Proceedings of the 5th International Conference on Soft Computing for Security Applications (ICSCSA-2025)
DVD Part Number: CFP256A9-DVD; ISBN: 979-8-3315-9490-9

Multi-Domain Fusion with Explainable Boosted Learning (MF-EBL) for Diabetes Prediction

S.Paramaguru
Research Scholar
Department of Computer Science
Vels Institute of Science, Technology & Advanced Studies
Pallavaram, Chennai-600117
guruvetri90@myyahoo.com

L.Ramesh
Assistant Professor
Department of Computer Applications (UG)
Vels Institute of Science, Technology & Advanced Studies
Pallavaram, Chennai-600 117
lramesh.scs@vistas.ac.in

***Abstract*— Diabetes is a chronic metabolic disorder that is rapidly increasing in the world, which makes early and accurate prediction critical for timely intervention. The study introduces a new predictive methodology called Multi-Domain Fusion with Explainable Boosted Learning (MF-EBL) for predicting diabetes using clinical, demographic, and lifestyle datasets. The multi-domain fusion inputs advanced methods of machine learning (ML) including XGBoost, Random Forest(RF), Support Vector Machine(SVM), Logistic Regression(LR), K-Nearest Neighbors (KNN), and Neural Networks that were evaluated according to accuracy, F1 score, AUC-ROC, and explainability or interpretability of the ML model. Overall, XGboost prediction model performed the highest in all methods studied (accuracy of 88.7%, F1-Score of 0.87 and AUC-ROC of 0.91). The SHAP is applied to understand the interpretability of the model to predict diabetes without losing model accuracy. The study `conducted a data analysis and feature selection for the all data included in the study to provide high confidence for a robust model, and gained insight into the key risk factors such as glucose levels, age and body mass index (BMI). The MF-EBL framework presents a robust and trustable solution for monitoring and early detection of diabetes with results that are meaningful to clinicians who are providing personalized care to patients and the results can assist with decision-making not only through monitoring and also with early risk assessment.**

Keywords— Diabetes Prediction, Machine Learning, XGBoost, Explainable AI, Feature Fusion.

I. INTRODUCTION

Diabetes mellitus is a chronic metabolic disease that includes a notable rise in blood glucose levels and may lead to long-term complications including cardiovascular disease, renal failure or neuropathy. On a global scale, diabetes is increasingly being recognized as a public health threat, particularly Type 2 diabetes. Laboratory tests and clinical decisions have historically formed the basis for current eurodiagnosis of diabetes, however, they may be limited in identifying individuals at high risk of developing diabetes early enough, especially if they are asymptomatic [1]

Recently, ML efficiently identifies latent patterns in data and may also support the automation of decision making based on existing data, has shown promise in medical diagnostics. Nevertheless, existing ML-based diabetes prediction models have limitations [2]. Most models utilize only structured clinical data, and the models have not adequately exploited contextual information from demographic, lifestyle and behavioral factors. Additionally, some models are poorly interpretable and therefore unsuitable for clinical decision support applications where trust and transparency are crucial. Many of these models rely on training data that is unbalanced or have limited generalizability across populations. In order to meet the above challenges, a new framework has been proposed MF-EBL, which not only combines multi-source data (clinical, demographic, lifestyle), but also uses enhanced ML frameworks capable of explainability (i.e. Shapley Additive Expectations [SHAP])[3]. Using multi-domain data to combine clinical, lifestyle and demographic sources of data will not only provide the most accurate predictive options, but will do so in a transparent way. The key objectives of this study are:

- To develop a predictive modelling framework for diabetes using multi-domain data fusion with clinical, lifestyle and demographic factors as comprehensive predictors of risk factors not considered in single-domain approaches.
- To assess and compare performance of the proposed model with six ML algorithms such as XGBoost, Random Forest, Support Vector Machine (SVM), Logistic Regression, K-Nearest Neighbors (KNN), and Neural Networks) through standard performance evaluations including accuracy, F1-score, and AUC-ROC.

The paper is structured as follows Section 2, offer a literature review of recent advancements related to diabetes prediction, while in Section 3 details the proposed MF-EBL which consists of data preprocessing, model selection, and SHAP for explainability. Section 4 presents the experimental results and comparative performance across the six machine learning models, while Section 5 concludes the study and suggests future work.

II. RELATED WORKS

Recent advancements in ML have resulted in recognizable improvements in diabetes prediction utilizing clinical and lifestyle data. Differences in modeling principles include neural networks, ensemble learning, and then some, are all utilized in improved predictive diagnostics and decision support.

Chou et al., (2023) examined eight distinct subject parameters, including age, body mass index, insulin level, diastolic blood pressure, plasma glucose level, number of pregnancies, sebum thickness, and diabetes pedigree function. Once all of the patient data had been sorted, the models of different types of neural networks were trained using Microsoft ML Studio. The prediction results were then used to compare the predictive power of the different diabetes parameters [4]. Olisah et al., (2022) proposed a strong framework for developing a diabetes prediction model to aid in the clinical diagnosis is proposed. From a viewpoint that improves their performances, the framework

incorporates the use of polynomial regression for missing value imputation and Spearman correlation for feature selection, respectively [5].

Modak et al., (2024) offers a novel diabetes prediction model that uses a variety of machine learning methods, such as RF, SVM, Naïve Bayes(NB), and LR. Researchers further improve prediction accuracy and robustness by utilising ensemble learning in addition to these fundamental methods [6]. Dharmarathne et al., (2024) presents the first self-explanatory interface for machine learning-based diabetic diagnosis. This interface is essential because it diagnoses patients and provides clear justifications for the choices taken, giving users a better understanding of their present medical situations [7].

Dashdondov et al., (2024) ,emphasizes how important thorough feature selection and outlier detection are to improving the prediction ability of diabetes risk models. Notably, during the COVID-19 pandemic, there was a notable rise in diabetes cases, which were specifically associated with male sex, advanced age, living in a rural area, having high blood pressure, and being obese. This highlights the need for improved public health policies for targeted prevention and early intervention [8]. Oliullah et al., (2024) employ a variety of ML techniques to identify diabetes early on, especially in women. The goal of this research is to use these techniques to give doctors useful tools for early disease detection, allowing for prompt interventions and better patient outcomes[9].

En-RfRsK, a voting classifier that combines three ML techniques, RF, radial SVM and KNN is proposed to predict the risk of diabetes mellitus. The performance of numerous, comparatively uneven models or trees is used by RF. The suggested method makes use of the benefits of these ML approaches [10]. NG et al., (2025) offers a brand-new ML-based intelligent diabetes mellitus prediction framework (IDMPF). The framework is the outcome of a thorough analysis of diabetes prediction models found in the literature [11]. Table 1 1ists the recent studies pertaining to the current study.

Table 1: Related works pertaining to the current study

+	Methods Used	Strengths	Limitations
Chou et al., (2023)	Neural Networks (various types), Microsoft ML Studio	Analyzed 8 diverse clinical parameters Comparison of multiple NN architectures	Limited focus on missing data or feature engineering No explainability framework
Olisah et al., (2022)	Polynomial Regression (imputation), Spearman Correlation (feature selection)	Robust pre-processing strategy Enhanced feature relevance via statistical correlation	No ensemble or deep learning methods used Interpretability not addressed
Modak et al., (2024)	RF, SVM, Naïve Bayes, Logistic Regression, Ensemble Learning	Model diversity improves prediction Ensemble boosts accuracy and robustness	Increased model complexity Lack of model transparency
Dharmarathne et al., (2024)	ML classification (unspecified) + Explainable Interface	Offers transparent decision explanations Focused on patient communication	Interface usability not validated clinically Specific ML models not detailed
Dashdondov et al., (2024)	Feature Selection, Outlier Detection, Statistical Analysis	Incorporates COVID-19-related risk insights Emphasizes feature refinement	Region-specific findings No details on prediction models
Oliullah et al., (2024)	Multiple ML models for early detection	Focus on women's health Supports early intervention for better outcomes	Gender bias may affect generalization Lacks interpretability tools
NG et al., (2024)	Ensemble (Voting Classifier): RF, Radial SVM, KNN	Combines strengths of individual classifiers Enhances predictive accuracy	Computational overhead Black-box nature limits interpretability
NG et al., (2025)	Intelligent Diabetes Mellitus Prediction Framework (IDMPF)	Built from comprehensive model analysis Framework integrates intelligent ML logic	Limited novel algorithmic contribution Validation scope not clarified

Although ML is being used more and more to forecast diabetes, many models are opaque and difficult to understand, which restricts their clinical use. Robust techniques for handling missing data and outlier detection are frequently overlooked by current methodologies. The majority of research is dataset-specific, which raises questions about how well the model can be applied to different populations. It is rare for early prediction techniques to give end users tailored feedback or

explanations. Furthermore, risk variables brought on by previous public health emergencies, such the COVID-19 pandemic, are not well integrated.

III. METHODOLOGY

The proposed model utilizes a diverse array of features, including: structured clinical data and lifestyle and demographic attributes; concentrations of relevant biomarkers (especially glucose and HbA1c) and, importantly, possibly even genetics and family history – all to increase the accuracy and reliability of diabetes predictions. The model will utilize a feature-level fusion framework (i.e. having various data sources amalgamate into a common representation prior to classification). The predictive modelling uses XGBoost[12] (an efficient gradient-boosted tree algorithm) for predictions, since it works well on complex patterns with high dimensional data and interactions in features. In addition, SHAP (SHapley Additive exPlanations), which provides indications of the contributions and reasoning for a prediction (for the stakeholder) will be employed for interpretability of the model's predictions. The overall flow diagram for the suggested methods is shown in figure 1

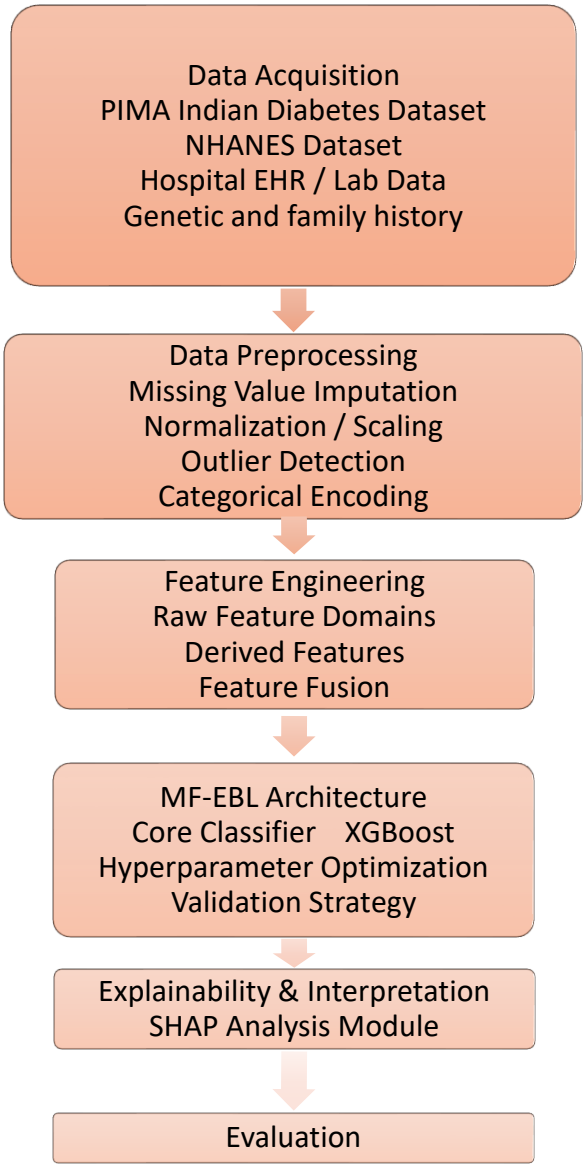

Figure 1: Proposed Model for Diabetics Prediction

A. Data Collection and Preprocessing

The model is designed using multiple complementary datasets to allow for applicability and robustness. The PIMA Indian Diabetes Dataset has well organized clinical data, which is especially beneficial for testing models in the South Asian population [13]. The NHANES (National Health and Nutrition Examination Survey) contributes a vast number of lifestyle and demographic variables, showing us how behavioral and social factors may affect diabetes risk. The additional lab-provided or hospital Electronic Health Record (EHR) data offers clinical data and biomarkers gathered in a real-world scenario for assessing clinical relevance and generalizability of the model, in potentially a wide variety of populations.

Preprocessing: A complete preprocessing pipeline is developed to verify data quality and model fitness:

- Imputation of missing values: Missing values are imputed according to the imputation strategy based on whether the feature is continuous or categorical, and filled with median or mode value. Regardless of the imputation strategy employed, bias is reduced and the integrity of the data is maintained.
- Standardized or Min-Max data: Feature values are standardized or min-maxed; both multidimensional values are within a comparable range to not put machine learning algorithms, like XGBoost, at a disadvantage, along with bettering the performance and increasing the chances of our data to converge.
- Outlier detection with Isolation Forest: An Isolation Forest is used for outlier detection, to locate and eliminate records that are outliers that are likely to skew the model training. Isolation Forest utilizes unsupervised learning, is appropriate for high dimensional data, and isolates instances based on being rare.
- Categorical encoding: Categorical features are encoded to machine-readable formats by one-hot encoding nominal features, and label encoding ordinal features so that the model can ingest and use categorical training instances.

B. Feature Engineering

To expand the input space and increase predictive power, the model draws from three sources of data: clinical, demographic, and lifestyle. This holistic modeling allows the model to consider both physiological and behavioral aspects of diabetes risk. Data can be considered in raw form, but derived features through domain-specific transformations and "calculations" (ratios, trends metrics, for instance) allow for added insight into risk-associated features. Some of the key features and its description are shown in table 2.

Table 2 Features and its description

Domain	Feature Name	Description
Clinical	Glucose	Fasting blood glucose level, a key biomarker for diabetes detection.
	BMI	Body Mass Index, indicates body fat based on height and

		weight.
	BP	Blood pressure, associated with metabolic syndrome and cardiovascular risk.
Demographic	Age	Patient's age; older individuals are generally at higher risk.
	Gender	Biological sex, influencing disease prevalence and response to treatment.
	Race	Ethnic background, used to account for genetic and socioeconomic differences.
Lifestyle	Exercise	Frequency or intensity of physical activity.
	Alcohol	Consumption level; excessive intake is a diabetes risk factor.
	Smoking	Smoking status; known to impact insulin sensitivity.
Derived	Glucose/BMI Ratio	Combines metabolic and body composition factors to assess proportional risk.
	Age-adjusted Risk Score	Weighted feature incorporating age-specific diabetes risk factors.
	HbA1c Velocity	Rate of change in HbA1c over time (if longitudinal data is available).

Recursive Feature Elimination (RFE): RFE is a method of feature selection that is considered a "wrapper" method because it uses a learning algorithm's ability to assess features as part of its input processing. RFE starts with all the features based on the performance of the chosen model. The model will be trained, and ranked the importance of each feature before recursively eliminating the least important features as defined by their importance. RFE continues recursively removing the least important features until the desired amount of features is left. The end product is a subset of features that has the most important features that lead to better accuracy and reduced chances of overfitting[14].

Mutual Information: Mutual Information is a filter method that measures the dependency between the target variable and each feature. Mutual information examines how much knowing a feature's value reduces uncertainty about the outcome by looking at the general relationships between the possible values of the features against the outcome--this means that mutual information measures largely linear and non-linear relationships among features and the target variable. Features are ranked based on mutual information metrics, and features with the most mutual information metrics are considered. The best features will be considered first for model inclusion, making sure that the selected inputs, or features, have good statistical association to diabetes risk level**.**

SHAP Importance Ranking: SHAP (SHapley Additive exPlanations) viables provide an interpretable, game-theoretic perspective on feature importance, assigning an individual contribution for each feature's influence on individual predictions. Because SHAP assesses the average of the individual contributions of features visually, it can be used to create a global ranking of features based on their mean effect. It also provides some benefit in meaningfully assessing predictive relevance, while still providing a certain degree of alignment to the interpretability goal of the model, given it would provide the clinician with a clear understanding of the influence of each feature on their decision-making. Performance and interpretability or transparency are balanced when these complementary approaches are used to guarantee that the final feature set is predictive and interpretable in a clinical setting[15].

C. Model Architecture – Explainable Boosted Learning:

The proposed model MF-EBL is based on the Explainable Boosted Learning (EBL) framework that combines advanced predictive modeling with interpretable modeling. The architecture utilizes XGBoost, which is a gradient-boosted decision tree algorithm that achieves the highest performance and scalability today and utilizes both numerical and categorical data and features. XGBoost is useful in healthcare applications for three primary reasons: It can handle missing data; it can model non-linear relationships in the data; and it can rank feature importance.

XGBoost is used as the core model of the EBL model because, amongst structured data tasks, it is the few models that can have confidence in its predictive ability. Also it's inbuilt regularisation mechanisms to mitigate against overfitting and its ability to address complex feature interactions and mixed data types. The framework of combining trees via ensemble learning allows XGBoost to model higher order interactions of the features, especially since they are typically not explicitly pre-processed or transformed.

The model is reported and optimised for performance with two hyperparameter tuning methods; Grid Search and Optuna. Unlike Optuna, which uses a more iterative and Bayesian optimised methodology, Grid Search will only do a constant exploratory search over a predetermined hyperparameter combinations space. Therefore, the precaution of using dual methods will leave little to chance, while also promising advancement in efficiency.

The final model is validated using stratified K-fold cross-validation. This method divides the dataset for validation purposes into K subsets (folds), accounting for the target variable class distribution in each fold, using stratification. In practice, stratification means that diabetic cases are represented proportionately to the same amount of representation as non-diabetic cases calculated as proportions of the sampled size. The model will train and

validate on different combinations of these folds to demonstrate stability and avoid a source of bias inherent to leveraging a single split. This combination of a powerful classifier, systematic optimization, and rigorous validation forms a robust and interpretable foundation for diabetes risk prediction.

D. Explainability Module

The explainability module utilizes SHAP, which is a cutting-edge technique for interpretability based on cooperative game theory. SHAP measures how much each feature contributed to the prediction of the model with insights both globally and locally. Globally it highlights which features influenced model decision making for the whole dataset to highlight the main risk factors and the risk factors that may point to biases in the model. Locally, it provides an explainer at the patient level showing how each individual feature pushed a specific prediction toward diabetic or non-diabetic. This type of detailed visibility is essential within healthcare systems as it will build clinician trust, provide greater transparency into the model, and enable more informed and personalized medical decision-making with a diabetic population.

IV. Results and Discussion

The performances of all ML models used in the MF-EBL (Multi-source Fusion/Explainable Boosted Learning) diabetes prediction framework are shown in table 3 using three evaluation metrics: Accuracy, F1-Score, AUC-ROC.

Table 3. Performance Analysis –Proposed MF-EBL

Method	Accuracy	F1-Score	AUC-ROC
XGBoost	88.7%	0.86	0.91
Random Forest(RF)	86.3%	0.83	0.89
Support Vector Machine(SVM)	83.0%	0.80	0.85
Logistic Regression(LR)	81.5%	0.78	0.83
K-Nearest Neighbors (KNN)	80.2%	0.76	0.81
Neural Network (MLP)	85.1%	0.82	0.87

XGBoost outperformed all the models in terms of accuracy (88.7%), F1 score (0.86) and AUC-ROC (0.91). This indicates that XGBoost exhibited good overall performance regarding classifying patients, a fair level of precision and recall, and good discrimination capability further downfield in the thresholds. RF performs next best, with decent metrics (accuracy: 86.3%, F1-score: 0.83, AUC-ROC: 0.89). This indicates that this algorithm would be an acceptable alternative but is still behind XGBoost in each category. SVM and neural network (MLP) demonstrated moderate performance values of 83% for SVM and 85.1% for MLP. While MLP is slightly better than SVM, both still trail the tree-based models. LT and KNN represent the lowest performing traditional approaches. Thus, it is not surprising that KNN has the lowest metrics of any model employed (accuracy: 80.2%, F1-score: 0.76, AUC-ROC: 0.81), which allows us to conclude it has little applicability in identifying the complexities of predicting diabetes. The results confirm the use of XGBoost as the main model in MF-EBL, showing the best performance across all measures, and support the benefits of multi-domain features with advanced, explainable learners.

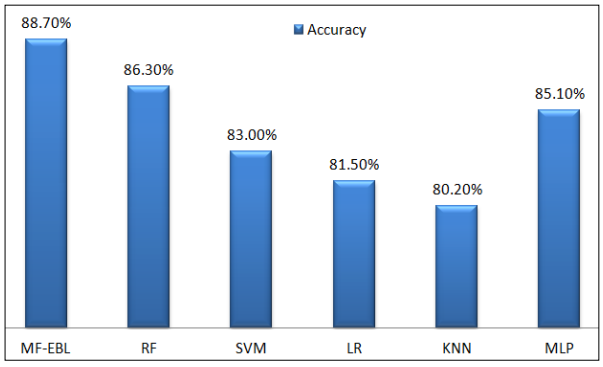

Figure 2 Performance Analysis –Accuracy

Figure 2 shows the classification accuracy of the proposed method with the contrasted methods The MF-EBL (Multi-Feature Explainable Boosted Learning) method demonstrated the highest classification accuracy (88.70%) of all the models presented showing it is more effective than those examined. Next, Random Forest (RF) and Multi-Layer Perceptron (MLP) are also high in terms of classification accuracy (86.30% and 85.10% respectively) that can be interpreted as fair to very good predictive ability. The Support Vector Machine (SVM), Logistic Regression (LR), and K-Nearest Neighbor (KNN) models were lower with accuracies of 83.00%, 81.50%, and 80.20% respectively which can be defined as acceptable classifiers for predictive outcomes. An ensemble and/or deep learning approach, MF-EBL, shows it was more effective and achieved higher classification accuracy in complex predictive tasks.

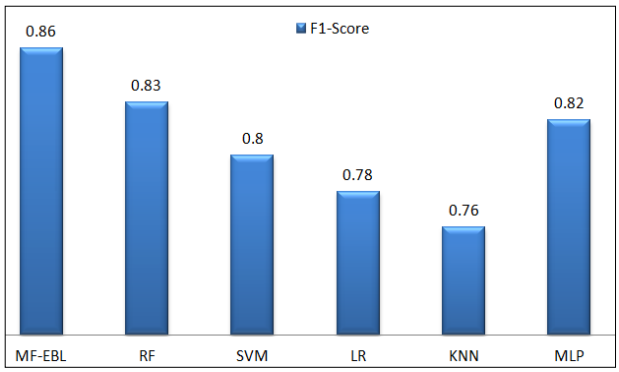

Figure 3 :Performance Analysis-F1-Score

The comparative analysis of F1-Scores of the suggested methods with 5 other ML models are shown in Figure 3. The MF-EBL (Multi-Feature Explainable Boosted Learning) model obtained the best F1-Score of 0.86, which is indicative of the models excellent balance of precision and recall. Between the other approaches, the RF model and MLP model attained the next-best scores of 0.83 and 0.82 respectively. The SVM model had a moderate F1-Score of 0.80 and the other remaining methodologies such as LR and KNN had F1-Scores of 0.78 and 0.76 respectively. Collectively these results indicate that advanced ensemble and deep learning models like MF-EBL can produce

stronger and more reliable classification results as they relate to complex predictive tasks.

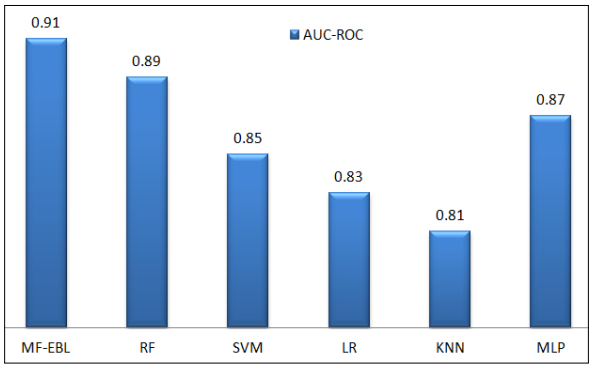

Figure 4: Performance Analysis-AUC-ROC

Figure 4 shows the AUC-ROC values of the suggested method with the contrasted methods .The suggested model developed (MF-EBL) achieved the highest AUC-ROC value (0.91) which is excellent discrimination. RF and MLP also achieved good AUC-ROC values (0.89 and 0.87). Other traditional models (namely, SVM, LR had moderate AUC-ROC values (0.85, 0.83) while KNN achieved the lowest AUC-ROC value performance (0.81). Altogether, the results suggest that ensemble and DL approaches, especially MF-EBL, exhibit better class separation ability than traditional models which allows these to be more applicable in addressing complex classification challenges.

Table 2: Features and its SHAP Values

Features	SHAP Value (mean)
Glucose	0.31
BMI	0.24
Age	0.18
Physical Activity	0.12
Family History	0.10
Blood Pressure	0.09
Insulin Level	0.07
Skin Thickness	0.05
Pregnancies	0.04
Cholesterol Level	0.03

Table 2 includes mean SHAP values for ten important features to a diabetes prediction model, and this allows the researchers to analyze effects of each feature on the model output. Glucose has the highest SHAP value of .31. This is not unexpected since glucose level is one of the predictor variables of diabetes, and elevated glucose is often the first indicator of the onset of diabetes. The next two most influential features are BMI (.24) and Age (.18), which can both indicate obesity and age, impacts the incidence of diabetes. Similarly, both Physical Activity (.12) and Family History (.10) make significant contributions to the model, which are indicators of lifestyle and genetic factors. Other clinical features, such as Blood Pressure (.09) and Insulin Level (.07) had relatively low influence and made only decisive contributions. Finally, Skin Thickness (.05), Pregnancies (.04) and Cholesterol Level (.03) collectively contributed almost nothing to the prediction outcome. The SHAP values provide transparency as to how the features outlined in the model contributed to the diabetic predictions from the model. The SHAP values are helpful for clinicians with understanding and explaining risk profiles for patients in regard to diabetes.

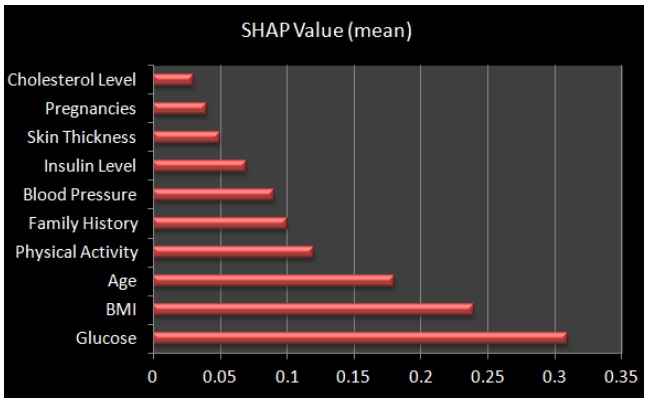

Figure 5 SHAP Analysis

Figure 5 shows the average SHAP values of ten features which contributed to the diabetes prediction model, showing the impact on the decision process relative to the other features. Glucose as expected, is the most influential feature with the most SHAP sides showing it's significance in the diagnosis of diabetes. BMI and Age come next, both having known risks associated with diabetes coming from obesity and age as a factor. And Physical Activity and Family History demonstrate a mix of lifestyle influences and genetic influences. Blood Pressure and Insulin Level are moderate influences that certainly have relevance to someone's metabolic health.. Skin Thickness, Number of Pregnancies, and Cholesterol Level had much lower SHAP values making only minor contributions to the model but still have predictive value. This interpretation through SHAP enhances transparency of the model helping make diabetes predictions more explainable and clinically relevant.

The unique aspect of the proposed methodology is its inclusive and pragmatic structure for a real-world health context. First, it offers multi-domain feature fusion, including clinical, lifestyle and demographics that provides a holistic view of the health of the patient. Second, it incorporates interpretable learning, which can build clinician and patient trust by providing transparency and explainable learning to understand its decisions. Third, it can do risk stratification, and provide personal feedback and personalized healthcare activities, based on risk profiles of the patients. Lastly, it is designed to be adaptable for real-world data, by appropriately handling noise, imbalanced datasets, and heterogeneous data, making it amenable for clinical use across the continuum of health and across different contexts.

V. CONCLUSION

This study has presented the MF-EBL (Multi-Domain Fusion with Explainable Boosted Learning) framework for accurate and explainable diabetes prediction. Using clinical, demographic and lifestyle characteristics and testing six machine learning models it was found that XGBoost was the best algorithm as it produced the strongest accuracy (88.7%), F1-score (0.86), and AUC-ROC (0.91). There was additionally an increased level of interpretability felt when using SHAP values which provided a transparent(ish) decision style with the supporting evidence of the main

contributing risk factors such as glucose levels, BMI and age. The implementation of data-driven explainable AI tools indicates the value of utilizing this technology to promote early diagnosis and assist with personalized healthcare. Future works will include the expansion of the data set in real-time longitudinal record form, and or, integrating multiple sources of data from wearables and devices; understanding the model's capabilities and performance within varying populations.

REFERENCES

[1] Ahamed, B. Shamreen, Meenakshi S. Arya, S. K. B. Sangeetha, and Nancy V. Auxilia Osvin. "Diabetes mellitus disease prediction and type classification involving predictive modeling using machine learning techniques and classifiers." Applied Computational Intelligence and Soft Computing 2022, no. 1 (2022): 7899364.

[2] Jaiswal, Varun, Anjli Negi, and Tarun Pal. "A review on current advances in machine learning based diabetes prediction." Primary Care Diabetes 15, no. 3 (2021): 435-443.

[3] Prendin, Francesco, Jacopo Pavan, Giacomo Cappon, Simone Del Favero, Giovanni Sparacino, and Andrea Facchinetti. "The importance of interpreting machine learning models for blood glucose prediction in diabetes: an analysis using SHAP." Scientific reports 13, no. 1 (2023): 16865.

[4] Chou, Chun-Yang, Ding-Yang Hsu, and Chun-Hung Chou. "Predicting the onset of diabetes with machine learning methods." Journal of Personalized Medicine 13, no. 3 (2023): 406.

[5] Olisah, Chollette C., Lyndon Smith, and Melvyn Smith. "Diabetes mellitus prediction and diagnosis from a data preprocessing and machine learning perspective." Computer Methods and Programs in Biomedicine 220 (2022): 106773.

[6] Modak, Sandip Kumar Singh, and Vijay Kumar Jha. "Diabetes prediction model using machine learning techniques." Multimedia Tools and Applications 83, no. 13 (2024): 38523-38549.

[7] Dharmarathne, Gangani, Thilini N. Jayasinghe, Madhusha Bogahawaththa, D. P. P. Meddage, and Upaka Rathnayake. "A novel machine learning approach for diagnosing diabetes with a self-explainable interface." Healthcare analytics 5 (2024): 100301.

[8] Dashdondov, Khongorzul, Suehyun Lee, and Munkh-Uchral Erdenebat. 2024. "Enhancing Diabetes Prediction and Prevention through Mahalanobis Distance and Machine Learning Integration" *Applied Sciences* 14, no. 17: 7480. https://doi.org/10.3390/app14177480

[9] Oliullah, Khondokar, Mahedi Hasan Rasel, Md Manzurul Islam, Md Reazul Islam, Md Anwar Hussen Wadud, and Md Whaiduzzaman. "A stacked ensemble machine learning approach for the prediction of diabetes." Journal of Diabetes & Metabolic Disorders 23, no. 1 (2024): 603-617.

[10] NG, Bhuvaneswari Amma. "En-RfRsK: An ensemble machine learning technique for prognostication of diabetes mellitus." Egyptian Informatics Journal 25 (2024): 100441.

[11] Abdelhafez, Hoda A., and Abeer A. Amer. "Machine Learning Techniques for Diabetes Prediction: A Comparative Analysis." Journal of Applied Data Sciences 5, no. 2 (2024): 792-807.

[12] Li, Wenguang, Yan Peng, and Ke Peng. "Diabetes prediction model based on GA-XGBoost and stacking ensemble algorithm." PloS one 19, no. 9 (2024): e0311222.

[13] Mishra, Soumya Ranjan, and Sachikanta Dash. "Machine Learning Based Diabetes Prediction Using the PIMA Indian Dataset." In 2024 2nd International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES), pp. 1-6. IEEE, 2024.

[14] Shams, Mahmoud Y., Zahraa Tarek, and Ahmed M. Elshewey. "A novel RFE-GRU model for diabetes classification using PIMA Indian dataset." Scientific Reports 15, no. 1 (2025): 982.

[15] Ahmed, Shamim, M. Shamim Kaiser, Mohammad Shahadat Hossain, and Karl Andersson. "A comparative analysis of lime and shap interpreters with explainable ml-based diabetes predictions." IEEE Access (2024).