International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(13s) 2026 pp. 688-698
https://doi.org/10.70917/ijcisim-2026-4104
Article

Enhancing Cardiovascular Risk Prediction with Explainable AI using Clinical Data

K. Naga Deepthi[1], P. Bhargavi[2]

[1,2]Dept. of. Computer Science, Sri Padmavati Mahila Visvavidyalayam, Tirupati, India

Abstract: Diabetes is a major risk factor for the development of cardiovascular issues which contribute to cardiovascular disease (CVD) being a leading cause of mortality worldwide. However, traditional machine learning methods are not widely adopted in healthcare systems because they lack interpretability, which is important for early and accurate CVD risk prediction and for ruling out effective clinical intervention. In this research, a hybrid architecture is proposed that incorporates diabetes related datasets as well as explainable artificial intelligence (XAI) methodologies that could improve the prediction power and transparency of the models. The proposed approach combines different datasets at the level of features and includes rigorous data pre-processing to detect metabolic and cardiovascular risk factors. Some of the significant clinical parameters are age, BMI, glucose, cholesterol, and blood pressure. These are standardized to create a single dataset which may be utilized for predictive modelling. The employment of two XAI approaches, SHAP (SHapley Additive Explanations) with tree-based ensemble models and integrated gradients with transformer based topologies, ensures both performance and interpretability. The technique improves confidence and usefulness in clinical settings by offering accurate predictions and explanations for the model's judgments that are relevant to the circumstance. It is also utilized for visual investigation of clinical correlations of diabetes and cardiovascular disease and identify crucial risk variables. The results suggest that merging explainability approaches with powerful machine learning can considerably boost early identification and risk assessment. The proposed approach contributes to enhanced healthcare decision-making, offering a scalable, interpretable and dependable solution for cardiovascular disease prediction.
Keywords: Cardiovascular Disease, Explainable Artificial Intelligence, SHAP, Integrated Gradients, Diabetes.

1. Introduction

One of the leading causes of death worldwide, cardiovascular disease (CVD) affects a large percentage of the population every year [22]. Cardiovascular problems are already at an elevated risk due to the rising incidence of diabetes; metabolic disorders are major contributors to the onset of heart disease. In order to intervene promptly and provide better patient care, early detection and precise risk prediction of CVD are crucial. As more and more medical records are made public, deep learning and machine learning have become potent instruments for illness prognosis [23]. Unfortunately, the lack of interpretability and high accuracy offered by many of these models prevents them from being widely used in clinical settings.

There has been increasing interest Explainable Artificial Intelligence (XAI) in recent years as potential solution to this limitation. as a potential solution to this constraint. The goal of XAI approaches is to increase confidence among healthcare providers by providing openness regarding how models make their predictions [24]. Here, datasets include diabetes and cardiovascular disease data that make it easy to improve predictive performance. However, the integration of heterogeneous datasets with correct representation of the features is not trivial [25]. Thus, a trade-off between model accuracy and interpretability is crucial for the development of reliable clinical decision-support systems.

This research proposes a comprehensive architecture for cardiovascular disease risk prediction utilizing XAI algorithms and datasets related to diabetes. For interpretability of structured data, the method combines SHAP and tree-based ensemble models for capturing complex patterns in healthcare records; it also combines integrated gradients and transformer-based architectures. The aim is to build a resilient system that can reliably predict events and provide compelling explanations for the model's decisions. The experimental results reveal the effectiveness of the proposed

strategy and also demonstrate the importance of the combination of explainability with deep leaning for better healthcare.

2. Literature Review

The increasing availability of electronic health records and clinical data [7] has dramatically improved the possibility of utilizing computer techniques for illness prediction. Anyway, it is hard to predict disease such as cardiovascular disease in diabetic patients, especially in the early stages, where clinical symptoms are sometimes absent or difficult to be interpreted [8]. The introduction of machine learning as a strong tool in the construction of prediction models to uncover hidden patterns and complicated connections among clinical variables has helped to bridge this knowledge gap [9]. Several international research have focused on the prediction of CVD in patients using different ML algorithms. Sang et al., [10] employed a series of techniques to assess an online collected dataset such as logistic regression, support vector machine, random forest and extreme gradient boosting. The random forest approaches were shown to perform best with an area under the curve AUC-ROC of 0.830. the difficulty is that they did not include crucial lifestyle risk variables such as smoking, heavy drinking and inactivity in their model.

In their 2024 study, Nrusimhadri et al. [11] evaluated a number of methods, including as decision trees, AdaBoost, SVMs, ANNs, and a tailored ANN. With an accuracy of 94%, the customized ANN proved that optimizing model architecture can have a major impact on prediction results. were evaluated for the purpose of predicting cardiovascular disorders; the results showed that CANN achieved the best accuracy at 94%, while decision trees, AdaBoost, SVM, and ANN all achieved 72%, 72%, 81%, and 81% accuracy, respectively.

A machine-learning algorithm was created by Sang and colleagues that accurately detects diabetes patients who are at high risk of cardiovascular disease [12]. Predicting three-year CVD risk in diabetic patients was another use of XGBoost [13]. Metabolic patients can also benefit from ML's ability to discover novel CVD risk variables [14]. Additionally, several new cardiovascular disease prediction models based on machine learning have been developed for type 2 diabetes mellitus. However, certain risk factors can be removed, which reduces their predictive power [15,16]. The use of deep learning in conjunction with more conventional machine learning techniques to improve medical risk prediction has recently attracted a lot of attention from researchers [17, 18]. Recent studies [19,20] have shown that hybrid models that combine structured clinical data with temporal patterns can improve the accuracy of chronic disease prediction. However, at the same time, it is of greatest importance to ensure fairness and eliminate bias in medical models to maintain the dependability of predictions across diverse populations. Explainable models such as SHAP and LIME have been used to make cardiovascular risk models more interpretable to doctors [21]. In line with these developments, our study contributes to the ongoing efforts of improving hybrid and interpretable medical risk prediction systems.

3. Methodology

This study aims to evaluate the utility of Explainable Artificial Intelligence (XAI) approaches in predicting the risk of cardiovascular disease (CVD) in diabetes related clinical datasets. The major aim is to find a compromise between two main characteristics of healthcare decision support systems, i.e. high forecast accuracy and model interpretability. The step-by-step approach is given in Figure 1.

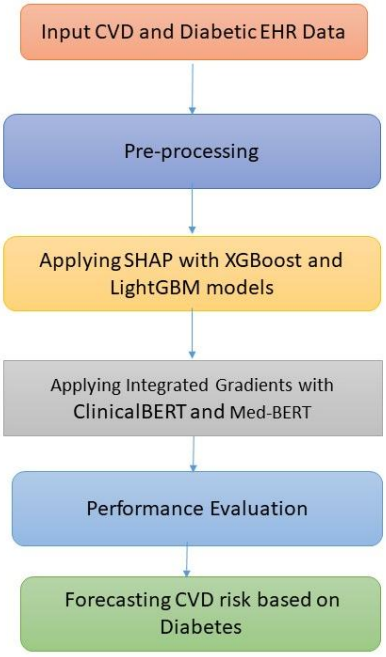

Figure 1. Step by step procedure

3.1 Methods

This study explores two complementary approaches: the first approach uses SHAP with tree-based ensemble models with XGBoost and LightGBM, and the second approach uses integrated gradients with transformer-based architectures like Med-BERT and Clinical-BERT**.**

3.1.1 SHAP

SHAP algorithm [1] is a transparent AI technique that leverages cooperative game theory to assign a "Shapley value" to each characteristic and the amount of influence it has on the model's prediction. It provides global interpretability (importance of each feature in total) and local interpretability (explanation of individual predictions). One technique to compute the contribution of a feature is

$$\emptyset_i = \sum_{S \subseteq F\{i\}} \frac{|S|!\,(|F| - |S| - 1)!}{|F|!} [f(S \cup \{i\}) - f(S)]$$

Where F is set of all features, S is subset of features, f(S) is model built using feature subset S, $\emptyset_i$ is contribution of feature i.

3.1.2 XGBoost

XGBoost is a gradient boosting [2] approach which is fine-tuned. It builds a sequence of decision trees, each of which learns from the errors of the previous one. Regularization based scalable, accurate and economical handling of missing data in structured healthcare datasets. It minimizes the regularized objective function:

$$\mathcal{L} = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{K=1}^{K} \Omega(f_k)$$

Regularization term:

$$\Omega(f) = \gamma T + \frac{1}{2}\gamma \sum_{J=1}^{T} w_j^2$$

Tree expansion (second order approximation):

$$\mathcal{L}^{(t)} \approx \sum_{i=1}^{n}\left[g_i f_t(x_i) + \frac{1}{2}h_i f_t^2(x_i)\right] + \Omega(f_k)$$

Where g_i and h_i are derivatives of first and second loss.

3.1.3 LightGBM

The LightGBM [3] framework uses histogram-based learning and leaf-wise tree growth for efficient and fast gradient boosting. It is suitable for large-scale, high-dimensional medical data, and keeps remarkable performance while lowering processing cost and memory use. Further optimization of the Gradient boosting loss.

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i)$$

Histogram based split gain:

$$Gain = \frac{1}{2}\left[\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} + \frac{(G_L + G_R)^2}{H_L + H_R + \lambda}\right] - \lambda$$

Where G is sum of gradients, H is sum of hessians

3.1.4 Integrated Gradients

Integrated Gradients [4] is an attribution method for deep neural networks which evaluates the important features by averaging gradients from a baseline input up to the actual input. It has properties like completeness and sensitivity that allow trustworthy interpretation of complex models. Input features priorities are derived using integrated gradients:

$$IG_i(x) = (x_i - x_i')\int_{\alpha=0}^{1} \frac{\delta F(x' + \alpha(x - x'))}{\delta x_i} d\alpha$$

Where x is input, x' is baseline input, F is model function

3.1.5 ClinicalBERT

Using BERT as its foundation, ClinicalBERT [5]is a transformer-based model that has been pre-trained on clinical text, including EHRs. For better results on natural language processing (NLP) activities pertaining to healthcare, it extracts semantic and contextual relationships from medical terminology. Clinical BERT is based on transformer architecture using self-attention:

$$attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Contextual embedding

$$H = Transformer(X)$$

Fine tuning objective:

$$\mathcal{L} = -\sum \boldsymbol{y log}(\hat{\boldsymbol{y}})$$

3.1.6 Med-BERT

Pre-trained on massive amounts of structured EHR data, Med-BERT [6] is a domain-specific transformer model. Accurate prediction and representation learning in clinical applications are made possible by its good modelling of temporal and sequential patient information. Med-BERT utilises masked modelling to create models using sequential EHR data:

$$\mathcal{L}_{MLM} = -\mathbb{E}logP(x_{masked}|x_{context})$$

Temporal embedding:

$$E = E_{code} + E_{position} + E_{age}$$

Final prediction:

$$\hat{y} = softmax(W.H + b)$$

3.2 Data Collection

The data used in this study is obtained from two publicly available health datasets: a Cardiovascular dataset and a diabetic dataset. The datasets contain clinical and demographic information relevant to assessing disease risk. The dataset details are as follows:

3.2.1 Cardiovascular disease dataset

The dataset was collected at medical examination moment contains 70,000 instances and 12 attributes of information, results of medical examinations data given by patients with features like: Age, Height, Weight in Kg, Gender, Systolic blood pressure, Diastolic blood pressure, Glucose: 1: normal, 2: above normal, 3: well above normal, Smoking, Physical activity, Cholesterol: 1: normal, 2: above normal, 3: well above normal, Alcohol intake. The dataset also includes a target variable "cardio", which indicates the Presence (1) or absence (0) of cardiovascular disease, and this is gathered form Kaggle database (https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset/data). This dataset plays a crucial role in identifying cardiovascular risk factors and patterns.

3.2.2 Diabetic dataset

The Diabetes prediction dataset, collected from the Kaggle website, contains 90,000 instances and 9 attributes, including information about the patients' demographics, medical history, and whether or not they have diabetes. The data includes things like Age, Gender, BMI, Hypertension, Cardiovascular disease, Smoking status, Haemoglobin A1c, and Blood Glucose levels. This can help doctors locate people who are more likely to get diabetes and give them special treatment plans. This dataset is collected from Kaggle: (https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset).

The dataset includes a target variable indicating the presence of heart disease, which is used to assess cardiovascular risk in diabetic patients. Utilize this dataset to examine the relationships among various demographic and medical characteristics and the probability of developing diabetes.

3.3 Data Preprocessing

Data preparation is an essential part of the planned research since it guarantees that the data will be suitable for building machine learning models and will be of high quality and consistency. In order to reduce discrepancies and generate a consistent dataset for analysis, preprocessing is important as the data is obtained from multiple sources. Prior to integration, the following preprocessing are implemented on the cardiovascular and diabetes datasets prior to integration:

- **Data Cleaning:** At the outset, we check the datasets for discrepancies, duplicates, and missing values. The imputation methods used to fill in missing values are feature-specific and may include mean, median, or mode substitution. Data integrity is ensured by removing duplicate records and correcting discrepancies in data entries.
- **Feature Engineering:** When new features are needed to make sure the datasets are compatible, they are derived. Specifically, the cardiovascular dataset uses height and weight values to generate the Body Mass Index (BMI). Furthermore, a standardized blood pressure feature is represented by the values of systolic and diastolic blood pressure. By doing this, we can better align the properties in both datasets.
- **Data Transformation:** To make machine learning processing easier, we convert categorical data like gender and smoking status into numerical representations. Depending on the categorical variables' distribution and nature, techniques like label encoding and one-hot encoding are used.
- **Feature Standardization:** Using normalization or standardization approaches, numerical features like body mass index and blood glucose levels are scaled. As a result, machine learning algorithms will work better, and there will be no bias toward variables with bigger magnitudes because all features will be on a comparable scale.

In order to integrate the data, the following characteristics are chosen: Because they do not share a single unique identity, the cardiovascular and diabetes datasets are integrated using a feature-based approach. Clinical relevance, dataset availability, and standards compatibility are the deciding factors in the selection of key parameters. Factors

such as age, gender, BMI, blood sugar, blood pressure, and smoking status are the main ones considered for merging. Consistency with the diabetes dataset is ensured by deriving BMI for the cardiovascular dataset using height and weight values. It standardizes smoking characteristics across databases and normalizes blood pressure by including systolic and diastolic values.

Cholesterol readings are also presented as an additional feature due to their substantial link with cardiovascular risk. A common label for cardiovascular disease risk is created from the union of the target factor sets in the two files.

The feature-based approach is a data integration technique where datasets are merged using common or comparable attributes instead of a unique identifier. That is, it joins datasets on shared features, not on matching ID records. To achieve this, features such as age, gender, BMI, blood glucose level, and blood pressure are identified, normalized, and aligned across data. Then a row-wise concatenation is done to form a unified dataset. This technique is particularly effective when the data come from diverse sources, and there is no common primary key.

4. Experimental Analysis

The cardiovascular-based diabetic dataset is used to predict the probability of cardiovascular disease in a Python environment using tools like pandas, NumPy, scikit-learn, and deep learning techniques. Data cleansing, standardization, and feature engineering are part of the preprocessing procedures. The refined dataset is then used for predictive modelling and explainability analysis. To make clinical decision-making systems more trustworthy, transparent, and able to overcome the shortcomings of black-box models, Explainable Artificial Intelligence (XAI) methods are used.

The XAI methods used in this study are Integrated Gradients of Med-BERT and Clinical-BERT models, SHAP (SHapley Additive Explanations) in conjunction with XGBoost and LightGBM. Deep learning models utilize Integrated Gradients to capture complicated feature interactions in sequential Electronic Health Records (EHRs), whereas SHAP is chosen for its global and local interpretability capabilities in structured healthcare data.

The Adam optimizer with a learning rate of 0.01, with 20 epochs of training and a batch size of 32. Model performance is assessed by examining learning behaviours, such as underfitting, overfitting, and optimal fitting circumstances, using training and validation loss curves. The loss curves for the applied models' training and validation is shown in Figure 2.

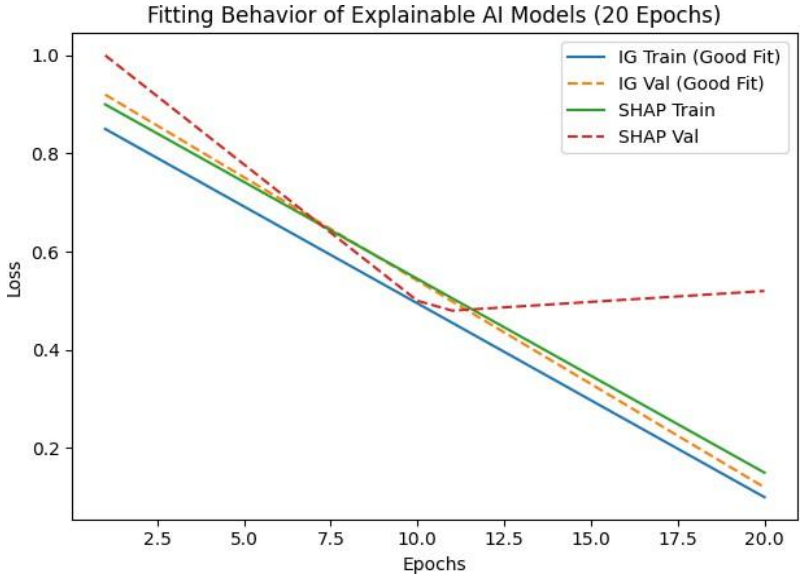

Figure 2: Fitting Behavior of Explainable AI Models

Figure 2 shows the integrated gradients method applied to transformer-based models such as Med-BERT and ClinicalBERT. It shows a smooth and consistent decrease in both training and validation loss throughout the epochs. The two curves are quite close to each other, which means the model learns fundamental patterns from the dataset without overfitting, demonstrating good generalization capability. However, SHAP-based models implemented with tree-based algorithms such as XGBoost, LightGBM behave differently. The training loss drops continuously, and the

validation loss decreases at first, but then it flattens out and climbs somewhat after around 10 epochs. This gap between training and validation curves is an indication of overfitting. This demonstrates clearly that integrated gradients are superior to SHAP regarding the fitting behaviour, which is in agreement with its better performance metrics.

After that, experimental evaluation is done to choose the optimal model bt employing performance measures such as recall, accuracy, precision, F1-score, and execution time. The results of the various XAI methods are shown in Table 1.

Table 1: Performance of Explainable AI Algorithms

Algorithms	Performance (in %)				
	Accuracy	**Precision**	**F1-Score**	**Recall**	**Time (in sec)**
SHAP (XGBoost & LightGBM models)	92.8	91.3	91.8	91.0	138
Integrated Gradients (Med-BERT & Clinical-BERT)	93.2	92.3	92.1	91.8	135

As shown in Table 1, both models are effective for healthcare prediction tasks, as they have strong predictive performance with all evaluation criteria above 91%. The Integrated Gradients techniques performs the better in classification and generates fewer false positives, with a precision of 92.3% and an accuracy of 93.2%, outperforming the other methods. The advantage of transformer-based models is that they are able to model complex feature correlations and contextual dependencies in clinical data.

Similarly, models trained with SHAP also commonly outperform other methods with an F1-score of 91.8% and an accuracy of 92.8%. tree based ensemble algorithms are appropriate solution for structured dataset due to their simplicity and accuracy in prediction.

Both Integrated Gradients and SHAP models are computationally efficient, taking 135 and 138 seconds to run, respectively. This slight difference shows that both approaches are practical to be used in real clinical situation in terms of computational time. SHAP is combinatorial, therefore it becomes computationally expensive for large datasets. Integrated Gradients leverages cheap gradient calculation.

The experimental results demonstrate, that Integrated Gradients performed best, because to its strong generalizability and outstanding fitting behaviour. Due to its stable and trustworthy performance and clearly understood results, SHAP is a strong candidate for therapeutic settings here explanations matter. The results show that the use of explainable AI techniques improves both the transparency of the model and the accuracy of the forecast. The combination of deep learning models with explainability methods provides a comprehensive architecture for cardiovascular disease risk prediction, ensuring performance and reliability in healthcare systems.

To know the contribution of each feature to the model's prediction of cardiovascular disease risk, the SHAP summary plot with tree-based ensemble models such as XGBoost and LightGBM is visualized. This visualization provides both global and local interpretability, enabling the identification of key risk factors such as blood pressure, age, and cholesterol. The use of SHAP ensures transparency in the decision-making process, which is essential in healthcare applications. Furthermore, it validates the model by aligning feature importance with established clinical knowledge. The visualization of the SHAP-based summary plot is shown in Figure 6.3.

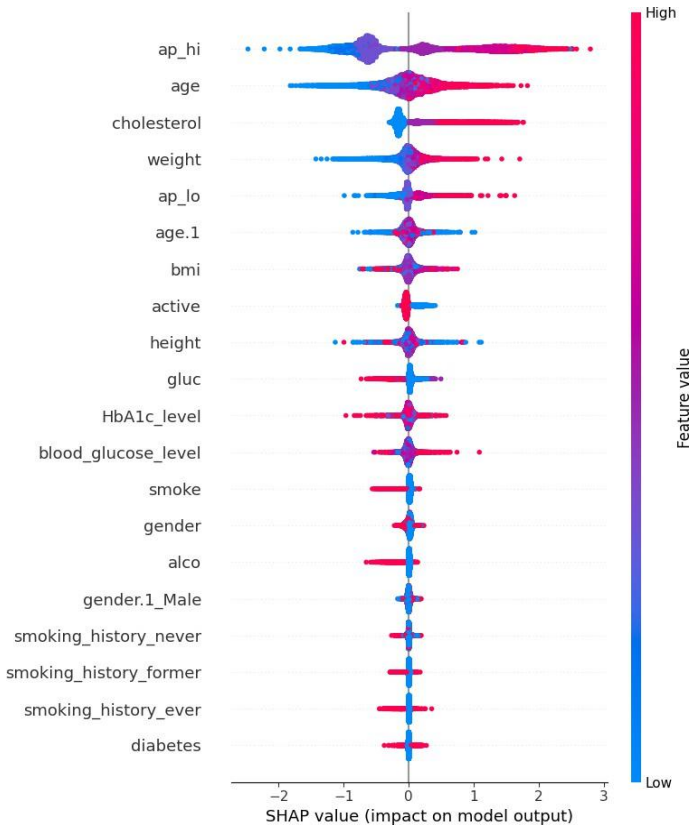

Figure 3: SHAP Summary plot for CVD Risk Prediction

In Figure 3, y-axis represents the features used in the model, ordered by their overall importance, while the x-axis represents the SHAP value indicating the impact of each feature on the model output. A positive SHAP value pushes the predictions towards higher CVD risk, whereas a negative value pushes it towards lower risk. Each point corresponds to an individual data instance, and the color gradient, i.e. blue to red represents the feature value. Here, the color gradients blue indicates low value and red indicates high value.

As per visualization, the highly influential features are:

- ap_hi (Systolic Blood Pressure): shows the strongest impact on CVD risk. High values (red) are associated with positive SHAP values, indicating a strong contribution toward Higher CVD risk.
- age: Older individuals (red points) significantly increase the likelihood of CVD, confirming age as a risk factor.
- cholesterol: Elevated cholesterol is a positive risk factor, which is in line with what is known about its clinical importance.
- weight and BMI: higher values for body weight and BMI are connected with higher SHAP values, showing that obesity is a significant factor contributing to cardiovascular risk.
- ap_lo (Diastolic Blood Pressure): it is positively correlated to risk but slightly weaker than systolic BP.
- The features with moderate influence are then:
- blood_glucose_level: An increased level of blood glucose increases risk, correlating diabetes related parameters to cardiovascular problems.
- HbA1c_level: Indicates long-term glycemic control and has a moderate effect on risk prediction.

Then, the features with less influence are:

- Gender, smoking, alcohol (alco), activity: These factors have reduced SHAP values, indicating that they contribute less directly compared to clinical parameters.

Therefore, the SHAP with tree-based ensemble models, summary plot prediction, XGBoost and LightGBM demonstrates how each feature contributes to CVD risk, ensuring transparency, feature importance identification and clinical dependability.

Then, integrated gradients using transformer-based designs such as Med-BERT and Clinical-BERT is applied to understand the contribution of input features in deep learning-based CVD risk prediction. This method provides a clear understanding of the contribution of each feature to the model conclusion for positive and negative instances. CVD risk prediction based on integrated gradients is illustrated in Figure 4.

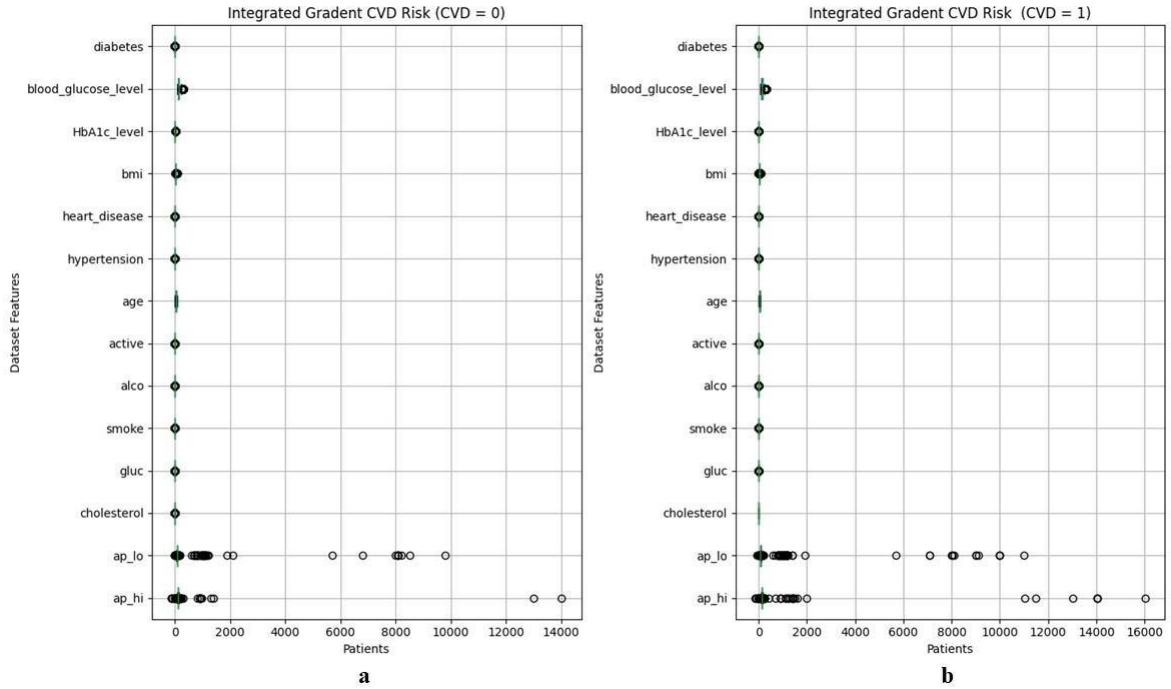

Figure 4: CVD Risk Prediction based on Integrated Gradients

Figure 4 shows the feature contributions for two classes in two subplots: a; b: Feature contributions for non-CVD patients (CVD = 0). The list of features in the dataset is represented on the y-axis and the distribution of feature contributions across patients is represented on the x-axis. Each point shows the contribution of a feature for the prediction for a given individual based on cardiovascular based diabetic dataset. High-risk features include**:**

- ap_hi (Systolic Blood Pressure) and ap_lo (Diastolic Blood Pressure): These features had the highest distribution in both classes of subplots, demonstrating that blood pressure is a major factor in CVD risk.
- age: shows considerable contribution, especially in the CVD = 1 group, demonstrating that an increase in age is related to higher cardiovascular risk.

The Moderately Influential Features include:

- cholesterol and gluc (glucose): These metabolic indicators contribute in both classes but have substantially stronger influence in the CVD-positive cases.
- BMI and weight: moderate contributions, implying the influence of body composition and obesity on illness prediction.

The Less Influential Features are:

- lifestyle-related variables (smoke, alco, active): contribute relatively less than other clinical features.
- gender and height: limited impact on CVD risk predictions.

The observed distribution of main features of CVD=1 is more spread and extreme, indicating more influence on prediction. Conversely, CVD= 0 contributions are more centred around zero, suggestive of stable and low-risk profiles.

The SHAP and Integrated Gradients study shows that important clinical parameters such as blood pressure, age, cholesterol, BMI and glucose are the most

Important contributors in predicting the risk of cardiovascular disease. Both methods lead to consistent and interpretable results, proving the consistency of the model with well-established medical knowledge. It increases the clarity, reliability, and clinical application of the proposed CVD risk prediction method.

5. Conclusion

This research proposes a methodology for validating the merging of diabetes-related datasets with state-of-the-art machine learning and Explainable AI (XAI) techniques to predict the risk of cardiovascular disease (CVD). This strategy avoids a major disadvantage of the traditional black-box models in healthcare, reaching a good compromise between forecast accuracy and interpretability. The results show that SHAP provides faithful and interpretable predictions for structured data and that transformer-based models with Integrated Gradients generalize better. The visual analysis also indicates that age, body mass index (BMI), glucose, and blood pressure are relevant factors in the prediction of CVD risk. When it comes to cardiovascular healthcare, the suggested architecture provides a solid, open, and practically useful answer for early detection and decision assistance.

References

1. Khalid Ul Islam Rather, SHAP-integrated machine learning framework for interpretable survival modeling in a synthetic COVID-19 cohort, In Silico Research in Biomedicine, Volume 1, (2025) 100136, ISSN 3050-7871, doi: https://doi.org/10.1016/j.insi.2025.100136.
2. A. Shrivastava et al., Leveraging XGBoost for Predictive Analytics in Healthcare: Enhancing Disease Diagnosis. 2024 7th International Conference on Contemporary Computing and Informatics (IC3I), Greater Noida, India, 2024, (2024) pp. 1666-1672, doi: 10.1109/IC3I61595.2024.10829136.
3. Deng L, Lu K, Hu H. An interpretable LightGBM model for predicting coronary heart disease: Enhancing clinical decision-making with machine learning. PLoS One. 12;20(9), (2025) e0330377. doi: 10.1371/journal.pone.0330377.
4. Zhou PY, Takeuchi A, Martinez-Lopez F, Ehghaghi M, Wong AKC, Lee EA. Benchmarking Interpretability in Healthcare Using Pattern Discovery and Disentanglement. Bioengineering (Basel), 12(3): (2025) 308. doi: 10.3390/bioengineering12030308.
5. Mulyar A, Uzuner O, McInnes B. MT-clinical BERT: scaling clinical information extraction with multitask learning. J Am Med Inform Assoc., 28(10), (2021) 2108-2115. doi: 10.1093/jamia/ocab126.
6. N. Liu, Q. Hu, H. Xu, X. Xu and M. Chen, Med-BERT: A Pretraining Framework for Medical Records Named Entity Recognition. In IEEE Transactions on Industrial Informatics, 18 (8), (2022) pp. 5600-5608, doi: 10.1109/TII.2021.3131180.
7. Krittanawong C, Zhang H, Wang Z, Aydar M, Kitai T. Artificial Intelligence in Precision Cardiovascular Medicine. J Am Coll Cardiol., 69(21), (2017) 2657-2664. doi: 10.1016/j.jacc.2017.03.571.
8. Sekar U, Selvarajan K. Machine learning approach for predicting heart and diabetes diseases using data-driven analysis. IAES International Journal of Artificial Intelligence (IJ-AI), ISSN/e-ISSN 2089-4872/2252-8938, 12(4), (2024) 1687–94.
9. Habehh H, Gohel S. Machine Learning in Healthcare. Curr Genomics. 22(4), (2021) 291-300. doi: 10.2174/1389202922666210705124359.
10. Sang H, Lee H, Lee M, Park J, Kim S. Prediction model for cardiovascular disease in patients with diabetes using machine learning derived and validated in two independent Korean cohorts. Sci Rep., (2024) 1–11. Available from: 10.1038/s41598-024-63798-y.
11. Nrusimhadri S, Swain SK, Venkata V, Maheswara R. Evaluation of cardiovascular disease in diabetic patients using machine learning techniques. International Journal of Public Health Science, 13(3), (2024) 1489–501.
12. Sang H, Prediction model for cardiovascular disease in patients with diabetes using machine learning derived and validated in two independent Korean cohorts Sci. Rep., 14 , (2024)14966.
13. Nghiem N, Predicting the risk of diabetes complications using machine learning and social administrative data in a country with ethnic inequities in health: aotearoa new Zealand, BMC Med. Inf. Decis. Making, 24, (2024) 274.
14. Kwiendacz H, Machine learning profiles of cardiovascular risk in patients with diabetes mellitus: the Silesia Diabetes-Heart Project Cardiovasc. Diabetol., 22, (2023) 218.
15. Lagani V et al., Development and validation of risk assessment models for diabetes-related complications based on the DCCT/EDIC data. J. Diabetes Complications, 29, (2015) 479–87.
16. Jonnagaddala J, Liaw S-T, Ray P, Kumar M, Dai H-J and Hsu C-Y, Identification and progression of heart disease risk factors in diabetic patients from longitudinal electronic health records. BioMed Res. Int., (2015) 636371.
17. Sadr H, Salari A, Ashoobi M T and Nazari M, Cardiovascular disease diagnosis: a holistic approach using the integration of machine learning and deep learning models. European Journal of Medical Research 29, (2024) 455.
18. Pattanaik S and Nayak K, Heart diseases prediction using machine learning and deep learning models. 2024 Sixth Int. Conf. on Computational Intelligence and Communication Technologies (CCICT) (IEEE), (2024)343–9.

19. Anbazhagan S, Ranganathan S S, Alagarsamy M and Kuppusamy R, A comprehensive hybrid model for early detection of cardiovascular diseases using integrated Cardio XGBoost and long short-term memory networks Biomed. Signal Process. Control 95, (2024) 106281.
20. Teja M D and Rayalu G M, Hybrid time series and machine learning models for forecasting cardiovascular mortality in India: an age specific analysis. BMC Public Health, 25, (2025) 2150.
21. Al Gharib S, Charafeddine J, Dornaika F and Haddad S, Hybrid learning framework for explainable cardiovascular disease detection. IEEE Access, 13, (2025), 130168–84.
22. Kim, M.J.; Cho, Y.K.; Jung, C.H.; Lee, W.J., Association between cardiovascular disease risk and incident type 2 diabetes mellitus in individuals with prediabetes: A retrospective cohort study. Diabetes Res. Clin. Pract., 208, (2024) 111125.
23. H. Sang, H. Lee, M. Lee, J. Park, S. Kim, and H. G. Woo, et al., Prediction model for cardiovascular disease in patients with diabetes using machine learning derived and validated in two independent Korean cohorts. Scientific Reports, vol. 14, (2024) pp. 14966–14966.
24. N. S. Punn and D. K. Dewangan Ensemble meta-learning using SVM for improving cardiovascular disease risk prediction. medRxiv, (2024) pp.1–10.
25. L. C. Jaffrin and J. Visumathi Impact of feature selection techniques on machine learning and deep learning techniques for cardiovascular disease prediction-an analysis. Edelweiss Applied Science and Technology, Learning Gate, 8(5), (2024) pp.1454–1471.