criteria, the proposed framework is intended to complement existing clinical practice by supporting earlier identification of patients at elevated risk and assisting clinicians in prioritizing confirmatory diagnostic testing[28].

Overall, these findings indicate that the proposed Hybrid ANN-XGBoost framework maintains excellent predictive performance under real clinical conditions and generalizes effectively to an independent African hospital dataset. The combination of high recall, outstanding precision, excellent discriminative capability, and complementary support to conventional diagnostic assessment highlights its potential as a reliable clinical decision-support tool for early diabetes risk prediction in resource-constrained healthcare environments [31].

5.3 Distribution Shift and Generalizability

Rather than representing a data inconsistency, these differences correspond to two distinct clinical settings: a general population screening environment and a specialized referral center. Such variations in disease prevalence, patient characteristics, and clinical context may substantially affect the generalization capability of machine learning models, particularly those trained under highly imbalanced conditions [19,20].

These findings indicate that the proposed framework achieves strong performance on the external clinical dataset across heterogeneous data distributions, while also demonstrating robustness under substantially different clinical conditions.

One possible explanation for this behavior lies in the complementary design of the proposed hybrid architecture. The ANN learns latent nonlinear representations from the clinical variables, whereas XGBoost exploits these enriched representations to construct robust decision boundaries. This combination may reduce sensitivity to variations in the original feature distribution by providing a richer representation of the underlying clinical information. Although this hypothesis is supported by the observed experimental results, additional multicenter studies involving more diverse patient populations will be necessary to further validate the robustness of the proposed framework under broader distribution shift scenarios.

These findings further emphasize the importance of **external clinical validation** in medical artificial intelligence. As highlighted by Kelly et al., evaluations performed exclusively on public benchmark datasets may overestimate predictive performance and provide limited evidence regarding real-world clinical applicability [18]. The successful validation on an independent African hospital dataset therefore provides additional evidence supporting the robustness and potential clinical utility of the proposed Hybrid ANN-XGBoost framework.

5.4 Interpretation of Feature Importance

This ranking is consistent with established clinical knowledge, as **HbA1c** and **blood glucose** constitute the principal biomarkers currently used for diabetes diagnosis and monitoring. Their dominant importance therefore provides independent support for the comparison with conventional clinical diagnostic criteria presented in Section 5.2. At the same time, the model also incorporates additional demographic and clinical variables, enabling individualized risk assessment that extends beyond fixed diagnostic thresholds.

Overall, the permutation importance analysis indicates that the proposed Hybrid ANN-XGBoost framework relies primarily on clinically meaningful predictors rather than on spurious associations. The strong agreement between statistical feature importance and established medical knowledge enhances the interpretability of the proposed framework and provides additional evidence supporting its clinical relevance.

5.5 ROC Performance and Discriminative Ability

This AUC indicates excellent discrimination between diabetic and non-diabetic patients.

As illustrated in Figure 6, the ROC curve rises rapidly toward the upper-left corner of the ROC space and remains well above the reference diagonal representing the performance of a random classifier. This behavior indicates that the proposed framework maintains a high true positive rate while keeping the false positive rate low across a wide range of decision thresholds.

The **AUC of 0.9823** further supports the robustness of the proposed framework despite the pronounced distribution shift between the development and external validation datasets. When considered together with the **95.46% recall** and the **96.21% F2-score**, these findings indicate that the proposed Hybrid ANN-XGBoost framework combines strong discriminative ability with high sensitivity, two essential characteristics for early diabetes risk prediction.

Overall, the ROC analysis complements the recall-oriented evaluation adopted throughout this study by demonstrating that the proposed framework maintains consistent classification performance across different decision thresholds. Together with the external clinical validation and the feature importance analysis, these results provide additional evidence supporting the robustness, interpretability, and clinical relevance of the proposed Hybrid ANN-XGBoost framework.

5.6 Limitations and Future Work

Despite the encouraging results obtained in this study, several limitations should be acknowledged.

First, although external validation was performed using an independent clinical dataset collected at the Endocrinology and Metabolic Diseases Center of CNHU-HKM in Cotonou, Benin, the evaluation remains

limited to a single healthcare institution. Consequently, additional multicenter studies involving larger and more diverse African populations are required to further assess the robustness and generalizability of the proposed Hybrid ANN–XGBoost framework across different clinical settings.

Second, the study relies on cross-sectional data; incorporating longitudinal patient records could improve early identification of at-risk individuals. Third, further ablation studies involving alternative ANN architectures and latent feature dimensions would better characterize the contribution of the representation-learning stage. Fourth, the added computational complexity of the hybrid framework, while justified by the observed performance gains, motivates future work on lightweight architectures and model compression for resource-constrained settings.

Finally, although Permutation Feature Importance improved the global interpretability of the proposed framework, complementary explainable artificial intelligence (XAI) techniques, such as **SHAP** and **LIME**, could provide local explanations for individual predictions, thereby improving transparency, facilitating clinical interpretation, and increasing clinician confidence in decision-support systems [21,29].

Future research will therefore focus on **(i)** multicenter external validation using larger African clinical datasets, **(ii)** systematic evaluation under additional distribution shift scenarios, **(iii)** optimization of the proposed hybrid architecture through comprehensive ablation studies, **(iv)** integration of advanced explainability techniques, and **(v)** prospective clinical studies to evaluate the impact of the proposed framework on routine diabetes screening and clinical decision-making.

In addition, the present study used the standard 0.5 classification threshold without probability calibration; future work could investigate threshold optimization and calibration strategies to further improve the clinical reliability of predicted probabilities.

6. Conclusion: -

This study proposed a **Hybrid ANN-XGBoost framework** for early diabetes risk prediction under class-imbalanced conditions. By combining the nonlinear representation learning capability of Artificial Neural Networks with the robust classification performance of XGBoost, the proposed framework enriches the original clinical feature space through latent representations, thereby improving diabetes detection while preserving the strengths of gradient boosting for structured clinical data.

A comprehensive experimental comparison involving Logistic Regression, Support Vector Machines, Artificial Neural Networks, Random Forest, Gradient Boosting, LightGBM, CatBoost, and standalone XGBoost demonstrated the effectiveness of the proposed hybrid architecture. On the large-scale benchmark dataset, the proposed framework achieved the highest recall and F2-score among all evaluated models, indicating that the integration of ANN-derived latent representations improves the identification of diabetic patients under class-imbalanced conditions.

The framework was subsequently evaluated using an independent clinical dataset comprising **2,182 patient records** collected at the Endocrinology and Metabolic Diseases Center of **CNHU-HKM** in Cotonou, Benin. Despite the pronounced distribution shift between the development and external validation datasets, the proposed framework maintained strong predictive performance, supporting its robustness and generalization capability under real-world clinical conditions.

Beyond predictive performance, this study makes three main methodological contributions. First, it demonstrates the value of combining ANN-based representation learning with gradient boosting for structured clinical prediction. Second, it highlights the importance of evaluating machine learning models under distribution shift using independent clinical datasets rather than relying exclusively on public

benchmark datasets. Third, it provides external clinical validation using real-world African healthcare data, addressing an important limitation of the current diabetes prediction literature.

Permutation Feature Importance identified **HbA1c** and **blood glucose level** as the two most influential predictors, while comparison with conventional diagnostic criteria based on **HbA1c** and **fasting blood glucose** further supported the clinical relevance of the proposed framework. Together, these findings enhance model interpretability and demonstrate that the proposed framework complements established clinical practice by integrating multiple patient characteristics into individualized diabetes risk prediction.

Overall, the proposed Hybrid ANN-XGBoost framework provides a robust and clinically relevant approach for early diabetes risk prediction. By combining hybrid representation learning, comprehensive benchmarking, external clinical validation, distribution shift analysis, comparison with conventional diagnostic criteria, and model interpretability, this study contributes to the development of more reliable and generalizable machine learning models for clinical decision support, particularly in resource-constrained African healthcare settings.

References

[1] World Health Organization. Diabetes. Geneva, Switzerland: World Health Organization. Available: https://www.who.int/news-room/fact-sheets/detail/diabetes

[2] Djiwa N'tèla O., Houngue P., Sarr C. Blockchain-based patient data collection: A case study in intensive care. Journal of Scientific and Engineering Research, 2025.

[3] Maniruzzaman M., Rahman M.J., Al-Mehedi Hasan M., et al. Comparative approaches of machine learning techniques for predicting diabetes. Healthcare Analytics, vol. 1, 2021.

[4] Kavakiotis I., Tsave O., Salifoglou A., et al. Machine learning and data mining methods in diabetes research. Computational and Structural Biotechnology Journal, vol. 15, pp. 104–116, 2017. https://doi.org/10.1016/j.csbj.2016.12.005

[5] Bhat S.S., Banu M., Ansari G.A., Selvam V. A risk assessment and prediction framework for diabetes mellitus using machine learning algorithms. Healthcare Analytics, vol. 4, Article 100273, 2023. https://doi.org/10.1016/j.health.2023.100273

[6] Dagliati A., Marini S., Sacchi L., et al. Machine learning methods to predict diabetes complications. Journal of Diabetes Science and Technology, vol. 12, no. 2, pp. 295–302, 2018. https://doi.org/10.1177/1932296817706370

[7] Chen T., Guestrin C. XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), pp. 785–794, 2016. https://doi.org/10.1145/2939672.2939785

[8] Lai H., Huang H., Keshavjee K., et al. Predictive models for diabetes mellitus using machine learning techniques. BMC Endocrine Disorders, vol. 19, Article 101, 2019. https://doi.org/10.1186/s12902-019-0436-6

[9] Ooi Ting Kee, Harun H., Mustafa N.A., et al. Cardiovascular complications in a diabetes prediction model using machine learning: A systematic review. Cardiovascular Diabetology, vol. 22, Article 245, 2023. https://doi.org/10.1186/s12933-023-01741-7

[10] Kopitar L., Kocbek P., Cilar L., Sheikh A., Stiglic G. Early detection of type 2 diabetes mellitus using machine learning-based prediction models. Scientific Reports, vol. 10, Article 11981, 2020. https://doi.org/10.1038/s41598-020-68771-z

[11] Febrian M.E., et al. Diabetes prediction using supervised machine learning. International Journal of Advanced Computer Science and Applications, vol. 14, no. 2, 2023.

[12] Hasan M.K., Alam M.A., Das D., Hossain E., Hasan M. Diabetes prediction using ensembling of different machine learning classifiers. IEEE Access, vol. 8, pp. 76516–76529, 2020. https://doi.org/10.1109/ACCESS.2020.2989092

[13] Sisodia D., Sisodia D.S. Prediction of diabetes using classification algorithms. Procedia Computer Science, vol. 132, pp. 1578–1585, 2018. https://doi.org/10.1016/j.procs.2018.05.122

[14] Mujumdar A., Vaidehi V. Diabetes prediction using machine learning algorithms. Procedia Computer Science, vol. 165, pp. 292–299, 2020. https://doi.org/10.1016/j.procs.2020.01.047

[15] Soni M., Varma S. Diabetes prediction using machine learning techniques. International Journal of Engineering Research & Technology, 2020.

[16] Kaur H., Kumari V. Predictive modelling and analytics for diabetes using a machine learning approach. Applied Computing and Informatics, vol. 18, nos. 1–2, pp. 90–100, 2022. https://doi.org/10.1016/j.aci.2018.12.004

[17] Esteva A., Robicquet A., Ramsundar B., et al. A guide to deep learning in healthcare. Nature Medicine, vol. 25, no. 1, pp. 24–29, 2019. https://doi.org/10.1038/s41591-018-0316-z

[18] Kelly C.J., Karthikesalingam A., Suleyman M., Corrado G., King D. Key challenges for delivering clinical impact with artificial intelligence. BMC Medicine, vol. 17, Article 195, 2019. https://doi.org/10.1186/s12916-019-1426-2

[19] Moreno-Torres J.G., Raeder T., Alaiz-Rodríguez R., Chawla N.V., Herrera F. A Unifying View on Dataset Shift in Classification. Pattern Recognition, vol. 45, no. 1, pp. 521–530, 2012. https://doi.org/10.1016/j.patcog.2011.06.019

[20] He H., Garcia E.A. Learning from Imbalanced Data. IEEE Transactions on Knowledge and Data Engineering, vol. 21, no. 9, pp. 1263–1284, 2009. https://doi.org/10.1109/TKDE.2008.239

[21] Lundberg S.M., Lee S.-I. A Unified Approach to Interpreting Model Predictions. Advances in Neural Information Processing Systems (NeurIPS), 2017.

[22] Breiman L. Random Forests. Machine Learning, vol. 45, no. 1, pp. 5–32, 2001. https://doi.org/10.1023/A:1010933404324

[23] Mustafa Tz. Diabetes Prediction Dataset. Kaggle, 2023. Available: https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset

[24] Kaliappan J., Saravana Kumar I.J., Sundaravelan S., Anesh T., Rithik R.R., Singh Y., Vera-Garcia D.V., Himeur Y., Mansoor W., Atalla S., Srinivasan K. Analyzing classification and feature selection strategies for diabetes prediction across diverse diabetes datasets. Frontiers in Artificial Intelligence, vol. 7, Article 1421751, 2024. https://doi.org/10.3389/frai.2024.1421751

[25] Li W., Peng Y., Peng K. Diabetes prediction model based on GA-XGBoost and stacking ensemble algorithm. PLOS ONE, vol. 19, no. 9, e0311222, 2024. https://doi.org/10.1371/journal.pone.0311222

[26] Fisher A., Rudin C., Dominici F. All Models Are Wrong, but Many Are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously. Journal of Machine Learning Research, vol. 20, no. 177, pp. 1–81, 2019.

[27] Fawcett T. An Introduction to ROC Analysis. Pattern Recognition Letters, vol. 27, no. 8, pp. 861–874, 2006. https://doi.org/10.1016/j.patrec.2005.10.010

[28] Topol E. High-performance Medicine: The Convergence of Human and Artificial Intelligence. Nature Medicine, vol. 25, no. 1, pp. 44–56, 2019. https://doi.org/10.1038/s41591-018-0300-7

[29] Ribeiro M.T., Singh S., Guestrin C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), pp. 1135–1144, 2016. https://doi.org/10.1145/2939672.293977