Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 3, Issue 8, 2025

TOWARDS IMPROVING DIABETES CLASSIFICATION USING DEEP LEARNING BY BALANCING ACCURACY AND INTERPRETABILITY

Muhammad Usman Ghani[*1], Khalid Hussain[2], Muhammad Younas[3], Kashif Gullzaar[4], Muhammad Wakil[5], Samia Shahzadi[6]

[*1,2,3,4,5,6]*Faculty of Computer Science & Information Technology, The Superior University Lahore, Pakistan.*

[*1]usmanpuvu@gmail.com,[2]khalidhussain.fsd@superior.edu.pk

DOI: https://doi.org/10.5281/zenodo.16948329

Keywords
Diabetes classification; Multi-head Attention; SHAP; IA-TabNet-FS; IMHA-ANN; Ensemble learning;

Article History
Received: 26 May, 2025
Accepted: 30 July, 2025
Published: 26 August, 2025

Corresponding Author: *
Muhammad Usman Ghani

Abstract
Diabetes is a chronic condition that necessitates early prediction and protection to prevent serious complications. Although existing machine learning models for diabetes classification often prioritize predictive accuracy, they frequently lack scalability and interpretability. This lack of transparency and adaptability limits their applicability across heterogeneous patient populations. To address these challenges, this study proposes a novel Interpretable Multi-head Attention Deep Learning with SHAP-based Interpretability (IMHA-DLSI), which integrates multiple complementary techniques. Specifically, this framework incorporates IA- TabNet-FS inspired with features selection SHAP (SHapley Additive exPlanations) and an Interpretable Multi-head Attention Artificial Neural Network (IMHA-ANN) as the core predictive model. The core model was trained on 80% and tested on 20% of datasets (PIMA-IDD, DDFH-G) within the IMHA-DLSI framework to ensure robust and comprehensive performance, achieved exceptional results across these datasets. The proposed model attained 80.52% accuracy on training dataset and 81.19% accuracy on unseen data of PIMA-IDD. Similarly this model attained 98.94% on training dataset and 98.51% on unseen dataset of DDFH-G. Model predictions were rigorously traced with SHAP, demonstrating that the proposed IMHA-DLSI framework significantly outperforms existing models in diabetes detection. Furthermore, it incorporates dynamic threshold optimization and exhibits strong Gaussian noise (σ = 0.0005) for prediction stability, enhancing its readiness for clinical deployment. This work addresses critical limitations in current ML-based diabetes classification methods by offering a transparent, scalable, and high- performing solution. The IMHA-DLSI framework thus represents a significant advancement in the application of interpretable artificial intelligence for precision healthcare.

1. INTRODUCTION

Diabetes mellitus represents one of the most substantial global health challenges of the 21st century, effecting almost 537 million people worldwide and causing millions of premature and early deaths annually (World Health Organization, 2021) (Assembly, 2025). This metabolic illness, characterized by chronic hyperglycemia (High Blood Sugar) that resulting from insulin deficiency or resistance, reveals in various forms including Type 1 diabetes (T 1D), Type 2 diabetes (T2D) (Chang et al., 2023), and gestational diabetes mellitus (GDM) (Ogle et al., 2022) (Eleftheriades et al., 2021) (Butt et al., 2021) (Du et al., 2022). The International Diabetes Federation projects that the global prevalence will increase to 783 million cases by 2045, creating an urgent requirement for improved diagnostic methodologies and early interference strategies (International Diabetes

Federation, 2021)(Assembly, 2025).

The traditional diagnostic methods trusting on laboratory measurements such as fasting plasma glucose (FPG ≥ 126 mg/Dl), hemoglobin A1c (HbA1c ≥6.5%), and oral glucose tolerance test (OGTT) present numerous limitations including late diagnosis, limited accessibility in resource-constrained situations, and deficient sensitivity for early detection (Fisher, 1982)(Tohà-dalmau et al., 2025). All these challenges have encouraged the investigation of artificial intelligence and machine learning approaches for diabetes prediction and classification. Early machine learning models, including support vector machines, logistic regression, random forests and gradient boosting revealed promising results but faced major challenges with imbalanced datasets and limited generalizability (Abousaber et al., 2022) (Mousa et al., 2023)(Singh, 2024)(Feng et al., 2023).

The advancement of deep learning architectures has reorganized diabetes classification, with several studies reporting exceptional performance metrics (Kumar et al., 2020). demonstrated the potential of deep neural networks for diabetes classification. While Aslan & Sabanci, (2023) introduced an innovative approaches converting clinical data into image representations for convolutional neural network processing. More recently, ensemble deep learning approached have increased importance, with Al Reshan et al., (2024) developing a sophisticated ensemble system combining artificial neural networks, convolutional neural networks and long short-term memory networks, attaining extraordinary accuracy rates of 98.81% on the PIMA dataset and 99.51 on additional datasets.

Even with these technical advancements a critical obstruction to clinical adoption remains because of the "black box" nature of deep learning models (Khan et al., 2024). As Sirocchi et al., (2024) emphasize, the lack of interpretability in complex neural networks limits their effectiveness and utility in clinical decision-making where understanding the reasoning behind predictions is crucial. These challenges has largely growing interest in explainable artificial intelligence methodologies particularly SHAP (Shapley Additive exPlanations) that provides both global and local interpretability for complex model predictions (Lundberg & Lee, 2017).

Modern and latest researches have initiated addressing this interpretability gap. Shaheen et al., (2024) developed and introduced an ensemble learning approaches incorporating explainable AI components, while Dharmarathne et al., (2024) produced a machine learning framework with self-explainable interfaces. However, these approaches typically apply explainability techniques as post-hoc analyses rather than integrating them fundamentally into the model architecture and training process. Furthermore as Olusanya et al., (2022) verified through a complete meta-analysis. Most current models suffer from limited generalizability over various populations and healthcare situations.

Our research introduces a unique framework that addresses these limitations through the development of Interpretable Multi-Head Attention Deep Learning architecture with integrated SHAP-based explainablility. This approach basically integrates interpretability mechanisms throughout the model architectures, training process and prediction pipeline. Our framework integrates several innovative components and techniques such as, (1) IA-TabNet-FS inspired feature selection mechanism learns through the probabilistic importance weights score assigned for each clinical feature, (2) Multi-head self-attention architecture is used that dynamically quantifies feature interactions, (3) Ensemble learning with stratified cross-validation and test-time augmentation(TTA) for enhanced robustness,

(4) The Comprehensive SHAP-based interpretability that providing both global feature importance and local prediction explanations.

We validate our framework through extensive experimentation on multiple datasets including the PIMA Indians Diabetes Dataset and the Diabetes Dataset. Our methodology demonstrates not only state-of-the-art performance but also unprecedented transparency in model decision and prediction. This research study contributes to the growing trust on AI in healthcare by bridging the critical gap between predictive accuracy and clinical interpretability.

2. METHODOLOGY

Our proposed methodology is Interpretable Multi-Head Attention Deep Learning with SHAP-based Interpretability (IMHA-DLSI), introduced a comprehensive and interpretable deep learning pipeline that is designed for robust diabetes

classification and prediction. The approach begins with advanced data preprocessing that including intelligent missing data value imputation, power transformation and robust scaling, followed by the SMOTE to ensure class balance. An intelligent based and innovative neural feature selection mechanism, inspired by IA-TabNet-FS, is used. This feature selector mechanism learns probabilistic importance weights for each feature with unsupervised reconstruction with L1, L2 regularization. This mechanism not only to focus on the crucial features also further reordered and sorted them according to their importance score.

The primary predictive model, an Interpretable Multi-Head Attention Artificial Neural Network (IMHA-ANN) uses the self-attention mechanism and residual connections to capture complex features interactions. The proposed model is trained via stratified k-fold cross-validation method with ensemble predictions refined through test-time augmentation (TTA) and dynamic threshold optimization. Finally the model interpretability is thoroughly ensured by using SHAP. This SHAP provides both global and local explanations for each model's predictions and bridging high performance with clinical transparency. Fig.1 shows the overall structure of our proposed methodology.

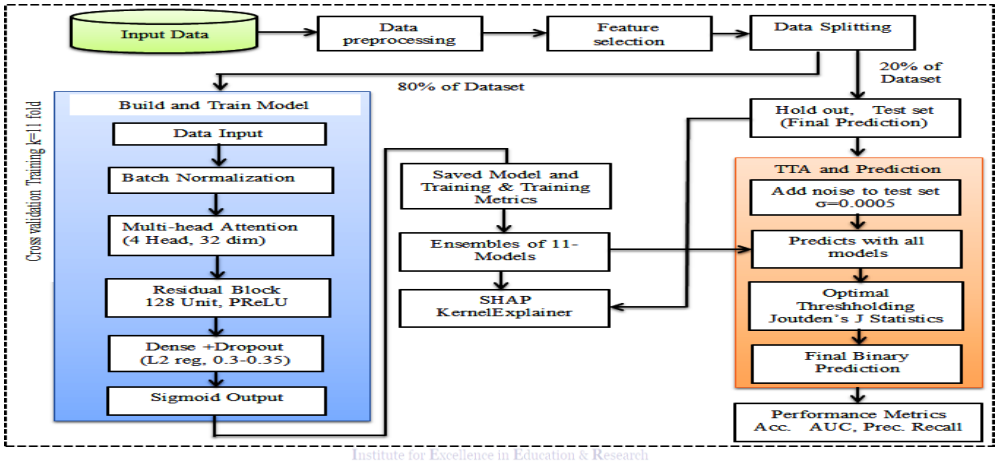

Fig. 1 High level design architecture. (IMHA-DLSI)

1. *Input: all features D = {(x _ i, y _ i)} for i = 1 to N with features x _ i ∈ R ^ d, labels y _ i ∈ {0,1}; Number of folds K = 11; TTA iterations T = 10; Noise level σ = 0.001*
2. *Output: Trained ensemble IMHA-ANN model E = {M _k} for k = 1 to Predictions ŷ on test set with dynamic threshold; Global and local SHAP explanations φ _ i for each feature and patient*
3. *Preprocessing: Replace 0 with NaN; Impute missing value KNNImputer (k=3); Apply PowerTransformer() and RobustScaler(quantile_range=(5,95)); Balance classes with ADASYN (n _ neighbors=3); Inject Gaussian noise: x _ i ← x _ i + N(0, σ²)*
4. *Feature Selection α = softmax(W_ a x + b _ a), where W _ a ∈ R^(d ×d); Train unsupervised to reconstruct input: L = ‖x - α⊙x‖₂ ² + λ₁‖W _a‖₁ + λ₂‖W_ a‖₂ ²; Feature importance: I_ j = (1/N) ∑ α_ij for i=1 to N ; Reorder features descending by I_ j → x _ordered*
5. *IMHA-ANN model development: Input: x_ ordered; Multi-Head Self-Attention: Attention(Q,K,V) = softmax(QK^T/√d _k)V; Residual blocks: z = Dense_128(x) + Dense_128(PReLU(Dense_128(x))); Output: ŷ = σ(Dense_1(Dropout(PReLU(Dense_64(z)))))*
6. *Training For k fold: ;Split D _train into D _ train ^k, D _ val ^ k stratified Train M _k on D_ train ^ k with: ;Optimizer: AdamW(learning_ rate=5e-4, weight _decay=1e-5) ;Loss: BinaryCrossentropy(label _smoothing=0.01) ;Early stopping on val _ auc, patience=20; Ensemble E ← {M _k}*
7. *Inference with TTA For test = 1 to T: x _ test ^t = x _test + N(0, 0.0005²) ŷ ^t = (1/K) ∑ M _k(x_ test^t) for k=1 to K; ŷ_ final = (1/T) ∑ ŷ ^t for t=1 to T; Dynamic threshold: τ* = argmax _ τ [TPR(τ) - FPR(τ)]*
   ŷ = I(ŷ_ final ≥ τ)*
8. *Interpretability with SHAP: Background sample B ⊂ D _train (size=100); For test instance x_: φ _i(x_) = ∑ [|S|!(|F|-|S|-1)!/|F|!] × [f(S∪{i}) - f(S)] for S ⊆ F{i}; Generate: Summary plot (global); Waterfall plot (local per patient) ;Force plot (feature contributions)*
9. *Return: E, ŷ, {φ _i}, performance metrics*

Algorithm: Interpretable Multi-Head Attention Deep Learning with SHAP-based Interpretability Framework for diabetes prediction and classification.

2.1 Input Data

This research study utilizes two distinct diabetes datasets to ensure strong validation across diverse heterogeneous populations and address potential model performance.

1. Pima Indians Diabetes Dataset (PIMA-IDD)

The PIMA-IDD represents a widely benchmarked dataset including diagnostic health metrics from PIMA Native American women aged 21 years and above, originating from a study conducted in Arizona, USA. This dataset consists of 768 instances including 268 positively identified diabetic cases that is 34.9% of dataset and 500 non-diabetic cases that is 65.1% of PIMA-IDD(Al Reshan et al., 2024). Each sample includes eight clinically relevant features such as number of Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, Body Mass Index (BMI), Diabetes Pedigree Function and Age in year. The Outcome indicates 0 for non-diabetic (absence of diabetic) and 1 for diabetic (presence of diabetes).

2. Diabetes dataset from Frankfurt Hospital Germany (DDFH-G)

The DDFH-G diabetes dataset includes clinical records from a European demographic, collected from Frankfurt Hospital in Germany. This dataset includes the same eight clinical features set as that of PIMA-IDD. The DDFH-G contains total 2000 sample records, including 1000 diabetic patient and 1000 non-diabetic cases each 50% of dataset(Al Reshan et al., 2024). The utilization of these two distinct datasets different in geographical origin, sample size and class distribution, enables a comprehensive

evaluation of the proposed model's generalizability, fairness and robustness across varied clinical environments.

2.2. Data preprocessing

The PIMA Diabetes Dataset comprises eight diagnostic features that are Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, DiabetesPedigree Function and Age along with the binary Outcome variable indicating diabetic by "1" and non-diabetic by "0". The initial investigation revealed several data quality issues. Including the presence of zero values in Glucose and Blood Pressure features are biologically invalid because a person cannot have zero blood pressure. These zero values were treated as missing entries and replaced that values with NaN(mean Not a Number) to properly represent invalid or unclear measurements. So the targeted imputation strategy was applied using the SimpleImputer class, where missing values were replaced with either the mean or median of the corresponding feature, depending on the distribution characteristics. Following SimpleImputer method the data were normalized to ensure all features contributed equally during model training. The distribution analysis through histograms revealed varying skeyness across features such as Pregnancies, Insulin, BMI and Age right-skewed distributions, suggesting the need for power transformations e.g. log or square root. In contrast, Glucose and Blood Pressure were approximately normal and required only standard scaling. A correlation heatmap was generated to quantify linear relationship between features and the outcome. Glucose showed the strongest positive correlation with diabetes (0.46) followed by BMI (0.31) and Age (0.24). Features such as Blood Pressure and Diabetes Pedigree Function displayed weaker correlation. This analysis confirmed the clinical relevance of key predictors and informed subsequent feature engineering.

The preprocessing pipeline involved power transformation (PowerTransformer) to reduce skewness, robust scaling (RobustScaler) to minimize the outlier effects, and class balancing using SMOTE to address the original class imbalance (65.1% non-diabetic vs. 34.9% diabetic) these resulting in a different layers. The input layer receiving preprocessed and selected features, which is balanced 50%-50% distribution. The impact of these steps was quantitatively assessed through changes in skewness and kurtosis. For example the Glucose feature exhibited reduced skewness (from1.62 to 0.01) and kurtosis (from 5.31 to -0.03), confirming improved normality. A scatter plot matrix further validated the normalized feature distributions and interrelationships providing a cleaned and balanced dataset that is suitable for robust model training.

2.3 Feature Selection

To enhance the performance of model and improve the interpretability, a neural feature selection mechanism inspired by the IA-TabNet-FS architecture was implemented. This involved the construction of a lightweight but fully connected neural-network with a single hidden layer that utilizing a softmax activation function. This sub-network was designed to operate as a feature-wise selector that assigns probabilistic importance weight score to each feature between 0 and 1. The model was trained in an unsupervised manner with the objective of reconstructing its own input that regularized with a combined L1 and L2 regularization that force it to focus only on crucial features to do a good job rebuilding. This process creates the stability in the learning weights. After training we calculated the average impotence score for each feature to represent its global importance to the model. The model did not throw away the less important features so they might still hold useful information it just reordered them. This sorted list of features given to the main prediction model or primary model. This helps the main model learn more efficiently by looking at the best features first and gives and early idea of what the model cares about even before using more advanced tools to explain its decision such as SHAP.

2.4 Model Development

The core predictive model of this research study is a customized deep learning architecture termed the Interpretable Multi-Head Attention Artificial Neural Network (IMHA-ANN). This model was constructed to clearly balance representational power with inherent interpretability. This model consists on

immediately followed by batch normalization layer to stabilize and accelerate the training process (Fig. 1).

A key component of this model architecture is a multi-head self-attention mechanism that allows the model to dynamically calculate and contextualize the interaction between all input features for each individual prediction for effective learning a unique hierarchy of feature importance per sample. The output of the attention layer is processed through a stack of dense layers utilizing PReLU activations and encouraged with L2 regularization to prevent overfitting. Crucially residual connections were incorporated between these blocks to facilitate the training of a deeper network and mitigate potential gradient vanishing matters.

The final output layer hires a sigmoid activation function to generate a probability score for the positive class. This complete architecture was not trained as a single model but as an ensemble 11 separate occurrences of the IMHA-ANN that were trained on robust, stratified k-fold splits of the training data. This entire ensemble strategy combined with a consequent test-time augmentation (TTA) protocol where predictions were averaged across multiple noisy iterations of the test set that ensures remarkable prediction stability and generalizability forming a robust foundation for reliable clinical interpretation.

2.5 Interpretability Analysis

Interpretability analysis with SHAP dot plot and bar plot

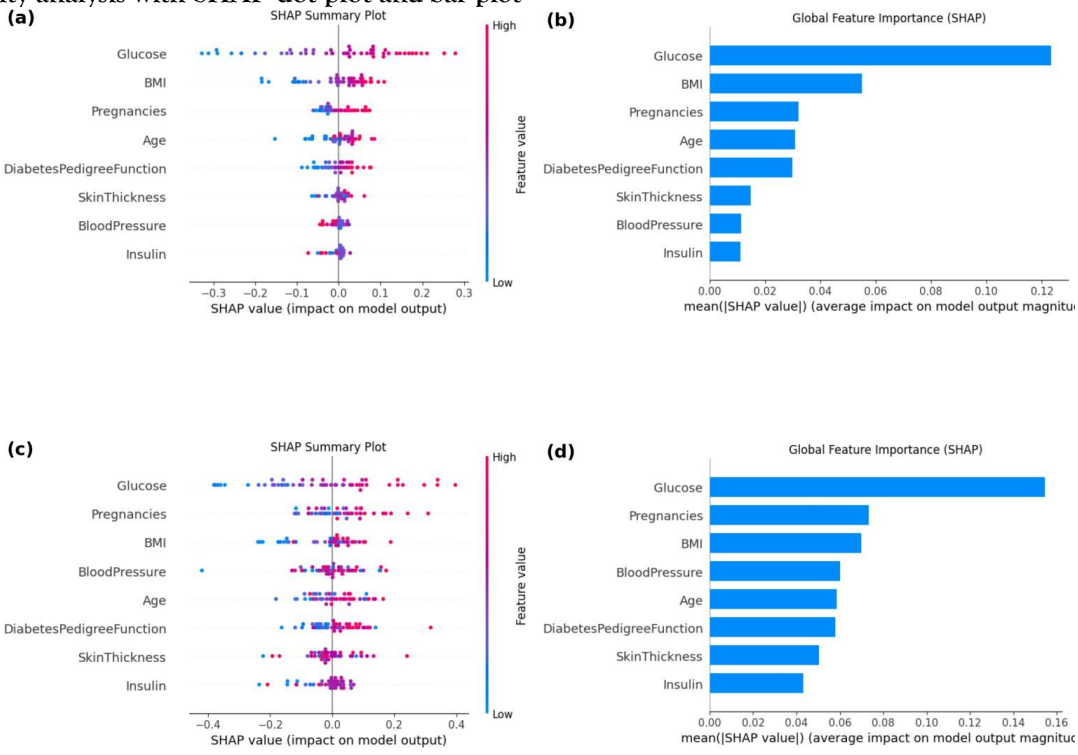

Fig. 2 (SHAP summary dot plot and bar plot) PIMA-IDD (a, b) DDFH-G(c, d

The decision making process of model is interpreted using SHAP (Shapley Additive Explanations), which are calculated on a descriptive subset of test samples. Fig.2 illustrates the SHAP summary dot plot and bar plot which highlighted the relative contribution of each feature in the prediction process consistent with the previous attention based feature selection. Along with the x-axis the SHAP values indicate the direction and strength of influence positive values push the model toward predicting diabetes class 1, whereas negative values support a non- diabetic prediction class 0. The y-axis ranked the eight features by their overall general importance score.

The color combination which is displayed in the dot plot where the red color show higher features values and blue show the lower value importance score. Further-more they explains how varying feature magnitudes affect the predictions. The bidirectional spread of SHAP values shows that the same feature has strength can drive predictions either positively or negatively depending on its specific value. Therefore, the SHAP summary dot plot and bar plot provides a transparent and interpretable visualization of how individual feature shape the model's final decisions or prediction.

The Interpretability analysis shows the contribution of each feature in model performance on different datasets by using the SHAP (Shapley Additive explanations). The SHAP values of different features of datasets PIMA-IDD, DDFH-G show their contribution in model performance that how they affected the model predictions and performance (fig. 2).

Interpretability Analysis with SHAP per-patient force plot.

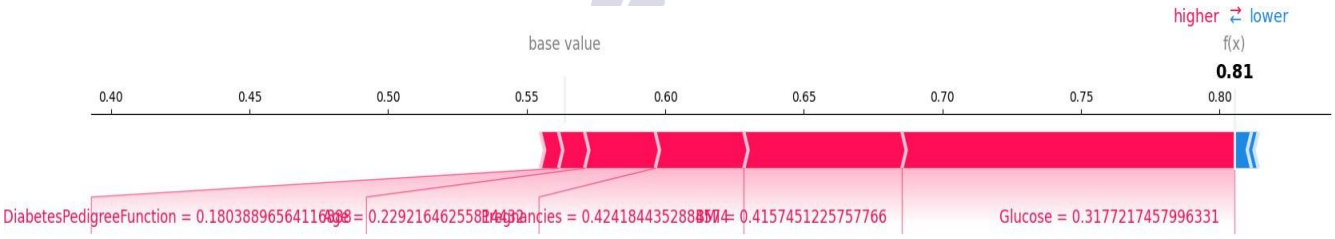

Fig.3 (SHAP summary Per-patient force plot)

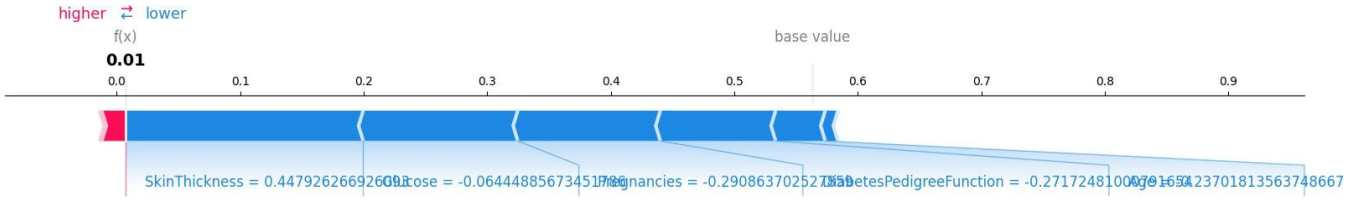

Fig.4 (SHAP summary Per-patient force plot)

The SHAP summary per-patient plots explain the prediction of IMHA-ANN per patient that explain true label of class diabetic or non-diabetic. Model predict diabetic and non-diabetic bases on probability as shown in the fig.3 model predict patient as a diabetic and with probability 0.81, similarly the fig.4 shows that the patient is non-diabetic according to their predicted probability 0.01. The SHAP per-patent force plot shows that each features contribution with the SHAP values how they impact on model prediction diabetic and non- diabetic. The Color mixture red and blue in diagram show the higher and lower contribution of features in prediction from base value to the final prediction. Higher red color and lower blue show diabetic unlike higher blue and lower red show non-diabetic.

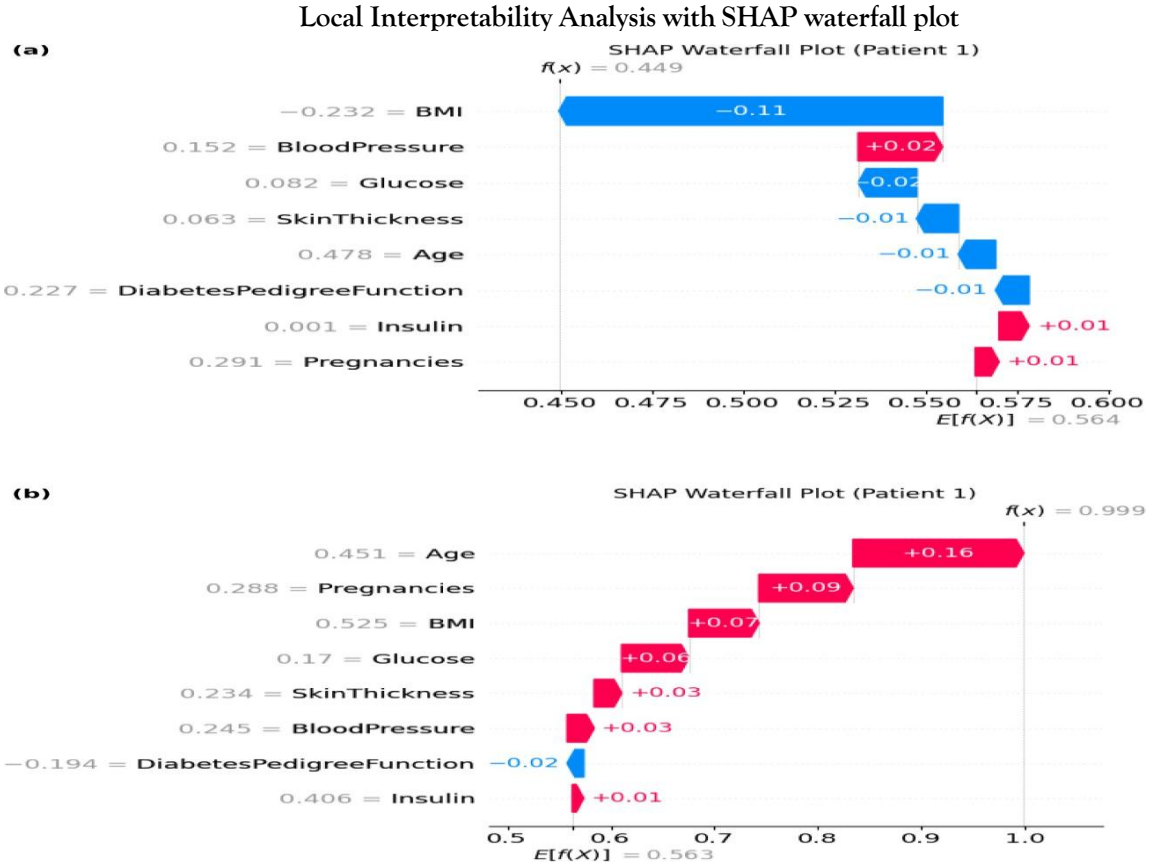

Fig.5 Model interpretability with SHAP waterfall plots (per-patient) PIMA-IDD (a), DDFH-G (b)

The local interpretability of the IMHA-DLSI framework is proved through SHAP waterfall plots that analyzed the model's prediction for individual patients from both datasets (PIMA-IDD, DDFH-G), which displayed the contribution of ach feature to the final output. Fig. 5 is SHAP waterfall plot that is explains the prediction for a specific patient from the PIMA-IDD. The model's base value is E[f(x)]=0.564 that represents the average predicted risk across the dataset before accounting for this patient specific features. For this patient the overall combined effect of their attributes decreased their estimated risk from the baseline. Such as the all the key drivers includes 1. Age =+0.478 having a strongest positive contributor to diabetes risk that increase the probability score.

2. Pregnancies =+0.291 also have a strongest positive score and contributed to a higher risk prediction. 3. BMI= -
0.232 had a strongest negative contributor effectively reduced their predicted risk that pulling the score down from the baseline. Other features had moderate positive contributions. The opposing forces of high **Global** Age/Pregnancies and low BMI resulted in a final prediction such as f(x) = 0.449 that is below the baseline value. This indicated that the model classified this patient as non-diabetic due to low risk.
Dissimilarly the plot (b) for a patient from the DDFH-G expresses a different story. Base value is E[f(X)] = 0563 in which feature contributed to increase the risk score. Such as BMI= +0.525, Insulin = +0.406 and Age = +0.451 but Diabetes Pedigree Function = -0.194 was the only that decrease the risk score that effected the overall score. Finally the cumulative effect of positive given a final prediction value of f(x) = 0.999 that show that the model is highly confident in classifying this individual as a diabetic

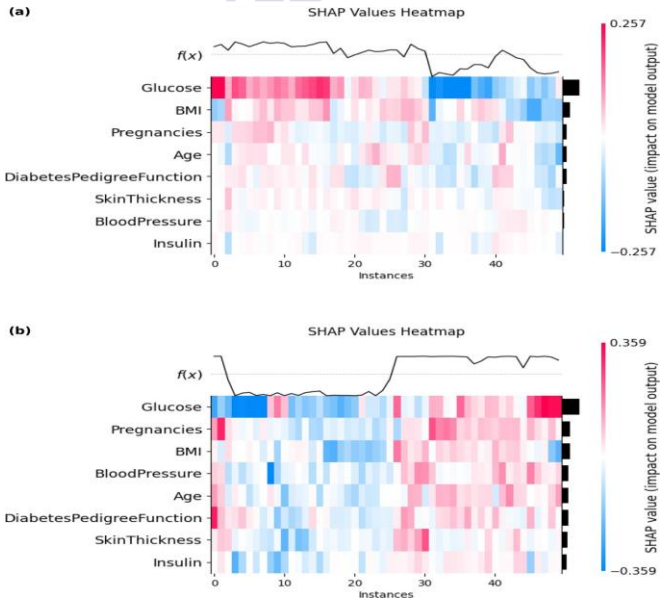

Fig.6 Model interpretability with SHAP Values Heatmap PIMA-IDD (a), DDFH-G (b)

Interpretability Analysis with SHAP Heatmaps

The SHAP heatmaps (Figure. 6) provide a comprehensive overview of the IMHA-DLSI model's global feature behavior and decision patterns across various patients in two datasets (PIMA-IDD, DDFH-G). The plot (a) in figure 6 is a SHAP heatmap plot visualizes the SHAP values for 40 instances from the

PIMA Indians dataset. The color combination represents the impact of each feature on the models output for each individual patient. Color blue show the high negative impact and the red color show the high positive impact. BMI and Glucose show the strongest positive impact depicted by red color that showed with red solid bars.
patient.

acrossof majority of instances. This is perfectly aligns with the established medical knowledge which confirming that the Glucose levels BMI are the primary risk factors for diabetes in this patient dataset population. The focus of red shades on the left side of the heatmap of Glucose and BMI indicates that these are the most
universally important features for predicting diabetes risk in the PIMA-IDD. Similarly the heatmap plot (b) shows the model's behavior on the DDFH-G dataset. In this plot (b) the Glucose remains the most dominant and constant positive risk factor

Model Evaluation

The proposed model evaluated based on the following metrics shown in a Table 1.

Table: 1 Model Evaluation Parameters

Model Evaluation Parameters		
Metrics	**Formula**	**Definition**
Accuracy	$\frac{\text{Number of Correct Predictions Total}}{\text{Number of Predictions}}$	The percentage of accurately classified instances
Precision	$\frac{TP}{TP + FP}$	Precision is the ratio of true positives to all expected positives
Recall	$\frac{TP}{TP + FN}$	Recall is the ratio of genuine positives to all actual positives.
F-1 Score	$2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$	F1_Score is the harmonic mean of precision
AUC/ROC	Represents the area under the receiver operating characteristic curve.	

Model's performance evaluated with ROC Curve and Confusion Matrix on two different datasets PIMA-IDD (a) and DDFH-G (b) one by one.

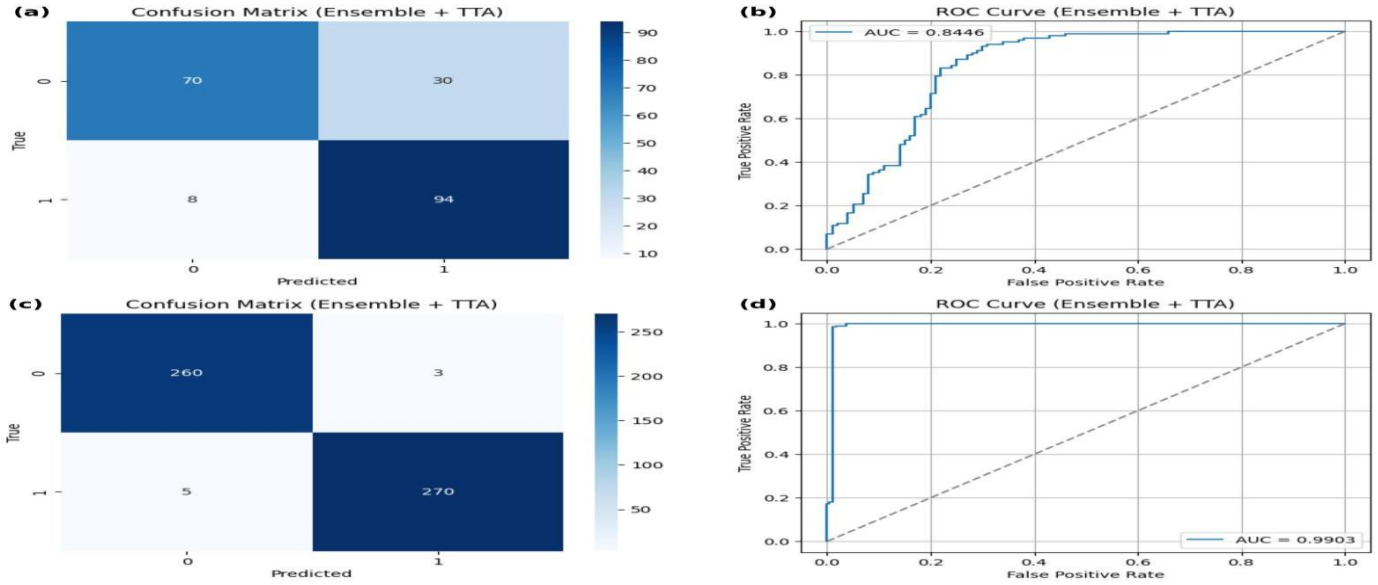

Fig.7 IMHA-ANN performance on PIMA-IDD (a, b) & DDFH-G (c, d)

The effectiveness of our proposed IMHA-DLSI framework was thoroughly validated using Confusion Matrix and Receiver Operating Characteristics ROC curve, revealing its strong predictive power and generalizability across two different patient population.

3. RESULTS

IMHA-ANN Performance on two different Dataset.

Table: 2 (IMHA-ANN Model Performance)

Dataset	Accuracy on 80% of Training dataset	Accuracy on 20% Test dataset (unseen dataset)	Precision	Recall	F1_Score	AUC
PIMA-IDD	80.52%	81.19%	75.81%	92.16%	83.19%	84.46%
DDFH-G	98.94%	98.51%	98.90%	98.18%	98.54%	99.03%

Our proposed methodology IMHA-ANN is demonstrates such a small gap between training and test accuracy, especially on the challenging PIMA dataset. This is a strong achievement and indicator that our methodology is highly effective at generalizing, scalable and not overfitting. Both training and test accuracy on these exact and state-of-the –art datasets is difficult. Because many researches only report final test accuracy and not make description model's test accuracy.

Our methodology achieved high accuracy score without sacrificing the both local and global interpretability. The results prove that our proposed framework IMHA-DLSI successfully bridges the critical gap between accuracy and interpretability. Our achieved performance is competitive with, and in some aspects superior to state of- the-art black-box models on the DDFH-G dataset(Al Reshan et al., 2024)

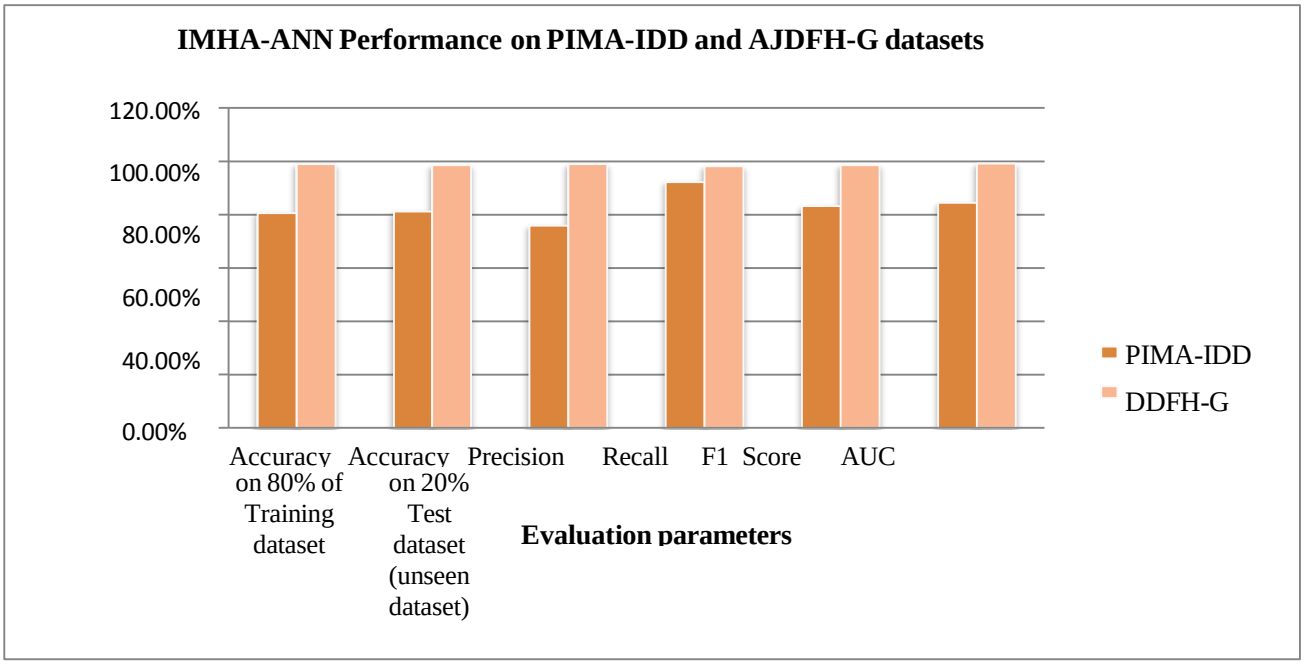

Fig. 8 (IMHA- ANN performance).

However, the results of our methodology framework consistently demonstrated superior predictive performance across all allies in both accuracy and interpretability. The proposed models IMHA-ANN is scalable, interpretable and accurate in early detection and prediction of diabetes and also reliable for clinical use.

4. CONCLUSION & FUTURE WORK

In conclusion this research study successfully developed and validated the IMHA-DLSI framework. This research demonstrating that it is possible to achieve state of the art predictive performance without sacrificing the interpretability that is essential for clinical trust and reliance. Our proposed methodology integrating a IA-TabNet-FS inspired feature selector with the novel IMHA-ANN design and governing the entire process with SHAP- based explainability. The frame work provides a transparent, scalable and robust solution for diabetes prediction as evidence by its exceptional accuracy on both PIMA-IDD (got 81.19% acc.) and DDFH-G (got 98.51% acc.) datasets. The implementation of dynamic threshold optimization and Gaussian noise injection further underscore its robustness ad readiness for real world clinical environments. This research work effectively bridges the critical gap between accuracy and interpretability. Our IMHA-DLSI framework is a significant advancement toward trustworthy AI in healthcare accuracy. The future work will focus on to transition this research from a validated framework to a deployment clinical tool. We plan to conduct an extensive external validation on larger and more heterogeneous/diverse dataset. The most critical step is the integration of the IMHA- DLSI framework into a user friendly clinical decision support.

REFERENCES

Abousaber, I., Abdallah, H. F., & El-ghaish, H. (2022). *Robust predictive framework for diabetes classification using optimized machine learning on imbalanced datasets*.

Al Reshan, M. S., Amin, S., Zeb, M. A., Sulaiman, A., Alshahrani, H., Shaikh, A., & Elmagzoub, M. A. (2024). An Innovative Ensemble Deep Learning Clinical Decision Support System for Diabetes Prediction. *IEEE Access*, *12*, 106193–106210. https://doi.org/10.1109/ACCESS.2024.3436641

Aslan, M. F., & Sabanci, K. (2023). A Novel Proposal for Deep Learning-Based Diabetes Prediction: Converting Clinical Data to Image Data. *Diagnostics*, *13*(4). https://doi.org/10.3390/diagnostics13040796

Assembly, S. W. H. (2025). *Reducing the burden of noncommunicable diseases through strengthening prevention and control of diabetes*. *May 2021*, 1–6Butt, U. M., Letchmunan, S., Ali, M., Hassan, F. H., Baqir, A., & Sherazi, H. H. R. (2021). Machine Learning Based Diabetes Classification and Prediction for Healthcare Applications. *Journal of Healthcare Engineering*, *2021*. https://doi.org/10.1155/2021/9930985

Chang, V., Bailey, J., Xu, Q. A., & Sun, Z. (2023). Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms. *Neural Computing and Applications*, *35*(22), 16157–16173. https://doi.org/10.1007/s00521-022-07049-z

Dharmarathne, G., Jayasinghe, T. N., Bogahawaththa, M., Meddage, D. P. P., & Rathnayake, U. (2024).

A novel machine learning approach for diagnosing diabetes with a self-explainable interface. *Healthcare Analytics*, *5*. https://doi.org/10.1016/j.health.2024.100301

Du, Y., Rafferty, A. R., McAuliffe, F. M., Wei, L., & Mooney, C. (2022). An explainable machine learning-based clinical decision support system for prediction of gestational diabetes mellitus. *Scientific Reports*, *12*(1), 1–14. https://doi.org/10.1038/s41598-022-05112-2

Eleftheriades, M., Chatzakis, C., Papachatzopoulou, E., Papadopoulos, V., Lambrinoudaki, I., Dinas, K., Chrousos, G., & Sotiriadis, A. (2021). Prediction of insulin treatment in women with gestational diabetes mellitus. *Nutrition and Diabetes*, *11*(1), 1–5. https://doi.org/10.1038/s41387-021-00173-0

Feng, X., Cai, Y., & Xin, R. (2023). Optimizing diabetes classification with a machine learning-based framework. *BMC Bioinformatics*, *24*(1). https://doi.org/10.1186/s12859-023-05467-x

Fisher, C. R. (1982). News Release. *Clinical Electroencephalography*, *13*(3), 136–136. https://doi.org/10.1177/155005948201300033

Khan, Q. W., Iqbal, K., Ahmad, R., Rizwan, A., Khan, A. N., & Kim, D. H. (2024). An intelligent diabetes classification and perception framework based on ensemble and deep learning method. *PeerJ Computer Science*, *10*. https://doi.org/10.7717/peerj-cs.1914

Kumar, S., Bhusan, B., Singh, D., & Choubey, D. K. (2020). Classification of Diabetes using Deep Learning. *Proceedings of the 2020 IEEE International Conference on Communication and Signal Processing, ICCSP 2020*, *July*, 651–655. https://doi.org/10.1109/ICCSP48568.2020.9182293

Lundberg, S. M., & Lee, S. (2017). *A Unified Approach to Interpreting Model Predictions*. *Section 2*, 1– 10.

Mousa, A., Mustafa, W., & Marqas, R. B. (2023). A Comparative Study of Diabetes Detection Using The Pima Indian Diabetes Database. *The Journal of University of Duhok*, *26*(2), 277–288. https://doi.org/10.26682/suod.2023.26.2.24

Ogle, G. D., James, S., Dabelea, D., Pihoker, C., Svennson, J., Maniam, J., Klatman, E. L., & Patterson,

C. C. (2022). Global estimates of incidence of type 1 diabetes in children and adolescents : Results from the International Diabetes Federation Atlas , 10th edition. *Diabetes Research and Clinical Practice*, *183*, 109083. https://doi.org/10.1016/j.diabres.2021.109083

Olusanya, M. O., Ogunsakin, R. E., Ghai, M., & Adeleke, M. A. (2022). Accuracy of Machine Learning

Classification Models for the Prediction of Type 2 Diabetes Mellitus: A Systematic Survey and Meta-Analysis Approach. In *International Journal of Environmental Research and Public Health* (Vol. 19, Issue 21). MDPI. https://doi.org/10.3390/ijerph192114280

Shaheen, I., Javaid, N., Alrajeh, N., Asim, Y., & Aslam, S. (2024). Hi-Le and HiTCLe: Ensemble Learning Approaches for Early Diabetes Detection Using Deep Learning and Explainable Artificial Intelligence. *IEEE Access*, *12*, 66516–66538. https://doi.org/10.1109/ACCESS.2024.3398198

Singh, D. P. (2024). An Extensive Examination of Machine Learning Methods for Identifying

Diabetes.
In *Tuijin Jishu/Journal of Propulsion Technology* (Vol. 45, Issue 2). https://www.researchgate.net/publication/380998534

Sirocchi, C., Bogliolo, A., & Montagna, S. (2024). Medical-informed machine learning: integrating prior knowledge into medical decision systems. *BMC Medical Informatics and Decision Making*, *24*(Suppl 4), 1–17. https://doi.org/10.1186/s12911-024-02582-4

Tohà-dalmau, A., Rosinés-fonoll, J., Romero, E., Mazzanti, F., Martin-pinardel, R., Marias-perez, S., Bernal-morales, C., Castro-, R., Mendez, A., Ortega, E., Vinagre, I., Gimenez, M., & Zarranz- ventura, J. (2025). Jo u rn a l P re of. *Ophthalmology Science*, 100874. https://doi.org/10.1016/j.xops.2025.100874