COMPARATIVE STUDY

Abstract- The prevalence of diabetes as a global health crisis necessitates innovative approaches for early detection and prevention. Our study presents a comprehensive examination of multilayer neural network architectures, including Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Multilayer Perceptrons (MLPs), and traditional Regression models, aimed at enhancing the accuracy of diabetes prediction. What sets this research apart is the extensive application of eXplainable Artificial Intelligence (XAI) tools such as SHAP (SHapley Additive exPlanations), LIME (Local Interpretable Model-agnostic Explanations), and ALE (Accumulated Local Effects) to elucidate the contribution of individual features and the inner workings of these machine learning models. This level of detailed feature importance analysis and elucidation of predictive mechanisms is uncommon in the complex realm of current machine learning applications. The imperative for interpretability and explainability is particularly pressing in the clinical context, where understanding model outputs is crucial. Our study offers a twofold contribution: it advances the methodology for diabetes prediction modeling with a strong emphasis on interpretability, and it centers the discussion around AI that clinicians can interpret and trust. The insights derived from this research could significantly impact diagnostic modeling in healthcare and establish a benchmark for future investigations into the harmonization of AI deployment with transparency.

Keywords- diabetes, explainable, XAI, machine learning, model, prediction

I. Introduction

The primary objective of this research was to harness the power of various neural network architectures to predict diabetes outcomes effectively. The uniqueness of this study lies in the amalgamation of different network configurations, each serving a distinct purpose in enhancing prediction accuracy.

A. Network Configurations

•CNN (Convolutional Neural Network): CNNs are renowned for their ability to process data with a grid-like topology, like time-series data or images. For our study, CNN was adept at capturing intricate patterns within the data, which is paramount for significant feature identification. By employing a series of convolutional layers, the network was able to hierarchically extract and amplify meaningful features from our dataset.

•MLP (Multilayer Perceptron): The MLP, a class of feedforward artificial neural network, was utilized because of its intrinsic capability to model non-linear relationships. With its multiple deep layers, it effectively specialized in modelling the complex interplay between features in the dataset.

•RNN (Recurrent Neural Network): The RNN was a crucial component of our methodology, especially when dealing with time-series data. Its unique architecture, designed to retain memory from previous inputs, made it exceptionally competent at capturing temporal dependencies, ensuring that no time-related nuance within our dataset was overlooked

B. Model Performance Evaluation

After rigorous training, we evaluated the model's performance with a battery of metrics, primarily focusing on accuracy and F1-score. While accuracy provided a holistic view of the model's prediction capability, the F1-score zeroed in on the balance between precision and recall, which is crucial in medical predictions to minimize the harmful effects of both false positives and false negatives.

C. Interpretation Mechanisms

Given the complexity of neural networks, there's an inherent challenge in making their predictions understandable to the human mind. To address this, we employed a suite of interpretative methodologies:

•Kernel SHAP is an interpretative method grounded in cooperative game theory. It seeks to quantify the contribution of each feature in a model by examining how the absence or presence of said feature can sway the model's prediction. The hallmark of Kernel SHAP [11] is its ability to offer "energy plots." These visual tools rank features based on their importance, revealing both their positive and negative contributions to a prediction. What makes Kernel SHAP commendable is its consistent and unbiased distribution of feature importance values and its versatility, being applicable across diverse models. However, there are downsides. The approach can be computationally demanding, especially when dealing with a plethora of features. Furthermore, Kernel SHAP operates under the presumption that features are independent, which might not consistently hold true.

•LIME stands as a beacon when one needs to decipher complex models. It operates by approximating such a model with a more comprehensible one, such as linear regression, but only within the vicinity of the prediction in question. The brilliance of LIME lies in its procedure: it modifies the data, procures predictions from the model, and then uses a simpler model to interpret those predictions locally. One of its visualization tools, the "Feature Importance Plots," details the weightage of each feature in the surrogate model, providing clarity on the factors driving specific predictions. Being model-agnostic, LIME [12] can interpret virtually any machine learning model. Yet, its explanations can sometimes be inconsistent, contingent on the surrogate model chosen. Moreover, when modifying data, LIME can occasionally produce unrealistic samples, which might undermine the reliability of the interpretation.

•Diving into the intricate interplay between features and predictions, ALE plots come to the fore. These are crafted to delineate how a specific feature influences a model's prediction, factoring in potential interactions with other features. By offering a granular perspective, ALE plots chart the trajectory of the model's predictions as particular features undergo changes. The plots' axes are straightforward: the x-axis showcases feature values, while the y-axis depicts the average effect on the model's outcome. ALE's strength lies in its adept handling of feature interactions, making its insights particularly reliable. Furthermore, its average-based approach ensures reduced sensitivity to outliers. However, like other methods, ALE [13] comes with its challenges. It presumes that a feature's distribution is independent, which may not always be accurate. Additionally, for models laden with features or intricate interactions, ALE can be computationally taxing.

D. Comparative Study

As a counterpoint to the neural network models, we also incorporated a Random Forest regression [16] model into our analysis. This addition served a twofold purpose: First, it helped us gauge whether the predictive prowess of our complex models genuinely outstripped that of simpler models. Secondly, it underscored the ongoing debate in the machine learning community about the trade-offs between model complexity, interpretability, and performance.

II. Results

A side-by-side evaluation of various models was conducted, using metrics like accuracy and F1-score. With the aid of eXplainable Artificial Intelligence (XAI) techniques, such as Kernel SHAP, LIME, and ALE, we extracted interpretative insights. These visual representations were paramount in discerning performance differentials and subsequently guiding our decision in choosing the most suitable model for practical diabetes prediction. In conclusion, our findings strongly advocate for a synergized approach when deploying machine learning models in healthcare, emphasizing the equilibrium between model intricacy and its interpretability. In the study, a combination of deep learning models and traditional algorithms was leveraged to predict the onset of diabetes in patients. Below is an overview of the models used, their significance, and their respective accuracies on the dataset.

A. Multilayer Perceptron (MLP)

The MLP is a deep artificial neural network characterized by multiple perceptrons. It consists of an input layer, one or more hidden layers, and an output layer. Its fully connected architecture processes the input signal, which aids in decision-making or prediction. The inherent architecture of an MLP [15] makes it suitable for this project as a comprehensive classifier. By taking diverse diabetes-related features as input (e.g., glucose level, insulin level), it provides a probability prediction related to the onset of diabetes. The MLP model achieved an accuracy of 75%.

B. Convolutional Neural Networks (CNN)

Designed primarily for image data, CNNs can also be applied to any structured grid or spatial data. Through convolutional layers, they filter input data for specific features and utilize pooling layers for data size reduction. The use of a CNN [14] in this study indicates the presence of structured grid or potential time-series data. Though it's unconventional for structured data, CNN's automated feature extraction capability offers a unique perspective. The CNN model recorded an accuracy of 70%.

C. Recurrent Neural Networks (RNN)

RNNs are known for their embedded loops, which allow for information persistence. Their architecture is particularly effective with sequential data due to its inherent "memory" that recalls previous sequence steps. The inclusion of RNNs in the study suggests the presence of sequential data, possibly tracking health metrics of patients over time. By "remembering" past data points, RNNs [17] use this sequential information to guide future predictions. In the dataset, the RNN model attained an outstanding accuracy of 80%.

D. Regression Analysis

Regression algorithms concentrate on the relationship between dependent and independent variables. Their goal is to estimate or predict the dependent variable based on the independent variable(s). In the context of this research, regression analysis elucidated how different health metrics correlate with the probability of developing diabetes. The regression model produced an F-score of 30%.

E. XAI Comparisons Between Models

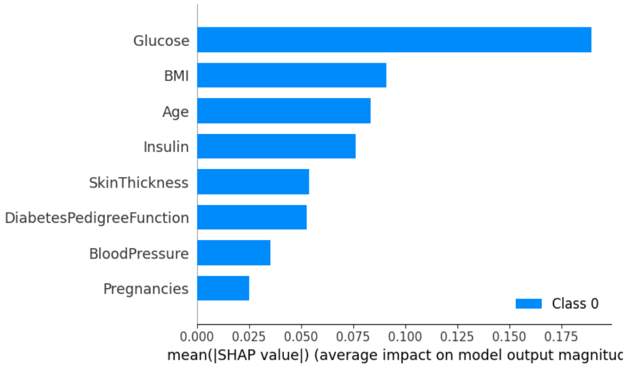

Fig.1: SHAP Summary Plot for CNN Model

Fig.1 presents a SHAP (SHapley Additive exPlanations) value-based bar chart, delineating the relative importance of predictive features within a diabetes classification model. The chart orders variables by descending influence, with 'Glucose' exhibiting the paramount mean SHAP value, denoting its predominant role in model output modulation. Subsequent features, 'BMI', 'Age', and 'Insulin', demonstrate progressively lesser impacts, while 'Pregnancies' contributes minimally. This graphical representation facilitates an analytical comprehension of feature pertinence in the predictive accuracy of the model, crucial for the enhancement of interpretability in machine learning applications within clinical prognostic contexts.

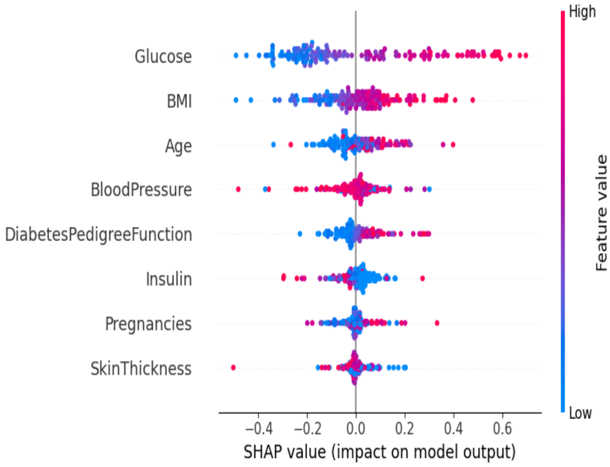

Fig.2: SHAP Summary Plot for MLP Model

Fig.2 displays a SHAP value scatter plot, illustrating the distribution of feature impacts on a predictive diabetes model. Each dot represents a SHAP value for an instance, with colour intensity denoting the feature value magnitude. The 'Glucose' level is discernibly the most significant predictor, with a wide spread of SHAP values, suggesting a strong differential impact on model predictions. The features 'BMI', 'Age', and 'BloodPressure' also show substantial variability in their SHAP values. This visualization provides a nuanced understanding of feature contributions, highlighting the heterogeneity in predictive determinants across the model's operational domain.

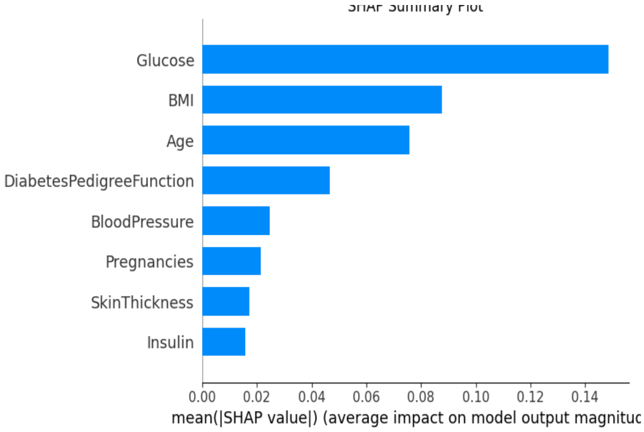

Fig.3: SHAP Summary Plot for Regression Model

Fig.3 illustrates a SHAP summary plot, which quantifies the average marginal contribution of each feature to the predictive model's output. The bar lengths represent the mean absolute SHAP values, conveying the average impact magnitude on the model's output. 'Glucose' is depicted as the most influential feature, with the highest mean SHAP value, indicating its significant role in the model's decision-making process. Following 'Glucose', 'BMI' and 'Age' are also shown as important predictors. The plot underscores the hierarchical significance of clinical features in the determination of diabetes risk, providing insights into the model’s interpretative framework.

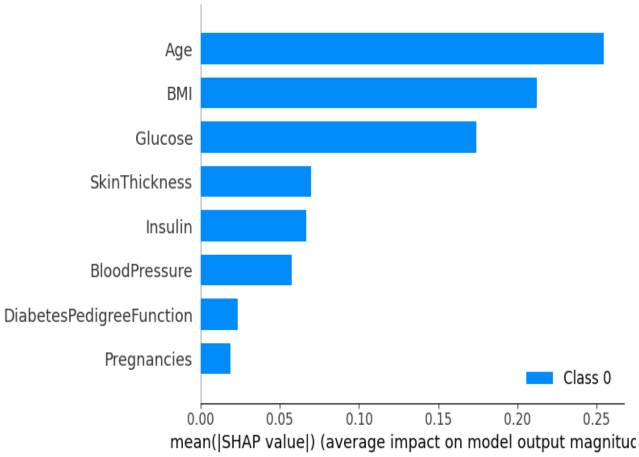

Fig.4: SHAP Summary Plot for RNN Model

Fig. 4 presents a SHAP summary plot, portraying the mean absolute SHAP values for each feature within a diabetes prediction model. The plot ranks the features by their mean SHAP values, signifying their average impact on the model's output for class 0. 'Age' is shown to have the highest mean SHAP value, asserting the strongest influence on the predictive outcome, followed by 'BMI' and 'Glucose'. This implies that older age contributes most significantly to the model's prediction of non-diabetic classification, with body mass index and glucose levels also being key determinants.

This visualization is essential for understanding the model's interpretability in clinical decision support.

In the realm of machine learning, delineating the significance of individual features contributing to a model's prediction remains paramount. SHAP (SHapley Additive exPlanations) summary plots have emerged as a vital tool in this context. Drawing their foundation from cooperative game theory, SHAP values introduce a standardized measure of feature importance by equitably apportioning the prediction outcome amongst the input features. Visually, SHAP summary plots elucidate these values, illustrating both the magnitude and direction of feature contributions. For this study, these plots prove indispensable, furnishing intricate insights into the decision-making mechanisms of the model, thus highlighting pivotal features and their influence on predictions. Such comprehension becomes vital, particularly in critical sectors such as healthcare, where the interpretability of models rivals the importance of their accuracy.

Comparative Examination of SHAP Summary Plots Across Diverse Architectures:

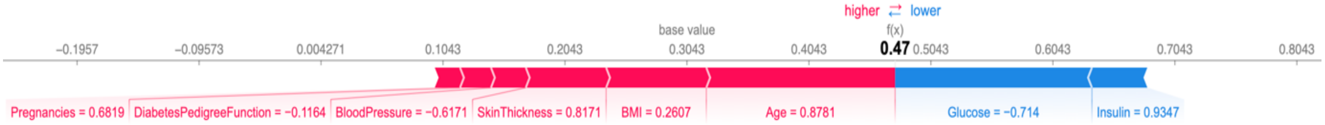

Fig.5: SHAP force plot for CNN

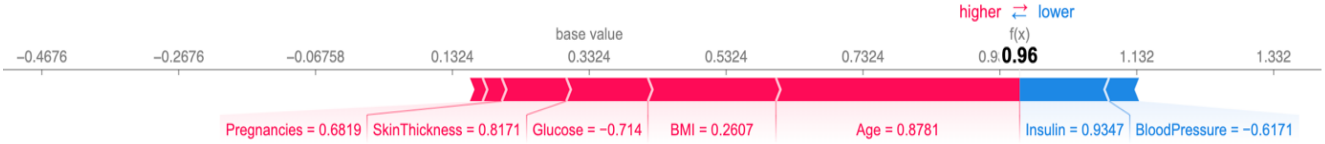

Fig 6: SHAP force plot for regression

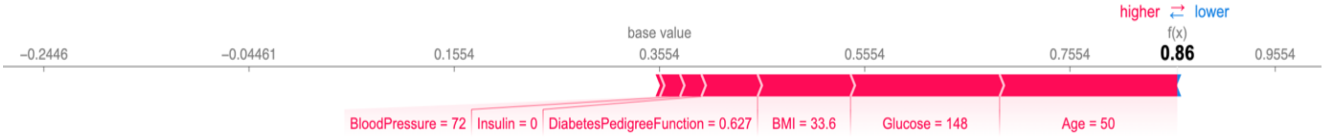

Fig. 7: SHAP force plot for RNN

Fig.1: SHAP Summary Plot for CNN Model: The CNN model underscores a pronounced influence from the 'Glucose' feature, subsequently trailed by 'BMI' and 'Age'. This infers that the convolutional layers inherent in CNNs may discern complex patterns associated with these features, attributed to their inherent capability in processing image or sequential data.

Fig.2: SHAP Summary Plot for MLP Model: Conversely, the MLP model elucidates prominence of features like 'Glucose', 'BMI', and 'Age', albeit with nuanced variations in magnitude. The inherent linear structure of MLPs might elucidate these differences, implying a distinct weighting of these features during prediction formulation.

Fig.3: SHAP Summary Plot for Regression Model: The regression model exhibits a predilection for 'Glucose', succeeded by 'BMI' and 'DiabetesPedigreeFunction'. Notably, while 'Glucose' and 'BMI' maintain consistent significance across models, the emergence of 'DiabetesPedigreeFunction' as an influential feature underscores the model's proclivity towards statistical correlations.

Fig.4: SHAP Summary Plot for RNN Model: RNNs, renowned for their proficiency with sequences, proffer a unique perspective. 'Age' emerges as the predominant feature, succeeded by 'BMI' and 'Glucose'. This posits that temporal or sequential patterns, potentially inherent in the data structuring or input mechanism, might influence feature significance in RNNs.

Across the quartet of architectures—CNN, MLP, Regression, and RNN—a unanimous accord emerges regarding the significance of 'Glucose', 'BMI', and 'Age'. Nevertheless, nuances in the sequence of significance and the extent of influence become evident, illuminating the distinct perceptions and weightings of features across varied model architectures. For the scope of this study, such comparative insights prove pivotal, facilitating informed model selection and optimization, while concurrently underscoring the imperativeness of tools like SHAP in engendering transparent and enlightened decision-making.

Fig.5 illustrates a force plot from a Convolutional Neural Network (CNN) model, which provides a visual decomposition of feature contributions to a specific prediction. The plot quantifies the directional influence of each feature, with red indicating an increase and blue a decrease in the model's output, relative to the base value. This allows for the dissection of the CNN's decision-making process at the instance level, offering insights into the model's interpretative mechanics.

Fig.6 displays a force plot for a regression model, depicting individual feature contributions towards a single prediction. Red shades indicate positive influence, pushing the prediction higher from the base value, while features absent denote neutral impact. This plot serves to elucidate the attributive weight of each variable, like 'Age' and 'Glucose', within the predictive framework.

Fig.7 features a force plot from a recurrent neural network (RNN) model, visualizing the directional influence of input features on a single predictive outcome. The plot delineates how each feature's value shifts the output from the base value towards a higher or lower prediction, with red illustrating positive and blue negative contributions, thereby dissecting the RNN's predictive dynamics at an instance level.

The evolution of machine learning and its ubiquitous application in diverse sectors necessitates a clear understanding of how models arrive at their decisions. SHAP (SHapley Additive exPlanations) force plots serve as an instrumental tool in this endeavor, offering a window into the inner workings of complex models. These plots, a visual representation of SHAP values, highlight the relative significance of features and their contribution to a model's decision-making process. In this research, the team employed SHAP [11] force plots to delve deeper into the decision rationale of various model architectures. Given the crucial nature of predictions in areas like healthcare, discerning the influential features becomes paramount.

Comparative Insights:

Fig.5: The waterfall plot related to the CNN model starts with a base value of 0.3043 and ends with a prediction value of 0.47. Each feature adds or subtracts a certain amount to the prediction. In this chart, 'Glucose' has the most negative impact, decreasing the prediction value, followed by 'Age', 'Insulin', and 'SkinThickness'. Conversely, 'BMI' increases the prediction value the most, followed by 'DiabetesPedigreeFunction' and 'BloodPressure'.

Fig.6: For the Random Forest regression, the base value is 0.3554, leading to a final prediction of 0.86. Here, 'Glucose' with a value of 148 has a significant positive contribution, followed by 'BMI' and 'Age'. 'BloodPressure', 'Insulin', and 'DiabetesPedigreeFunction' contribute negatively to the prediction, pulling the value down.

Fig.7: The RNN model starts with a base value of 0.3324 and culminates in a prediction of 0.96. This time, 'Insulin' and 'Glucose' are the primary positive contributors. In contrast, 'BloodPressure' and 'SkinThickness' exert the most substantial negative impact on the prediction.

Comparative Analysis:

Feature Importance Across Models: 'Glucose': Consistently crucial across all three models but affects differently. In the CNN, it reduces the prediction value, but in Regression and RNN, it significantly increases it. 'BMI': It's a significant positive contributor in both the CNN and Regression models. Its absence as a top influencer in the RNN model is noteworthy. 'Age': Has negative influence in CNN and RNN, while it positively contributes in the Regression model. 'Insulin': While it has a negative impact in the CNN and Regression models, it stands out as the top positive contributor in the RNN.

Model Consistency: While all three models predict a higher than base value likelihood, their interpretations of features vary. For instance, 'Glucose' contributes negatively in CNN but positively in Regression and RNN [17]. Such discrepancies emphasize that different algorithms can perceive the significance of features differently, even when applied to similar data.

Visualization Insights: Force plots offer a step-by-step breakdown of how each feature modifies the base value to arrive at the final prediction. Features pushing the value right are positive contributors, and those pushing left are negative ones. This visualization aids in discerning the magnitude and direction of each feature's influence.

Conclusion: Interpreting model predictions, especially when multiple models are in play, requires careful analysis. The SHAP values, represented by these waterfall plots, provide granular insights into how each feature influences the outcome. These visualizations highlight that while some features consistently impact predictions across models ('Glucose'), others vary in their influence ('Insulin', 'BMI'). Such discrepancies reinforce the importance of understanding and trusting the model in use. The choice of model can significantly affect feature interpretations, emphasizing the need for a contextually relevant model selection in real-world scenarios.

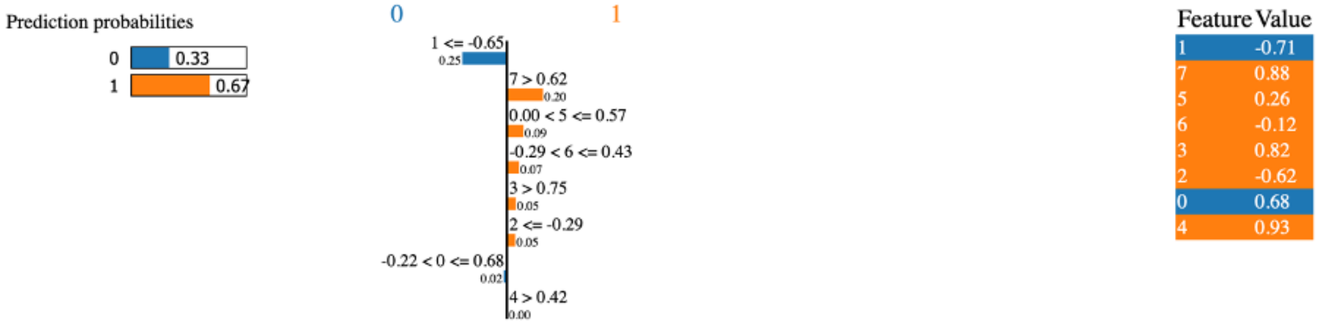

Fig.8: LIME for CNN

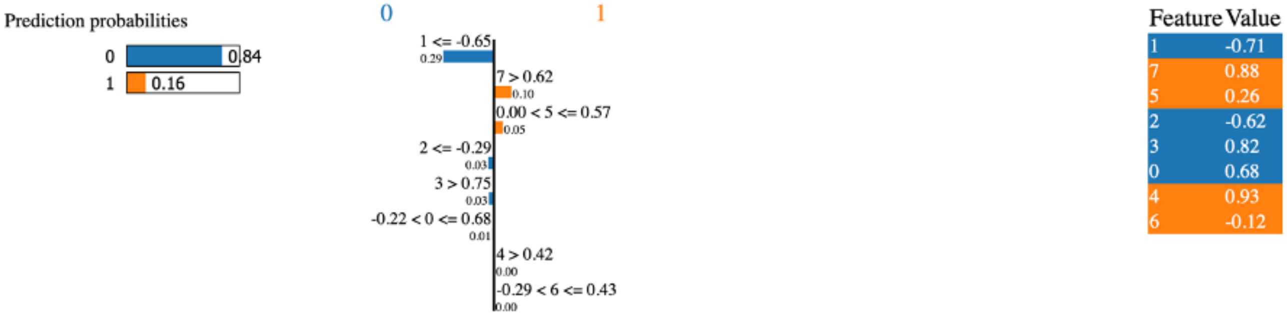

Fig.9: LIME for MLP

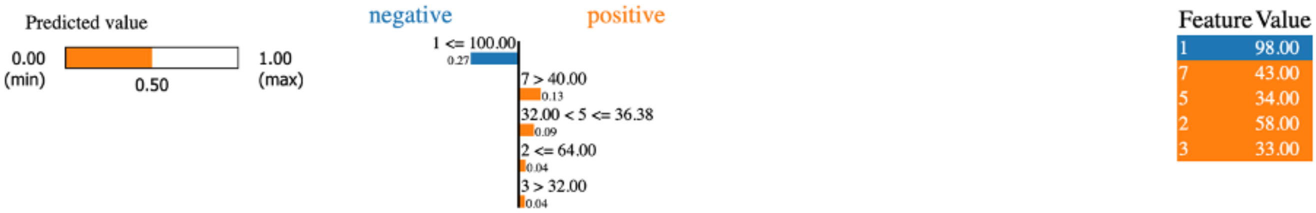

Fig.10: LIME for Regression

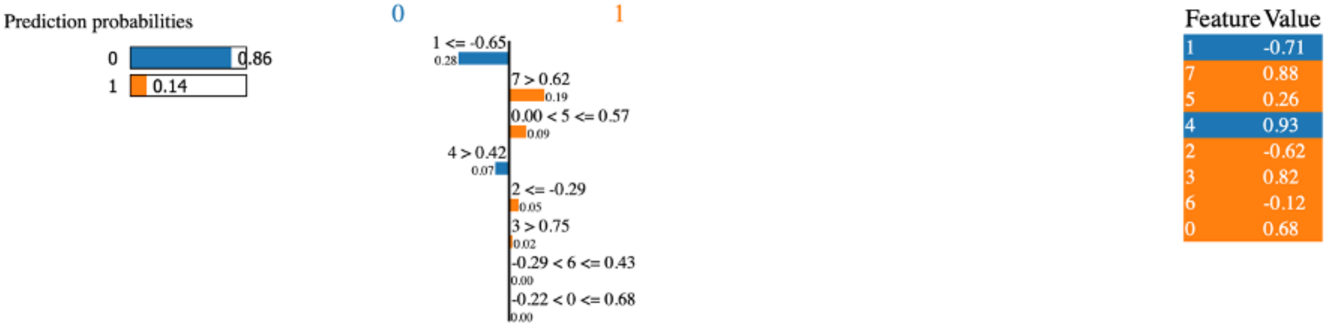

Fig.11: LIME for RNN

LIME, standing for Local Interpretable Model-Agnostic Explanations, emerges as a formidable tool for demystifying the predictions made by machine learning classifiers. It breaks down the complex decision boundaries carved out by advanced models into more intuitive linear relationships, casting light on the influential features for individual predictions. Through its interpretable visual aids, LIME conveys the essence of feature significance—green hues for features bolstering the predicted class, and red for those detracting from it. These visual narratives are not merely abstract; they also quantify prediction probabilities, offering a peek into the model's confidence in its assertions. The foundational prediction value, or the intercept, sets the stage before individual feature influences come into play. Moreover, the actual values for the instance at hand are laid bare, allowing for a tangible connection between data points and predictions.

In the scope of our research, as we delve into the LIME interpretations for models such as CNNs, MLPs, regression algorithms, and RNNs tested on initial datasets, a deeper comprehension of the decision-making processes of each model is unearthed. By comparing and contrasting these models, one can discern both uniformities and discrepancies in their reasoning patterns, providing a rich canvas for our research narrative.

Fig. 8 from the CNN model encapsulates its prowess in image-related tasks. The diagram plots "Prediction probabilities" for two classes, labeled "0" and "1," with the CNN tipping the scales towards class "1" at a probability of 0.67. This is portrayed alongside feature values, with feature "4" taking center stage due to its substantial positive contribution, as evidenced by the color-coded representation of the feature influences—blue for negative and orange for positive.

Fig. 9, depicting the MLP model, affirms its architecture as a versatile feedforward neural network. Although the MLP shares a similar visual language with the CNN in expressing prediction probabilities, the narrative differs with class "0" taking the lead at a probability of 0.84. Here, individual features play a divergent tune, with feature "1" dampening the prediction's chances, while feature "7" enhances it significantly.

Fig. 10 presents a regression model employing a Random Forest approach. This model diverges from the classification path and ventures into predicting continuous outcomes. The "Predicted value" scale is broad, and the model leans towards a lower predicted value, denoting a negative inclination. The feature values are numerically explicit, with feature "1" marked as the linchpin in this prediction scenario.

Finally, Fig. 11 aligns with the RNN model, underscoring its finesse with sequential data, a trait indispensable for time series analysis and language processing. The model's prediction heavily favors class "0" with a high probability of 0.86. The right side of the figure reiterates the importance of features "4" and "7," both exerting a strong positive pull on the predictive outcome.

LIME's forte lies in its capacity to illuminate the roles of individual features in model predictions, thus bolstering transparency and reliability in machine learning. Our comparative analysis across the CNN, RNN, MLP, and regression models points to a notable consistency in the influence of features such as 1, 7, and 4, despite varying prediction probabilities. These coherent interpretations, married with the insights drawn from LIME [13], advocate for its critical role in refining models, sculpting features, and bringing much-needed clarity to the domain of machine learning, highlighting its value in both research contexts and practical applications.

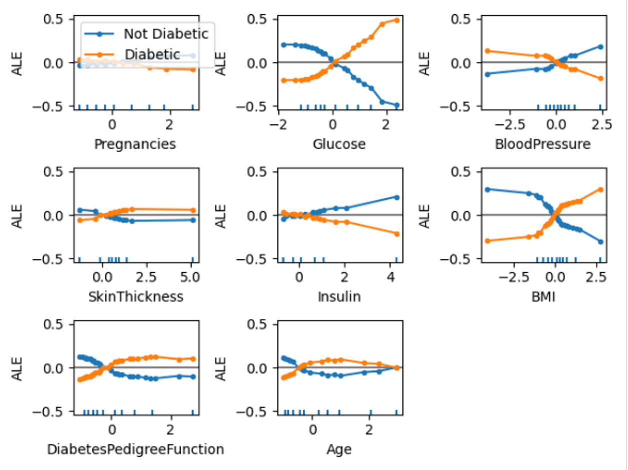

Fig. 12: ALE for CNN

Fig. 12 illustrates a series of Accumulated Local Effects (ALE) plots for a Convolutional Neural Network (CNN) model, specifically applied to diabetes prediction. Each plot is dedicated to a distinct feature—Glucose, BMI, and Age—revealing the impact of incremental changes in these features on the model's predictions. The plots differentiate the predictions for diabetic outcomes (represented by an orange line) versus non-diabetic outcomes (indicated by a blue line). The ALE values present a refined understanding of each feature's influence on the prediction outcome, distinctly independent of other variables. This collection of ALE plots provides a transparent representation of the model's behavior, showcasing how it responds across a range of values for each feature. Such visualizations are instrumental for researchers, offering a lucid interpretation of the CNN's predictive dynamics, where each feature's effect is carefully delineated, fostering an enhanced comprehension of the model's decision-making process in the critical domain of diabetes classification.

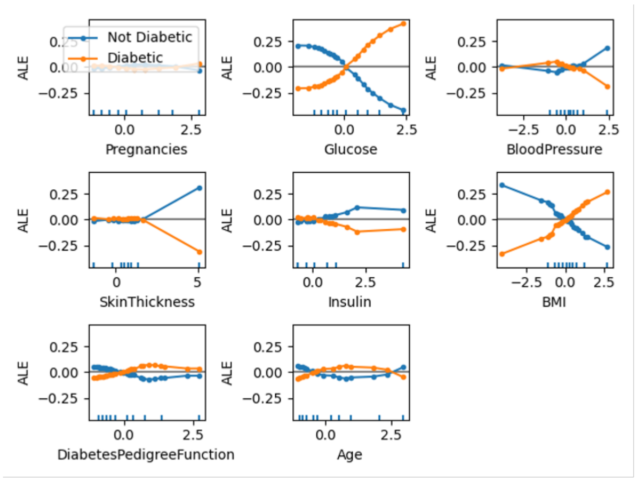

Fig. 13: ALE for MLP

Fig. 13 displays ALE plots for a Multi-Layer Perceptron (MLP) model, capturing the impact of different clinical features on diabetes prediction. The ALE values indicate the average model output change over a range of feature values, with distinct trends for diabetic versus non-diabetic classifications, highlighting how each clinical measure influences the MLP's prediction outcomes.

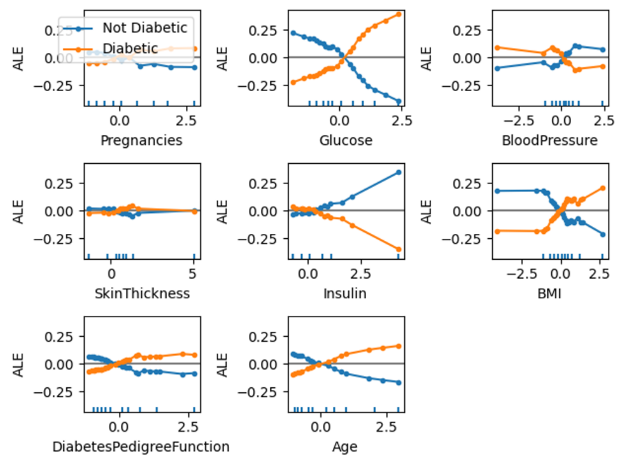

Fig. 14: ALE for RNN

Fig. 14 presents ALE plots for an RNN model, focusing on the 'Glucose' feature as a significant predictor for diabetes classification. The 'Glucose' plot illustrates a stark divergence in its influence on the classification of diabetic versus non-diabetic outcomes, underscoring its critical importance in the RNN's predictive performance.

Accumulated Local Effects (ALE) plots are a sophisticated instrument for visualizing and understanding the relationship between predictor variables and a model's output. When we analyze the ALE plots through the lens of three distinct machine learning architectures—CNN, MLP, and RNN—an intricate tapestry of the importance of various predictors for diabetes prediction unfolds.

For the CNN model, the ALE plots unravel various patterns. The number of pregnancies suggests a slight increase in diabetes likelihood, while glucose levels show a stark contrast in impact between diabetic and non-diabetic classes, affirming glucose's central role in the prediction of diabetes. Blood pressure presents an interesting trend: extreme values on either end are linked to a higher diabetes probability. Skin thickness contributes a negligible variation, implying a minor effect, whereas rising insulin levels show a significant increase in diabetes prediction. The plots for BMI indicate that both very low and high values are more frequently associated with diabetes. In contrast, the Diabetes Pedigree Function and Age reveal more uniform prediction probabilities, hinting at a less pronounced impact on the model's predictions.

Shifting the focus to the MLP model, the narrative alters slightly. Glucose, pregnancies, and BMI continue to be key features, with evident changes in prediction probabilities. Yet, features such as blood pressure and age may display different patterns from the CNN model, necessitating a nuanced approach to prediction interpretation.

The RNN model's ALE plots reinforce the prominence of features like glucose levels, pregnancies, and BMI, mirroring the CNN model's insights. Conversely, attributes like skin thickness maintain a low profile across all models, suggesting a lesser relevance in diabetes prediction tasks.

In essence, while each model exhibits its unique characteristics, features such as glucose levels, pregnancies, and BMI consistently stand out as significant predictors across the evaluated architectures. Grasping these subtle yet critical differences and similarities is crucial for leveraging these models to their fullest potential in practical, clinical settings.

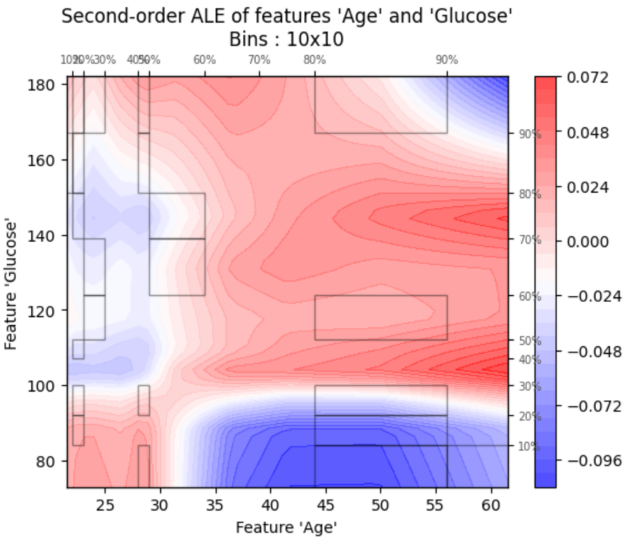

Fig.15: ALE for regression

Fig. 15 features a second-order ALE plot analyzing the interaction between 'Age' and 'Glucose' features within a regression model. The visualization employs contour lines and color gradients to articulate the joint effect of these two features on the model's predictions. Areas depicted in red signal a positive ALE value, suggesting that higher 'Age' and 'Glucose' levels jointly amplify the model's output, which could be associated with an elevated probability of diabetes. In contrast, regions shaded in blue denote a negative ALE value, where a combination of lower 'Age' and 'Glucose' levels leads to a reduction in the model's output.

The graph paints a picture of the interplay between 'Age' and 'Glucose' in influencing the model's predictive behavior. With 'Age' on the horizontal axis covering a span from 25 to 60 years and 'Glucose' on the vertical axis ranging from 80 to 180, the color-coded backdrop acts as a guide to the magnitude of their interaction effect. Notably, a younger age bracket (circa 25 years) with higher glucose levels (approaching 180) casts a positive influence on the model's output, while an older demographic (close to 60 years) with lower glucose levels (around 80) correlates with a negative influence. The interlacing pattern of blue and red hues across the plot captures the complex nature of this interaction, with both 'Age' and 'Glucose' intricately shaping the model's output across their respective spectrums.

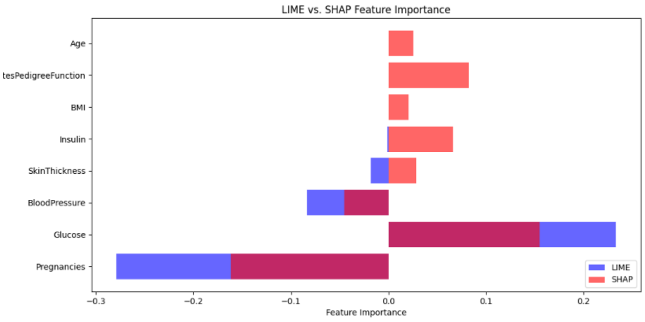

Fig. 16: LIME vs SHAP for CNN

Fig. 16 contrasts LIME and SHAP feature importance in a bar chart for a CNN model. This visual comparison brings to light the relative importance assigned to each feature by the two interpretability methods. In the chart, LIME's evaluations are represented by blue bars, whereas SHAP's assessments are depicted in red. A notable divergence is observed in the influence attributed to 'Pregnancies' and 'Glucose' by each method, with the bars reflecting significant discrepancies in their computed importance. This contrast not only highlights the method-specific perspectives on feature significance but also underscores the need for careful consideration when interpreting such analyses, as different techniques can yield varied insights into the model's decision-making process.

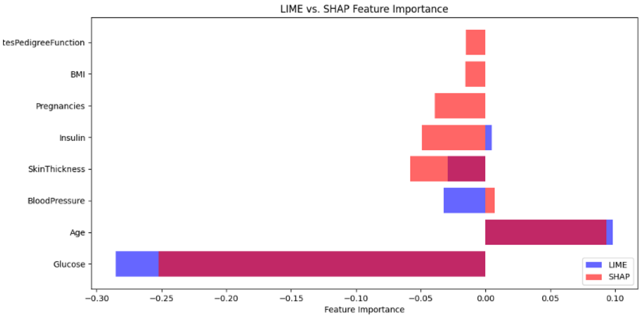

Fig. 17: LIME vs SHAP for MLP

Fig. 17 offers a comparative analysis of feature importance for an MLP (Multi-Layer Perceptron) model, as discerned by LIME and SHAP—two prominent techniques for interpreting machine learning models. The bar chart presents a side-by-side comparison, where LIME’s interpretation is indicated by blue bars, revealing a strong negative impact for 'Glucose'. On the other hand, SHAP's red bars indicate a higher positive importance assigned to features such as 'BMI' and 'Pregnancies'. This discrepancy not only highlights the unique approaches each method takes in evaluating feature influence but also reflects the inherent complexity in understanding machine learning models. The divergent interpretations exemplify the variability that can arise from different explanatory methodologies, emphasizing the importance of cross-verifying feature importance to obtain a more holistic understanding of the model's behaviour.

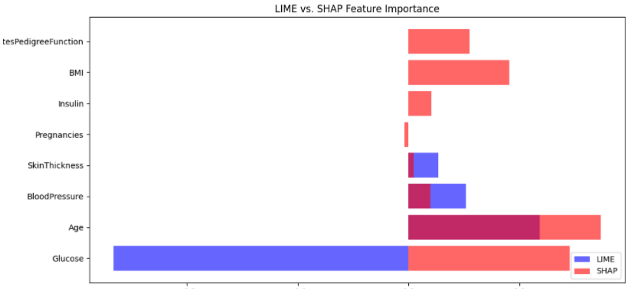

Fig.18: LIME vs SHAP for Regression

Fig. 18 depicts a comparative visualization of feature importance for a regression model, as evaluated by LIME and SHAP. This plot showcases the striking contrast in the interpretation of 'Glucose', which LIME analysis marks with a significant negative impact—represented by the blue bar. In juxtaposition, SHAP assigns 'Glucose' a smaller positive impact, indicated by the red bar. The bar chart further reveals the differential weightings of features like 'Age' and 'Blood Pressure', with each showing varying levels of importance across the two methods. This visualization underscores the contrasting lenses through which LIME and SHAP view the model's decision-making process, offering a nuanced perspective on the interpretability of predictive features within the regression framework.

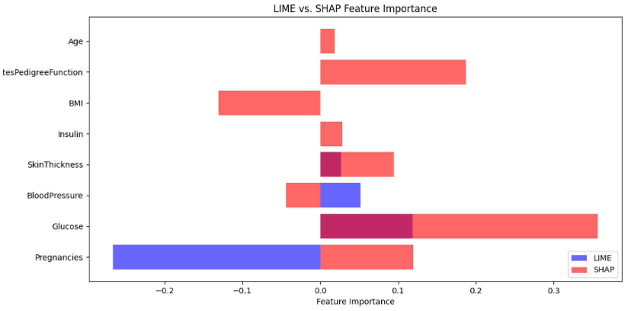

Fig. 19: LIME vs SHAP for RNN

Fig. 19 delineates a side-by-side analysis of feature importance for an RNN model, informed by LIME and SHAP methodologies. In this insightful plot, the varying bar lengths visually articulate each feature's estimated impact on the model's predictions, with LIME's blue bars and SHAP's red bars offering a stark illustration of their interpretive differences. The feature 'Pregnancies' commands notable emphasis in LIME's analysis, whereas 'Glucose' stands out in SHAP's, albeit with SHAP ascribing greater significance to 'Glucose'. This graphical representation serves as a compelling narrative on the diversity in local interpretability provided by these two prominent methods.

In a comparative lens, the CNN model's feature importance is scrutinized by both LIME and SHAP, unearthing convergences and divergences. While agreeing on the salience of features like glucose and pregnancies, LIME and SHAP part ways on others, such as the Diabetes Pedigree Function, where LIME sees more weight, contrasting with SHAP's prioritization of age and insulin.

The MLP model's examination further accentuates the distinct perspectives of LIME and SHAP. A shared acknowledgment of the significance of glucose and pregnancies exists, but their views split on other features like skin thickness and blood pressure, with LIME casting blood pressure in a significantly negative light, an assessment not shared by SHAP.

The regression model's interpretive journey reveals a harmonious understanding between LIME and SHAP over several features, yet they present a split verdict on the influence of age, with SHAP leaning towards a more negative attribution.

Turning to the RNN model, glucose’s influential status is corroborated by both methods, reflecting its critical predictive power. However, the weightage accorded to pregnancies by LIME displays a sharper negative tilt compared to SHAP. Additionally, other features like age and skin thickness receive varied importance ratings, indicative of the nuanced complexities in feature interpretation.

In synthesis, this comparative delve into LIME and SHAP's analyses across a suite of machine learning models—CNN, MLP, Regression, and RNN—unravels the intricate fabric of model interpretability. Glucose and pregnancies consistently emerge as features of note, yet the appraisal of other features is contingent on the chosen interpretive tool. Such diversity underscores the imperative of employing a spectrum of interpretability tools for a comprehensive and multi-faceted understanding of the models’ decision-making processes, which is crucial in the advancement of machine learning research.

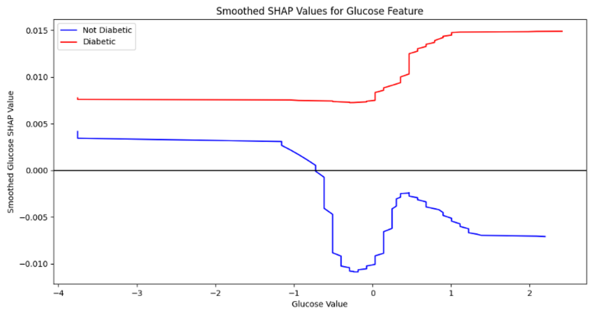

Fig.20: SHAP vs Glucose for MLP

Fig. 20 showcases a smoothed SHAP value plot for the 'Glucose' feature in an MLP (Multi-Layer Perceptron) model, which discerns between diabetic and non-diabetic predictions. The plot elegantly captures the relationship between the normalized values of 'Glucose' and their corresponding SHAP values. For non-diabetic predictions, illustrated in blue, we observe that as glucose levels ascend, the SHAP values become more negative, implying a lower probability of diabetes. Conversely, in the case of diabetic predictions, indicated in red, SHAP values exhibit an upward trend alongside rising glucose levels, signaling a heightened likelihood of diabetes. This visualization effectively conveys the dichotomous impact that glucose levels exert on the model's predictive behavior for the two distinct outcomes, delineating a clear gradient of influence that 'Glucose' levels have over the model's decision-making process.

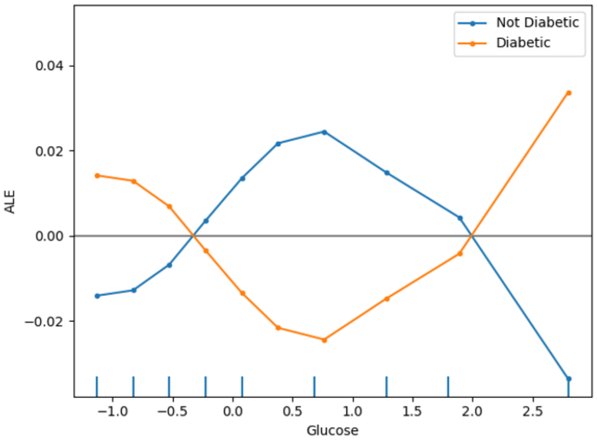

Fig.21: ALE vs Glucose for MLP

Fig. 21 illustrates an ALE (Accumulated Local Effects) plot for the 'Glucose' feature within an MLP (Multi-Layer Perceptron) model, which bifurcates the predictive outcomes into diabetic (depicted by the orange line) and non-diabetic (shown by the blue line) categories. The graph eloquently captures the 'Glucose' feature's value influence on the model's prediction output. For instances leading to a diabetic prediction, there is a discernible positive relationship; as the glucose values climb, so does the ALE value, reinforcing the association with an increased likelihood of diabetes. Conversely, the ALE value trends downward with rising glucose levels for non-diabetic predictions, signifying a negative correlation with the likelihood of not having diabetes. This visual contrast elucidates the variable impacts that glucose levels exert on the model's predictions for different classes.

The comparative analysis with SHAP and ALE for glucose values in an MLP model offers a tale of two narratives. The SHAP graph paints a relatively clear picture: higher glucose values suggest a stronger prediction towards diabetes, with a significant demarcation around the zero axis. In contrast, the non-diabetic trend through SHAP values is predominantly negative. Meanwhile, the ALE graph provides a more complex story, with fluctuations suggesting that the impact of glucose values on the model's predictions is less linear and more nuanced, crossing the zero threshold around the midpoint. These insights demonstrate that SHAP provides a more direct interpretation of glucose's role, whereas ALE reveals a layered, multifaceted impact of glucose levels, showcasing the intricate dynamics at play in model interpretation.

III. Conclusion

Our exploration into the predictive modelling of diabetes using advanced neural network architectures and classic regression analysis has yielded promising insights. Through the integration of XAI techniques, we have not only enhanced the predictive accuracy but also peeled back the layers of complexity that often shroud machine learning models, offering a window into the 'why' and 'how' of their predictions.

The use of SHAP, LIME, and ALE has been instrumental in demystifying the feature-specific influences across different models, including CNNs, RNNs, MLPs, and Regression models. By applying these interpretability tools, we've provided a granular analysis of feature importance that contributes to the robustness and reliability of our models in clinical settings. This nuanced understanding is pivotal, given the life-altering implications of diabetes diagnoses and the need for clinicians to trust and effectively utilize AI insights in patient care.

Furthermore, our research stands as a testament to the synergetic potential of AI and human expertise. By centering on interpretability, we ensure that AI serves as an aid, not an enigma, to healthcare professionals. The outcomes of our study advocate for a new standard in healthcare analytics, one where clarity and transparency are not afterthoughts but fundamental components of the design and implementation of AI.

In conclusion, this project not only advances the field of diabetes prediction through AI but also sets a precedent for future research to build upon. It underscores the importance of explainable and interpretable AI, paving the way for its broader adoption in healthcare and beyond, ensuring that AI's profound capabilities are fully leveraged to enhance human decision-making.