Explainability of the predictions of Machine and Deep learning models for the Heart health

Dr. B Bharghavi [1] and M.V. Sri Lasya [2]

[1] Gitam Deemed University, Hyderabad
[2] Gitam Deemed University, Hyderabad

Abstarct:

Machine learning and deep learning models have become increasingly prevalent in the field of healthcare, particularly in the prediction of heart health outcomes. However, the inherent complexity of these models can make them difficult to interpret and explain to end-users, such as clinicians. This research paper explores the importance of model explainability in the context of heart health prediction, examines the current state of explainable artificial intelligence (XAI) techniques, and discusses the challenges and potential solutions for improving the transparency and interpret-ability of these models.

Keywords: Machine learning, Deep Learning,Explainability.

1 Introduction :

The use of machine learning and deep learning algorithms in the healthcare industry has grown exponentially in recent years, driven by the availability of large datasets and the potential to improve diagnostic accuracy, clinical decision-making, and patient outcomes. One area where these models have shown particular promise is in the prediction of heart health, where they can be used to identify individuals at risk of developing cardiovascular diseases, such as coronary artery disease, heart failure, and arrhythmias.

But it is quite important and imperitive to understand and interpret the resultrs of the machine learning and deep learing models for better understanding and interpreting the models and thereby draw necessary insights so that the diagnosis improves.

2 Literature Survey :

In the recent years, the application of machine learning and deep learning in health care has seen significant advancements. These modes have been utilized to predict various health outcomes, including cardiovascular diseases,due to their ability to handle and analyze large and complex datasets. However, the black-box nature of these models can pose a challenge for the interpretability and trust, especially in the critical fields like healthcare.

Explainable AI (XAI) Techniques:

Explainable AI aims to make the decisions of ML and DL models understandable to humans. Various XAI techniques have been proposed, which can be broadly categorized into model-specific and model-agnostic methods:

1. Model-Specific Methods: These methods are tailored for specific types of models. For example:

- Decision Trees: Inherently interpretable as they provide a clear decision path.
- Rule-Based Models: Provide human-readable rules derived from data.
- Attention Mechanisms in Neural Networks: Highlight important features or inputs that influence the model's decisions.

2. Model-Agnostic Methods: These methods can be applied to any model, such as:

- LIME (Local Interpretable Model-agnostic Explanations): Generates local approximations of the model to explain individual predictions.
- SHAP (SHapley Additive exPlanations): Uses game theory to attribute the contribution of each feature to the model's prediction.
- Feature Importance Analysis: Measures the impact of each feature on the model's output.
- DiCE (Diverse Counterfactual Explanations): Gives the ‘what-if’ predictions of the model’s by tweaking the feature values to a certain extent to get the expected outcome.

Several machine learning techniques have been explored for predicting heart health, each with varying degrees of success and interpretability:

1. Traditional Machine Learning Models:.

- Logistic Regression: Commonly used for binary classification problems, offering a balance between performance and interpretability

- Decision Trees: Provide a visual representation of decision rules, making them easy to interpret.
- Support Vector Machines (SVMs): Effective for high-dimensional data but less interpretable.

2. Ensemble methods:

- Random Forests: Combine multiple decision trees to improve accuracy and robustness but can be harder to interpret.
- AdaBoost: Enhances the performance of weak learners but increases complexity.
- Gradient Boosting Machines (GBMs): Powerful predictors but often considered black-box models.
- Histogram Gradient Boosting Machines : It is more faster edition of the Gradient Boosting machines but can be considered as a black box model.

3. Deep Learning Models:

Multilayer Perceptrons (MLPs): Used for their ability to capture complex patterns in data, though they are less interpretable.

Challenges and Potential Solutions.

The main challenges in implementing XAI techniques for heart health prediction include

1. Trade-off Between Accuracy and Interpretability: More complex models often provide better performance but at the cost of interpretability.
2. Data Quality and Representation: Ensuring the data used is representative and of high quality is crucial for both model accuracy and interpretability.
3. Integration with Clinical Workflow: XAI tools must be seamlessly integrated into clinical workflows to be useful to practitioners
4. Potential solutions to these challenges involve developing hybrid models that balance performance and interpretability, improving data preprocessing techniques,and creating user-friendly interfaces for XAI tools.

3 Methodology:

The data is collected from kaggle, one of the best sources of datasets for machine learning and deep learning tasks were collected. For the link to the dataset, kindly refer [1] in the References.

3.1 Data Acquisition and Preprocessing

The dataset of dimensions 70,000 rows and 12 columns. The features were all numeric in nature and the data is encoded. The target feature ‘cardio’ is in the binary format where ‘1’ indicates that there is a heart related condition and ‘0’ otherwise. Some

features like systolic pressure and diastolic pressure, smoke and alcohol were also considered.

Data preprocessing: The initial phase of Exploratory Data Analysis has shown that the data is having some missing values and outliers were present in two records. The missing values were dropped and the outliers were imputed using the quartiles of the individual records.

The size of the dataset came down to 64,502 records.**The data has been scaled using the MinMaxScaler as we do not want our models to give us skewed results towards the values having high numerica value. So all the features are brought to the same scale.**

We tried to improve the model's performance by employing one of the most important feature transformation techniques : Principal Component Analysis. The untransformed and the transformed data were kept in separate variables to understand and pick up the better performing model.

Model Construction : The task was the supervised machine learning one which is classification. The classifiers considered are Decision Tree, Logistic Regression and Support Vector Machine.

The ensembling models like Random Forest, AdaBoost and Histogram- Gradient Boosting trees were also considered. A mulitlayer perceptron is also constructed for this dataset as it is more complicated compared to the machine learning models and explainability will make more sense for such black box models.

The simpler models tend to overfit to the training data and their results can sometimes be unreliable as they do not have greater level of accuracy. The complex models , often have greater level of generalization and produce greater level of accuracy but they are quite difficult to understand and interpret.

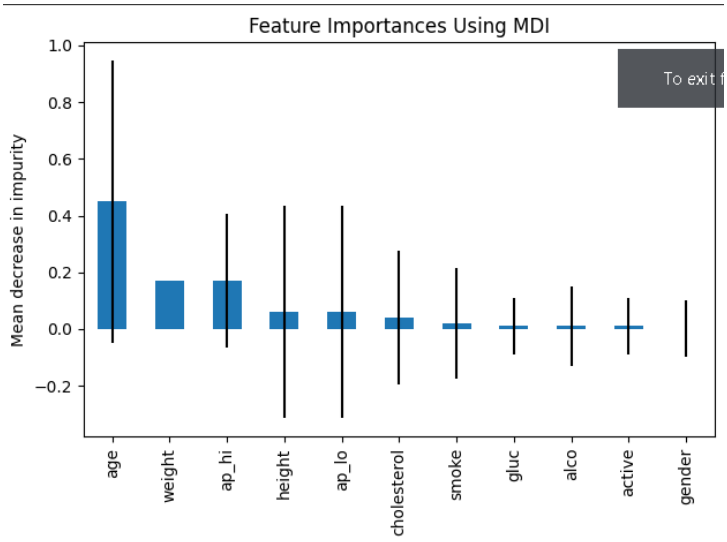
The results of the training process and their respective accuracies can be found in the following table:

MODEL	TRAINING ACCURACY	TESTING ACCURACY
DECISION TREE	99.98	62.20
LOGISTIC REGRESSION	72.52	71.47
SUPPORT VECTOR MACHINE	72.33	71.23
RANDOM FOREST	99.98	69.90
ADAPTIVE BOOSTING	99.98	71.92
HISTOGRAM BASED GRADIENT BOOSTING MACHINE	74.15	72.13
MUTLILAYER PERCEPTRON	72.92	72.49

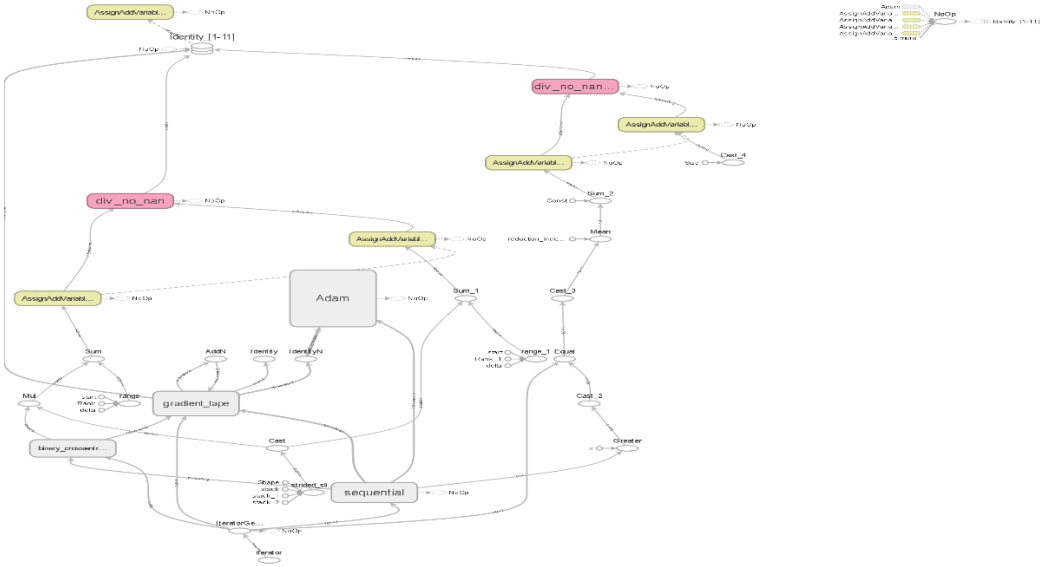
The results of the training and testinp process of the models' with the principal component analysis is as follows:

MODEL	TRAINING ACCURACY	TESTING ACCURACY
DECISION TREE	99.99	62.53
LOSISTIC REGRESSION	72.32	71.96
SUPPORT VECTOR MACHINE	72.13	72.01
RANDOM FOREST	99.98	68.08
ADAPTIVE BOOSTING	72.85	72.09
HISTOGRAM BASED GRADIENT BOOSTING MACHINE	73.98	72.63

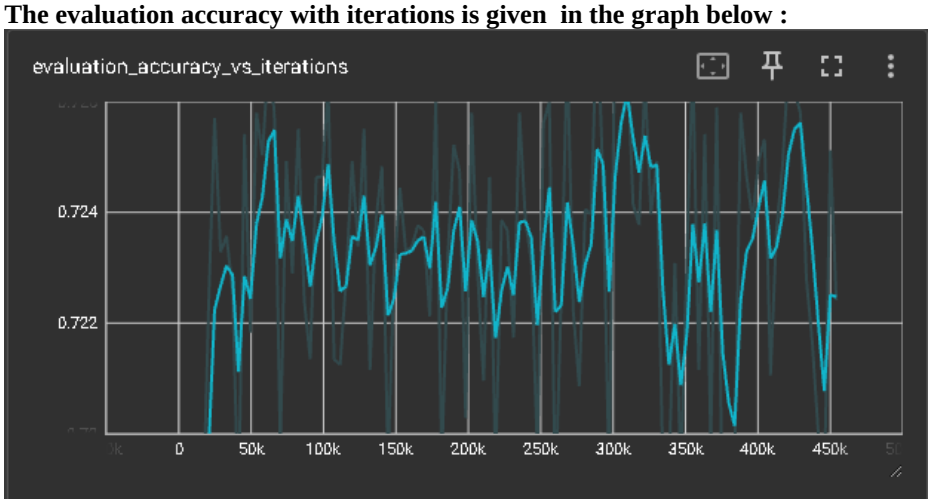
Initially feature importance was calculated from the AdaBoost classifier and the plot looked like this :



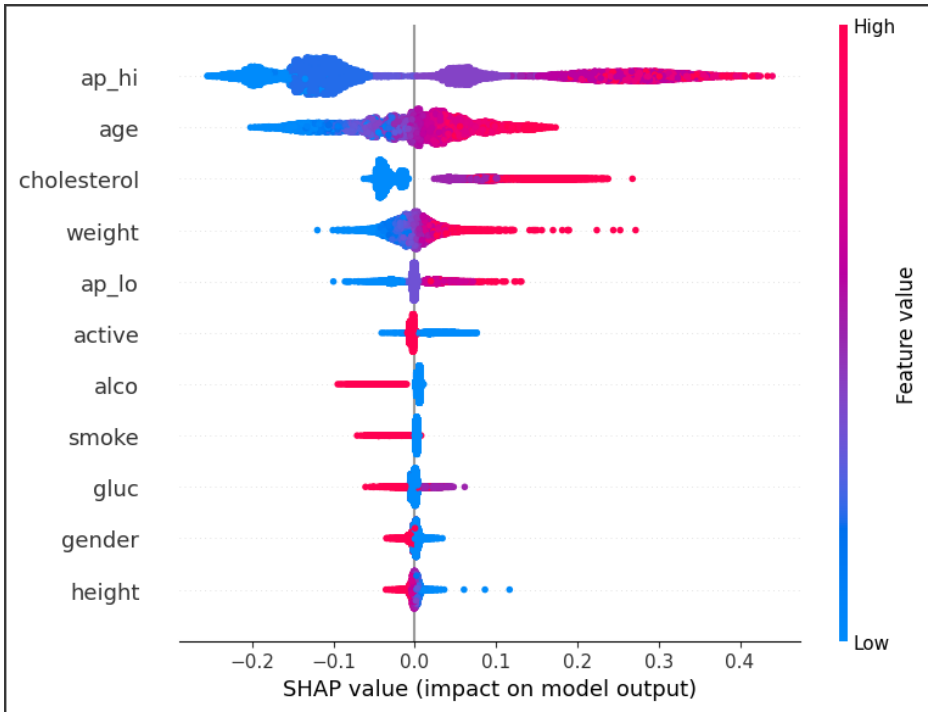
The operation level graph of the tensorflow neural network from the tensorboard looked as follows:



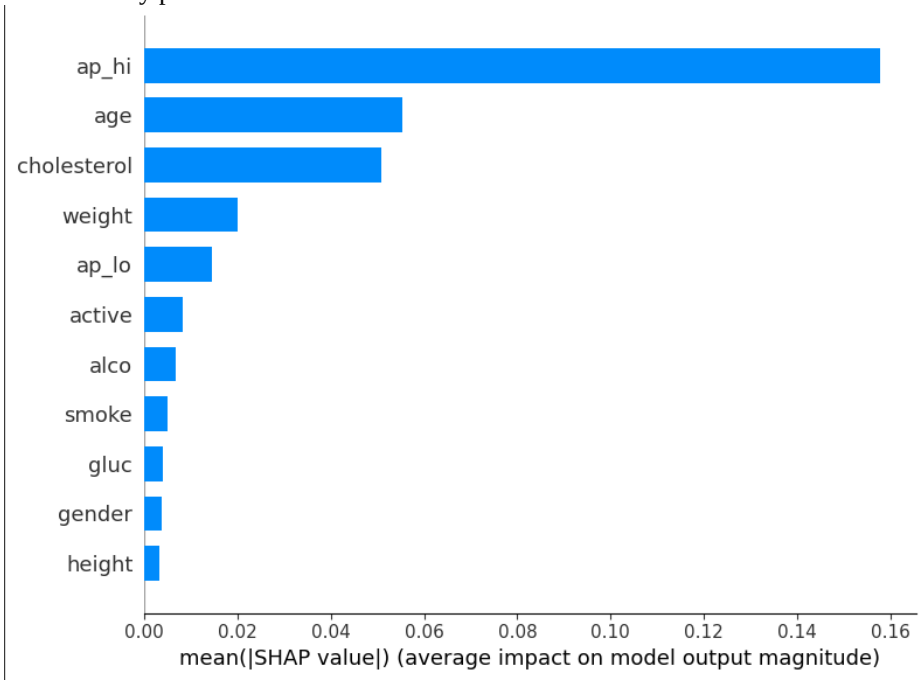
The evaluation accuracy with iterations is given in the graph below :



3.2 SHAP(SHapley Additive Predictions) : SHAP is a model agnostic explanation technique used to understand the model's interpretability. The model used with SHAP to interpret is the multilayer perceptron. The outputs looked as follows :



The summary plot in the bar format is :



3.3 **DiCE(Diverse Counterfactual Explanations) :** Diverse counterfactual explanations is a powerful model agnostic library which gives us the posible counterfactuals in a record to get the desired outcome. For example: one of the records in the dataset looked as follows :

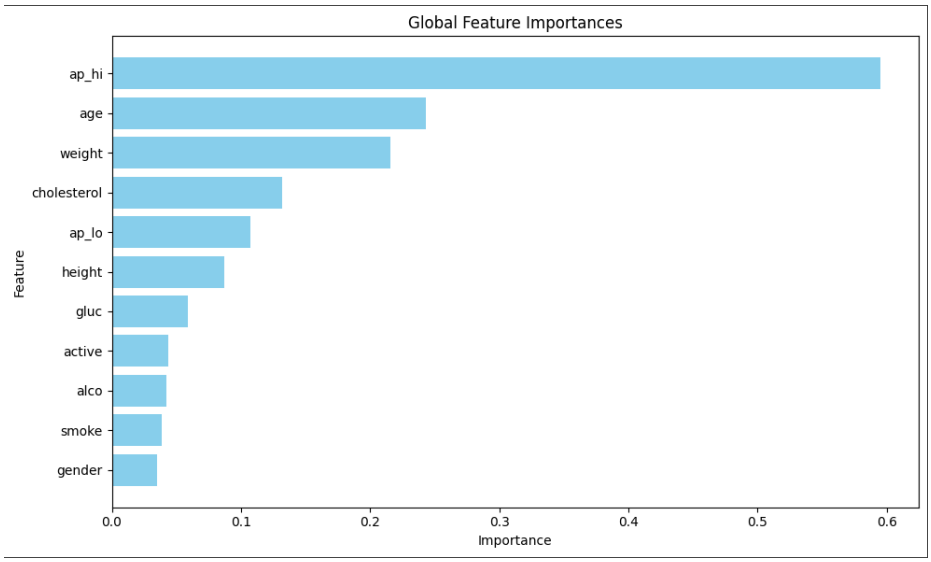
	age	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke	alco	active	cardio
0	0.688657	0.0	0.517949	0.306878	0.25	0.25	0.5	0.0	0.0	0.0	1.0	0

The outcome of DICE outcome looked as follows :

Diverse Counterfactual set (new outcome: 1)

	age	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke	alco	active	cardio
0	-	-	-	-	-	0.948	-	-	-	-	-	1.0
1	-	-	-	-	0.726	-	0.0	-	-	-	-	1.0
2	-	-	-	-	0.749	-	-	-	-	-	-	1.0
3	-	-	-	-	0.384	-	-	-	-	-	-	1.0
4	0.8248965	-	-	-	-	-	-	1.0	-	-	-	1.0

DiCE can be used to calculate the feature importance like the local and global feature importance. For a sample of 100 records, global feature importance was calculated and a bar plot is plotted. The results are turned out to be as follows :



4 Experimentation results :

The analysis of the data ended up explaining the fact that there is a huge importance to the feature ‘ap_hi’ and ‘age’ in the predictions. The detailed explanation can be :

- ‘ap_hi’ is the feature which stands for the systolic blood pressure. This is a feature of high importance and it could be one of the main reasons for having a heart related issue.
- ‘age’ can be an influential factor in having heart related issues and the risk of having a cardiac issue is high when you are older.
- ‘Cholesterol’ and ‘weight’ are two features with almost equal importance and if not taken care properly can be detrimental for the heart.

Conclusion :

The results and the model’s predictions are clearly indicating that the chance of having a cardiac issue is going to go down if healthy lifestyle is maintained.

The explainability of the black-box models is a task of importance. It is of great importance for us to work on the concept of explainability and come up with better tools and techniques to reduce the tradeoff between the performance and interpretability.

References :

1. Link to the dataset :
https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset

2. Effective Heart health prediction using Machine Learning techniques, Cihintan M.Bhatt, Parth Patel, Tarang Ghetia, Pier Luigi Mazzeo.

3. A LIME- Based Explainable Machine Learning Model for predicting the severity level of COVID-19,Freddy Gabbay, Shirly Bar-Lev, Ofer Montano and Noam Hadad.

4. Leveraging Knowledge Graphs for Enhancing Machine Learning based Heart Health Prediction , Majlinda Llugiqi, Fajar J. Ekaputra, Marta Sabou.

5. Understanding Deep Learning (Still) Requires Rethinking Generalization, Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht and Oriol Vinyals.

6. A Survey on the Explainability of the supervised machine learning, Nadia Burkart, Marco F. Huber.

7. CFDB : Machine Learning model analysis via Databases of Counterfactuals, Idan Meyuhas, Aviv Ben Arie, Yari Horesh, Daniel Deutch.

8. Layer-wise relevance propagation for deep neural network architectures, Sebastian Bach, Gregoire Montavon, Klaus- Robert Muller,Wojciech Samek.

9. Explainable AI : A review of Machine Learning Interpretability Methods, Pantelis Linardatos, Vasilis Papastefanopoulos and Sotiris Kotsaintis.

10. Explaining machine learning classifiers through diverse counterfactual explanations, by Ramaravind K.Mothilal, Amit Sharma and Chenhao Tan.