[1]Krishna Modi
[2]Ishbir Singh
[3]Yogesh Kumar

Optimizing Diabetes and Heart Disease Prediction through Machine Learning Algorithms incorporating Lifestyle Factors

Journal of Electrical Systems

Abstract: **-** Lifestyle diseases are becoming significant global public health concern. These diseases include hypertension, diabetes, heart diseases, asthma, obesity etc. This paper explores the use of ML models to predict lifestyle diseases, focusing on diabetes and heart disease. We utilized publicly available datasets—PIMA Diabetes and Cleveland Heart Disease to develop eight distinct ML models: k Nearest Neighbors (kNN), Logistic Regression (LR) , Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), Deep Neural Network (DNN), ADABoost, and XGBoost. Our approach emphasizes data preprocessing techniques to ensure high-quality input for model training, including the handling of missing values, and standard scaling for normalization. The importance of each feature was assessed using the ANOVA F-value method. We used stratified sampling for splitting dataset to maintain equal class distribution. Our findings indicate that DNN and XGBoost achieved the highest predictive performance on the PIMA Diabetes dataset, with recall values of 0.89 and 0.92, respectively along with AUC scores of 0.836 and 0.83, respectively. For the Cleveland Heart Disease dataset, AdaBoost emerged as the most reliable model, demonstrating a precision of 0.85, a recall of 0.909, and a high AUC of 0.924. Overall, this research highlights the potential of ML techniques in improving the early detection of lifestyle diseases, while also addressing the challenges of dataset quality and model interpretability.

Keywords: Lifestyle factors, Diabetes prediction, Heart disease prediction, Ensemble learning, Deep Neural Network (DNN), Lifestyle diseases.

I. INTRODUCTION

Lifestyle diseases are becoming a major public health concern globally. Lifestyle diseases are result of personal behavior and environment. Symptoms of these diseases are developed due to improper food intake, physical inactivity, smoking, alcohol consumption or stress. Few common lifestyle diseases are hypertension, diabetes, heart diseases, obesity, high blood pressure etc.

The rise of lifestyle diseases is interrelated to modernization. With advancement of lifestyle, people are shifting from traditional and physically demanding lifestyle to more comfortable living. This also includes increased consumption of processed food, high sugar, high salt, unhealthy fats, and exposure to pollution and toxins. The increase of new technologies and automated systems have undoubtedly revolutionized various aspects of daily life and reduced physical work in numerous fields. While these advancements brought convenience and efficiency, they have also inadvertently contributed to a lifestyle, which resulted in rise of lifestyle-related diseases like hyper-tension, diabetes, heart diseases, obesity and high blood pressure. Furthermore, the pressure of modern life contributes to mental health issues, which can turn in to physical health problems.

A large portion of population globally is affected with lifestyle diseases. These diseases play a substantial role in global mortality rates as well. Addressing lifestyle diseases requires a multiple approaches. Awareness programs, educating people about healthy living, encouraging physical activities and healthy diet can reduce the burden of these diseases and can improve overall public health. On the other hand, early prediction and detection are very effective preventive measures. By utilizing ML and AI, a person can have personalized health assistance and assessment. By accurately predicting the onset of these diseases, individuals and healthcare professionals alike can implement timely and appropriate preventive measures, ultimately working towards the betterment of public health and well-being worldwide.

Current studies demonstrate that ML techniques are playing significant role in the field of medical research, and are offering the capacity to improve disease prediction and diagnosis.

In this paper, we have presented a study on the total eight ML algorithms for the prediction of lifestyle diseases. Our aim is to develop accurate prediction models for a range of lifestyle diseases. Our research in this paper

[1]*Department of CSE, Indus Institute of Technology and Engineering, Indus University, Ahmedabad 382115, India

(E) : krishnamodi1994@gmail.com

[2] Department of ME, Indus Institute of Technology and Engineering, Indus University, Ahmedabad 382115, India

[3] Department of CSE, School of Technology, Pandit Deendayal Energy University, Gandhinagar, Gujarat, India

focuses on two major lifestyle diseases: diabetes, and heart disease. We have assessed the performance of our study in measures of precision, recall, Accuracy, AUC and F1-score.

The objectives of this research are to develop advanced models, which can identify a person at possibility of lifestyle diseases; and provide healthcare practitioners, particularly physicians, with a reliable and efficient tool for disease detection and prediction. By integrating AI-based solutions into medical practice, we aim to streamline healthcare processes and improve patient outcomes.

A. Selection of Articles

We have searched relevant articles on online databases PubMed, IEEE Xplore, and Scopus. Keywords and phrases used in the searching includes "lifestyle disease detection", "early prediction of diabetes", "heart disease", "machine learning in medical diagnosis", "deep learning in healthcare", "classification models for healthcare" and so on. The search was restricted to journal articles and conference proceedings. We have selected papers employing AI and ML techniques, including but not limited to deep learning, transfer learning and ensemble methods. Articles published within the last five years were given priority to ensure the inclusion of recent advancements in the field.

B. Organization of the Article

Following the introduction, paper is structured as follows: Section 2 highlights a comprehensive literature review focusing in the domain of lifestyle disease prediction and AI-based approaches. Literature review provides insights into current advancements, existing methodologies and gaps in the research. Section 3 presents the methodology used in our research, which includes dataset collection, feature extraction, data preprocessing, and data splitting. It also outlines various models we have implemented. Section 4 gives overview of performance parameters used for the evaluation of these models, ensuring robust assessment of their effectiveness. In Section 5, we compare the results of various models and discussed the performance of our proposed model. Lastly, Section 6 summarizes the paper with key findings, highlighting significance of our research and its potential implications for future studies in the field.

II. RELATED WORK

The rise of lifestyle diseases has become a major global health concern. Traditional diagnostic methods often involve time-consuming procedures. Consequently, many researchers are now utilizing ML and AI to develop effective and accurate diagnostic tools. Early prediction and diagnosis of lifestyle diseases play a crucial role in contemporary medical research, with existing models encompassing statistical methods and various ML techniques.

Approximately 53.7 Crore people aged 20 to 79 were living with diabetes in 2021. This number is estimated to rise to 64.3 Crore by 2030 and 78.3 Crore by 2045. It is estimated that diabetes caused 67 lac deaths in 2021 [6]. Alarmingly, about half of those living with diabetes remain unaware of their condition, highlighting the critical need for early detection and the identification of at-risk individuals to provide necessary support.

Singh et al. (2023) enhanced the performance of diabetes diagnosis and prediction through the implementation of ML approaches, specifically ANN and deep learning algorithms [17]. Their proposed framework improves classification results, contributing to early detection and better cure of diabetes. In their study, Wee et al. (2023) emphasized the importance of feature selection and dimensionality reduction in predictive models. They compared the accuracy of LR, SVM, RF, and ANN, both with all features and after feature selection, highlighting the significance of high-quality data in achieving optimal model performance [19].

Chang et al. (2022) classified diabetes using various ML algorithms, aiming to develop effective models capable of accurately identifying individuals based on relevant health indicators [4]. Yasar (2021) highlighted a feature selection process using swarm-based algorithms, demonstrating how optimized feature subsets improve classification accuracy [20]. Tanim et al. (2024) proposed DeepNetX2, a deep neural network integrated with Explainable AI techniques to enhance interpretability while maintaining high accuracy in diabetes diagnosis [18]. This paper also discusses the interrelation between diabetes and heart diseases. Mathukiya et al. (2024) provided a comparison of multiple ML techniques in diabetes detection, concluding that Random Forest performed exceptionally well, yielding high accuracy [12].

Heart disease continues to be a major cause of death globally. Begum et al. (2024) outlined an IoT-based system using deep learning to predict heart disease in real time by examining sensor data available from wearable devices

[2]. Several studies [8], [11], [13] have demonstrated the efficiency of ML techniques such as LR, kNN and RF, in processing vast amounts of medical data to provide accurate predictions of heart disease. Ahmed et al. (2023) reported that the RF model significantly outperformed traditional methods in predicting heart disease risk factors [1]. Similarly, Bhatt et al. (2023) highlighted the effectiveness of kNN in handling large datasets, achieving high accuracy in their predictions [3]. Jebamalar et al. (2024) [7] emphasized the robustness of LR in classifying patients with varying risk levels, while Kavitha et al. (2021) [9] demonstrated that a combination of these ML techniques could further improve prediction accuracy. Together, these studies underscore the potential of ML approaches in progressing early detection and diagnosis of heart disease.

III. METHODOLOGY

After analyzing and studying variety of papers on diseases prediction, we came up with a unique framework that can be applied on diseases dataset, which involves following key steps: data collection, data processing, feature selection or extraction, data splitting, and model validation as demonstrated in Fig. 1. Table 1 outlines the details of datasets used for this research. After collecting proper dataset, process starts with data processing, in which data visualization and missing value treatment is performed. Rather than training algorithm on dataset directly, here we have used ANOVA feature selection method and Chi Square based test to find relevancy of features to the diseases. After this, data is partitioned into training and testing sets as illustrated in Table 4. Data is trained on different eight algorithms. These algorithms hyper parameters are set as described in Table 3. The final stage involves performance evaluation, where the model's effectiveness is assessed across multiple metrics, providing comprehensive insights into its predictive capabilities.

A. Data Description

Comprehensive analysis of lifestyle diseases, covering four distinct types was conducted in this study. We have included a total of four datasets that encompass lifestyle and health-related factors, which facilitates research into the diagnosis of lifestyle diseases.

The Cleveland Heart Disease dataset available at UCI repository [5] comprises of 13 features, including clinical and diagnostic attributes, to predict the presence of heart disease in 303 individuals. Kaggle's PIMA Diabetes dataset [15] provides nine features related to diabetes risk factors, assisting in the classification of diabetes presence indicated by the binary variable among 768 individuals. The summary of these datasets is given in Table1. Table 2 and Table 3 represent the statistical analysis of continuous attributes for PIMA diabetes and Cleveland Heart Diseases respectively. We have also employed violin plots and boxplots to visualize distribution of continuous features with respect to the target attributes, which are illustrated in Fig. 2 and Fig. 3 for PIMA diabetes and Cleveland Heart Diseases respectively.

B. Data Preprocessing

Quality data is essential for training any algorithm effectively, and data preprocessing plays a key role in generating such quality data. Data preprocessing typically involves three critical phases. In the first phase, handling missing values, we address issues such as incomplete or missing data points. In the PIMA dataset, some numeric attributes such as Skin Thickness (Skin), Body Mass Index (BMI), and Glucose (Plas) contain values of 0 in a few records, indicating missing data. We have replaced these missing values with the mean value of the respective attribute to ensure the model is trained effectively. For categorical (string) data, records with missing values are deleted to maintain data integrity and quality. In second phase, to speed up the algorithm's calculations and normalize the data within a specific range, standard scaling (Also known as z-score normalization) is performed.

Table 1 Dataset description

Dataset No	Disease	Source	Total Features	Target Attribute	Number of Samples
1.	PIMA Diabetes	Kaggle [15]	9	Class (Binary)	768
2.	Cleveland Heart Disease	UCI Repository [5]	13	Target – Heart Disease(Binary)	303

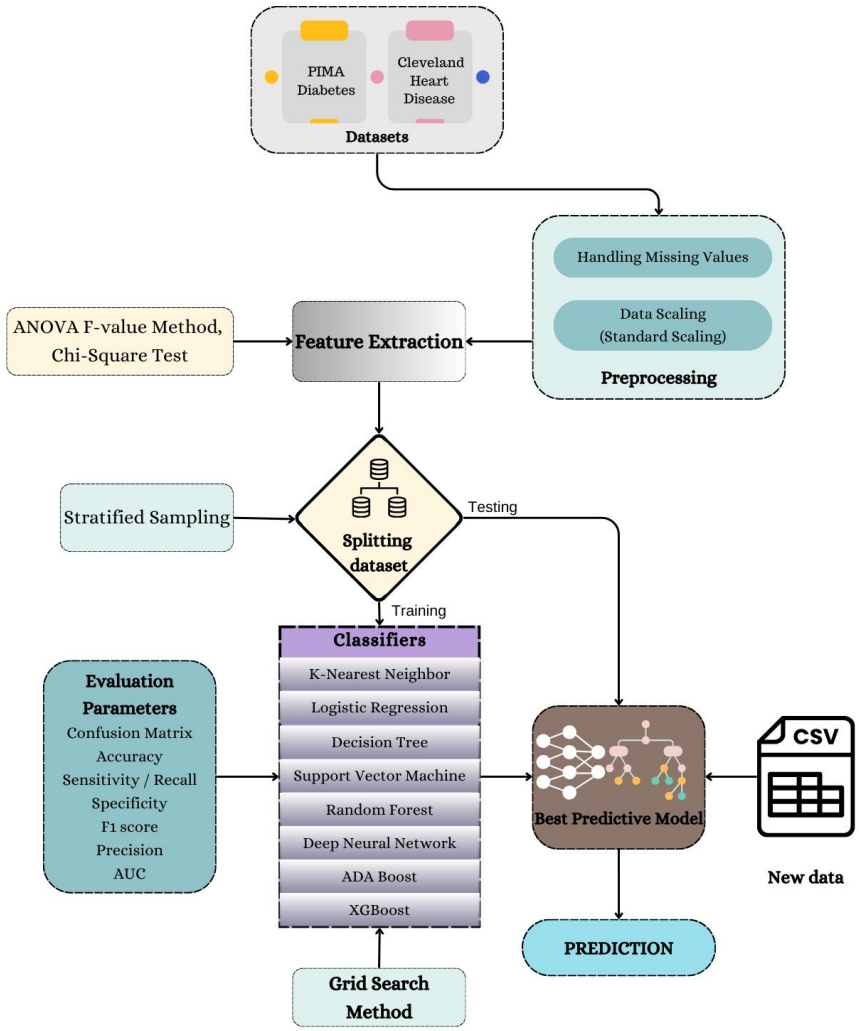


Fig. 1 System Design for Prediction

Table 2 Statistical analysis of attributes in PIMA diabetes

	class	preg	plas	pres	skin	test	mass	pedi	age
Mean	0	3.3	110	68.2	19.7	68.8	30.3	0.43	31.2
	1	4.87	141	70.8	22.2	100	35.1	0.55	37.1
Median	0	2	107	70	21	39	30.1	0.336	27
	1	4	140	74	27	0	34.3	0.449	36
Standard deviation	0	3.02	26.1	18.1	14.9	98.9	7.69	0.299	11.7
	1	3.74	31.9	21.5	17.7	139	7.26	0.372	11
Minimum	0	0	0	0	0	0	0	0.078	21
	1	0	0	0	0	0	0	0.088	21
Maximum	0	13	197	122	60	744	57.3	2.33	81
	1	17	199	114	99	846	67.1	2.42	70

Table 3 Statistical analysis of Cleveland heart disease on continuous attributes

	target	age	trestbps	chol	cp	thalach	oldpeak
Mean	0	56.6	134	251	0.478	139	1.59
	1	52.5	129	242	1.38	158	0.583
Median	0	58	130	249	0	142	1.4
	1	52	130	234	2	161	0.2
Standard deviation	0	7.96	18.7	49.5	0.906	22.6	1.3
	1	9.55	16.2	53.6	0.952	19.2	0.781
Minimum	0	35	100	131	0	71	0
	1	29	94	126	0	96	0
Maximum	0	77	200	409	3	195	6.2
	1	76	180	564	3	202	4.2

C. Feature Extraction

For building successful predictive model, correlation of each attribute with the target attribute is crucial. It enables us to determine the importance of each attribute, which is key to achieving accurate predictions. We have utilized ANOVA F-value method, specifically 'f_classif' class from 'sklearn.feature_selection' in Python, along with 'SelectKBest' class for feature selection. Alternatively chi square method can also be used with 'SelectKBest' class to select important feature. Feature selection also helps in reducing the dimensionality. This approach provides us with a high level of confidence in our analysis, ensuring that we only include relevant features in our predictive model. Co-relation of each attribute is displayed in Fig. 4 for each dataset.

D. Data splitting

We have used Stratified sampling for categorical datasets to split data so that data is evenly distributed as per the labels. We have used 4:1 ratio for training and testing data.

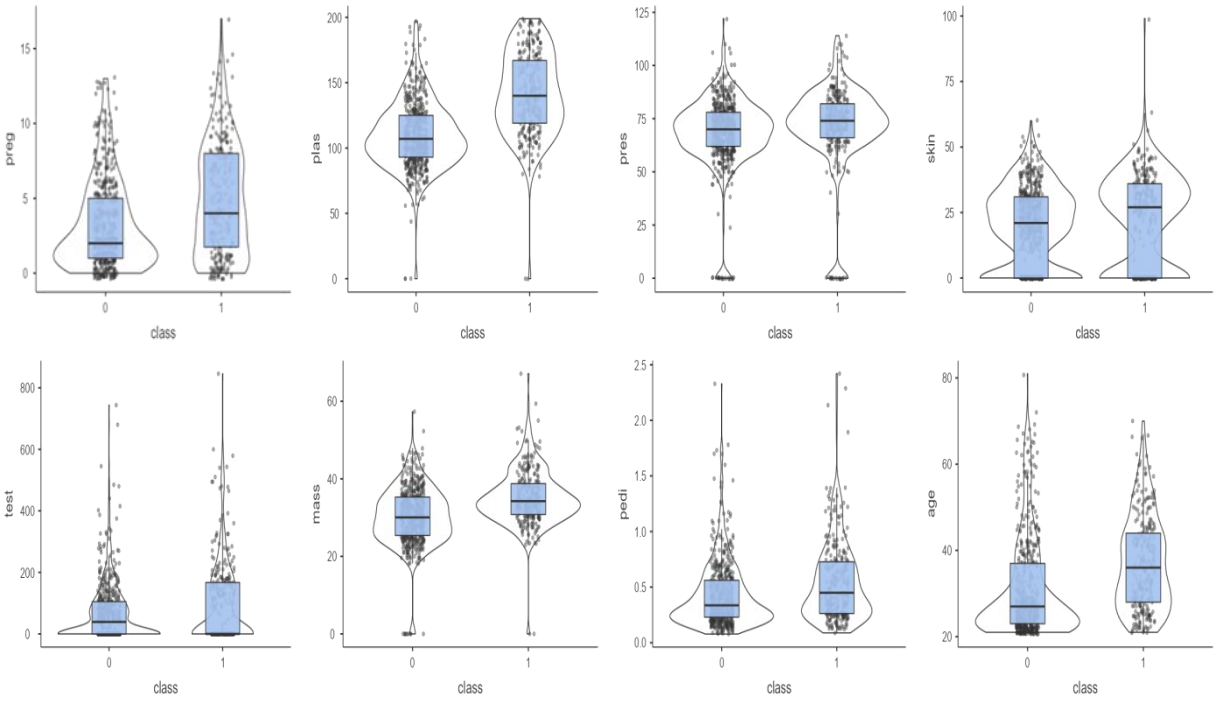


Fig. 2 Box plots and violin plots with respect to target attribute of PIMA diabetes

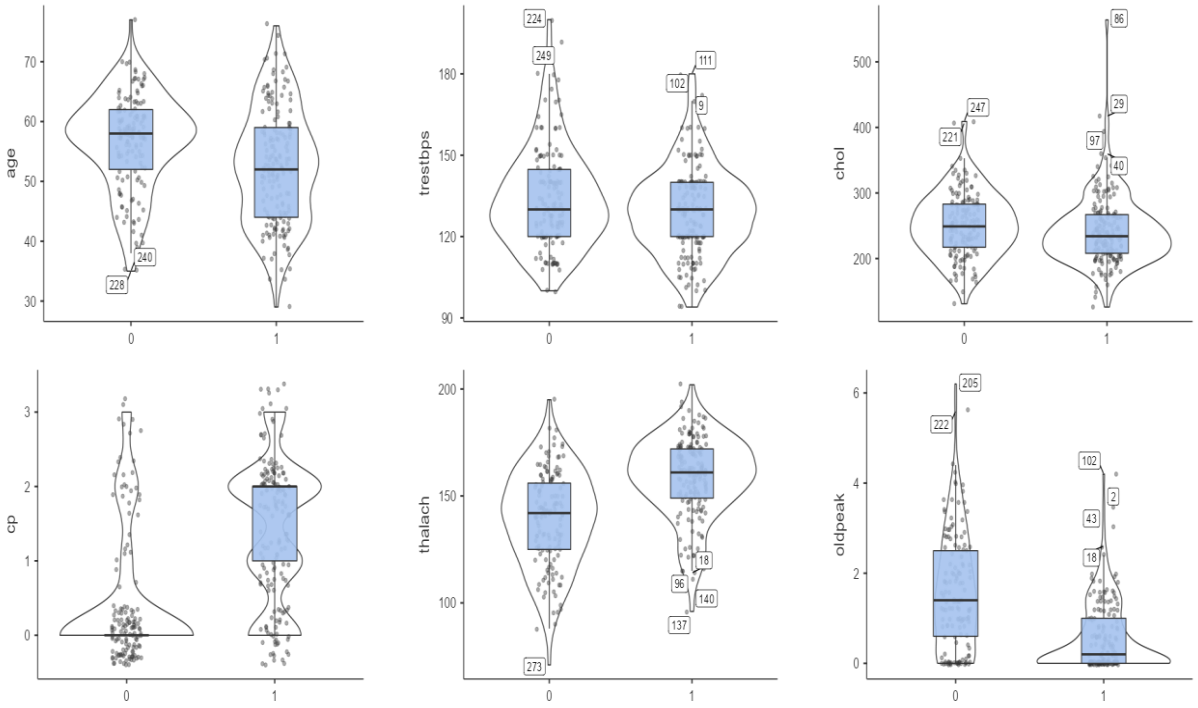


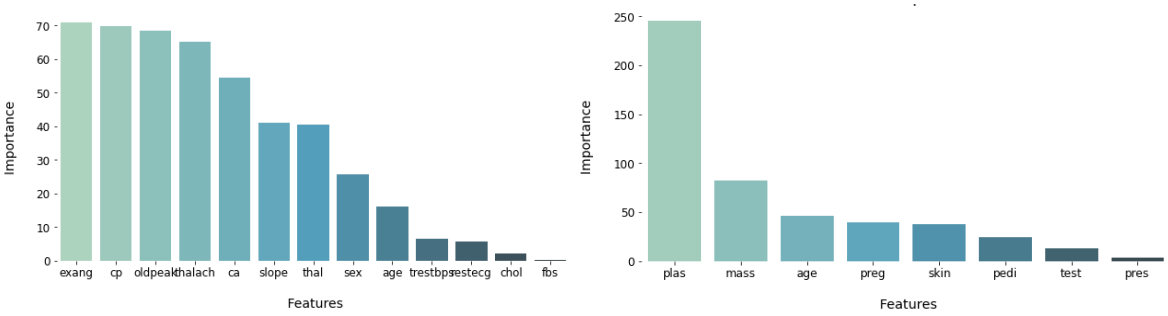
Fig. 3 Box plots and violin plots of continuous features with respect to target attribute of Cleveland Heart Disease

Logistic Regression finds the probability that a given input x belongs to a particular class. Probability of given input for particular class is given by (2) and decision boundary of logistic regression is determined by (3).

$$P(y = 1 \mid x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_d x_d)}} \quad (2)$$

where β0,β1,…,βd are the parameters to be estimated.

$$\hat{y} = \begin{cases} 1 & \text{if } P(y = 1|x) \geq 0.5 \\ 0 & \text{otherwise} \end{cases} \quad (3)$$



Cleveland Heart Disease PIMA Diabetes

Fig. 4 Co-relation of each attribute with target

In Decision Tree, internal node contains the test for decision and each branch represents the outcome of the test. Coming to the end, leaf node finally represents the class. This model is very easy to interpret. The decision rule for a node is based on a measure of Gini impurity. Gini impurity for a binary classification problem is given by:

$$\text{Gini}(D) = 1 - \sum_{i=1}^{C} (p_i)^2 \quad (4)$$

where pi is the proportion of samples that belong to class i in dataset D.

Random Forest uses bagging method. It constructs a number of decision trees during the training. Each tree is trained on a random subset of the data. When new data needs to be classified, each tree makes a prediction, and

final decision is decided by voting among all trees. The prediction of a Random Forest for classification is given in (5).

$$\hat{y} = \text{mode}\{\hat{y}_1, \hat{y}_2, \dots, \hat{y}_n\} \quad (5)$$

where $\hat{y}_n$ is the prediction of the nth decision tree.

Support Vector Machine constructs hyperplane(s) in a high-dimensional space to separate different classes. For a linear SVM, (6) is used for the decision function and class is predicted with (7).

$$f(x) = \beta_0 + \sum_{i=1}^{n} \alpha_i K(x_i, x) \quad (6)$$

Where, $K(x_i, x)$ is the kernel function, and α_i are the model parameters obtained from training.

$$\hat{y} = \begin{cases} 1 & \text{if } f(x) \geq 0 \\ -1 & \text{otherwise} \end{cases} \quad (7)$$

Deep Neural Network has multiple layers of neurons, with each neuron applying activation function to its input neurons. For a feed forward DNN with L layers, the output of the lth layer $a^{(l)}$ is given by (8).

$$a^{(l)} = \sigma(W^{(l)} a^{(l-1)} + b^{(l)}) \quad (8)$$

Here, $W^{(l)}$ represents weights of lth layer, $b^{(l)}$ represents biases of lth layer, and σ is the activation function.

ADABoost is combines the predictions of weak classifiers and create a strong classifier. XGBoost is a boosting method that uses tree models. It builds trees sequentially. Each tree corrects the error that is observed in the previous ones. The prediction function for XGBoost is shown in (9).

$$\hat{y} = \sum_{k=1}^{K} f_k(x) \quad (9)$$

Where f_k is a function in the space of regression trees.

E. Hyper parameter tuning

For machine learning classification tasks, several algorithms with distinct hyper parameters are employed. Grid search method is used to tune hyper parameters. The K-Nearest Neighbors (KNN) algorithm utilizes various values of K (the number of neighbors) ranging from 2 to 7, using Euclidean distance for proximity calculations, with a leaf size of 30, implemented through sklearn.neighbors. Logistic Regression applies L2 regularization with a regularization parameter (C=1) and uses the lbfgs optimization algorithm, available in sklearn.linear_model. The Decision Tree classifier, using the Gini impurity as the splitting criterion, is experimented with max depths from 3 to 6, implemented via sklearn.tree. Random Forest models, part of sklearn.ensemble, grow 100 to 175 trees, with Gini impurity used for splitting. Support Vector Machines (SVMs) employ multiple kernels such as linear, RBF, polynomial, and sigmoid with C=1, handled by sklearn.svm. For deep learning, a Deep Neural Network (DNN) with varying iterations (100 to 250) is used, adapting the hidden layer size to each dataset. It applies the Adam optimizer with a learning rate of 0.001 and an alpha of 0.001, implemented through sklearn.neural_network. ADABoost leverages decision trees (maximum depth of 4) as base estimators for multiclass classification with the SAMME boosting algorithm, testing up to 125 estimators and a learning rate of 1.0 (sklearn.ensemble). XGBoost, a powerful gradient boosting library (xgboost), uses up to 175 estimators, a maximum depth of 3, and a learning rate of 0.1. For multiclass problems like obesity datasets, the softmax objective function is applied, while binary classification tasks use the logistic function. These algorithms, tuned with appropriate hyperparameters, form a robust ensemble for classification tasks. All experiments were conducted using Python 3.11 on the Spyder IDE 5.4.3, within the Anaconda environment.

IV. PERFORMANCE PARAMETERS

Data models applied to lifestyle disease detection have demonstrated encouraging results. However, relying solely on accuracy is insufficient. To ensure the effectiveness of these models, it is crucial to assess their performance using suitable evaluation metrics. The following key metrics are important for gauging their execution.

A. Accuracy

Accuracy is a fraction of correctly predicted instances to the total instances. This parameter does not work well if dataset in imbalanced.

$$\text{Accuracy} = \frac{\text{Total Correct Predictions}}{\text{All Predictions}} = \frac{N_{TN} + N_{TP}}{N_{TN} + N_{TP} + N_{FP} + N_{FN}}$$

B. Precision

Precision is ratio of True Positive instances to total instances predicted as Positive. Precision is particularly used when dataset is imbalanced.

$$\text{Precision} = \frac{\text{Actual Predicted Positive}}{\text{All Predicted Positive}} = \frac{N_{TP}}{N_{TP}+N_{FP}}$$

C. Recall

Recall is ratio of True Positive instances to the total positive instances. For imbalanced data recall is also used to measure performance of model.

$$\text{Recall} = \frac{\text{Actual Predicted Positive}}{\text{All Predicted Positive}} = \frac{N_{TP}}{N_{TP}+N_{FP}}$$

D. F1 Score

F1 Score is the harmonic mean of precision and recall.

$$\text{F1 Score} = \frac{2}{\frac{1}{\text{precision}}+\frac{1}{\text{recall}}}$$

E. ROC Curve (Receiver Operating Characteristics)

ROC curve is used to plots True Positive Rate (sensitivity) and False Positive Rate at various threshold values. Value of Area Under ROC curve represents the model's performance. Higher value represents better performance.

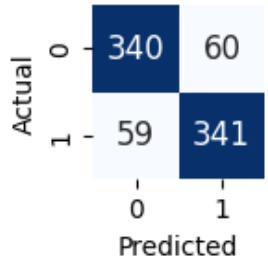
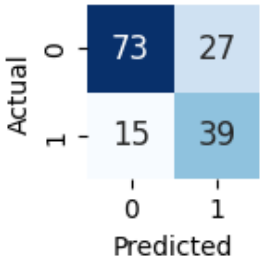
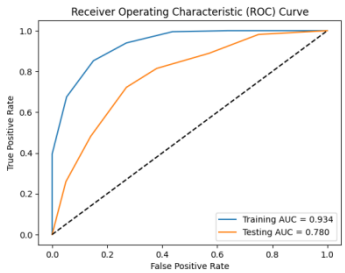
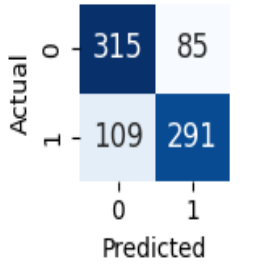
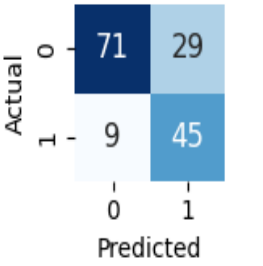
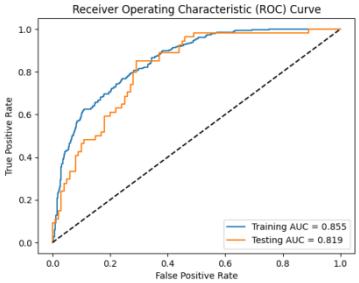
V. RESULTS

The results highlight the predictive performance of various machine learning models on PIMA Diabetes, and Heart Disease. We evaluated the models based on key metrics such as precision, recall, F1 score, accuracy, and AUC, considering both the training and testing phases.

Table 5 and Table 6 contain the results of the models and Fig. 5 offers a graphical representation of these results for PIMA Diabetes. For Heart diseases dataset, results and graphical analysis is included in Table 7, Table 8 and Fig. 6.

These results clearly states that we are getting better predictive results for models using ensemble methods, including Random Forest, XGBoost and ADABoost models. As ensemble methods are averaging the prediction of multiple models, the risk of overfitting, bias and variance are reduces. In addition to ensemble methods, DNN has also demonstrated its strong predictive capabilities.

Table 4 Confusion Metrics and ROC Curve for PIMA Diabetes

Algorithm	Confusion Metrics		ROC curve
	Training	Testing	
kNN			
LR			

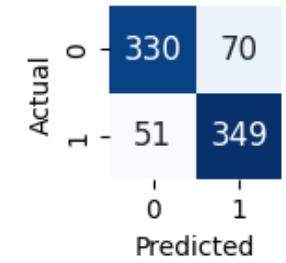
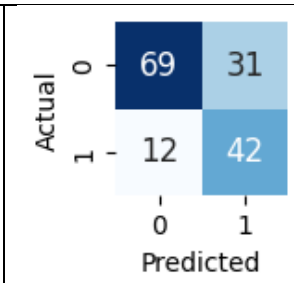
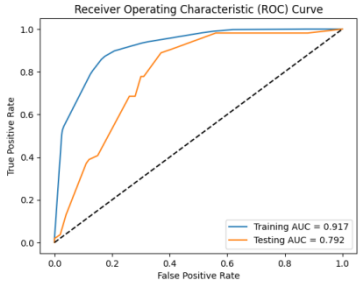
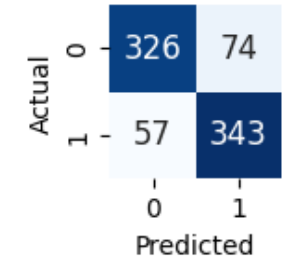
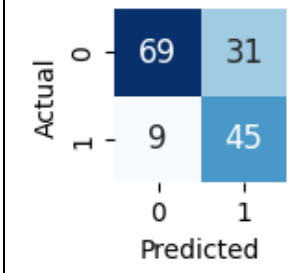
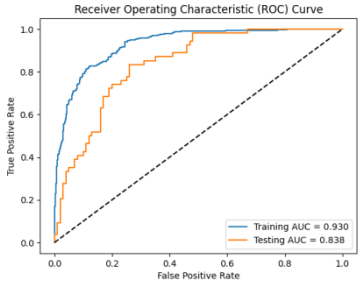
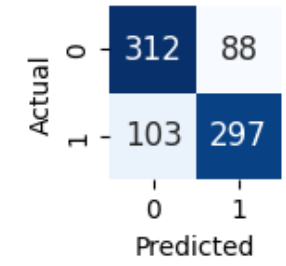
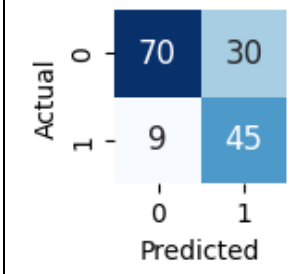
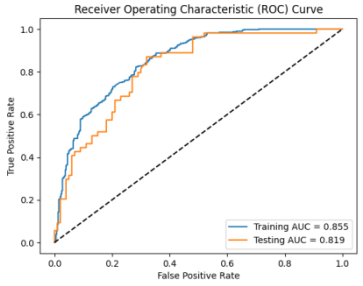
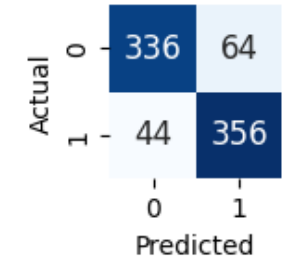
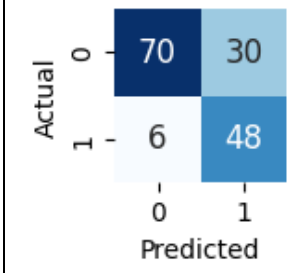
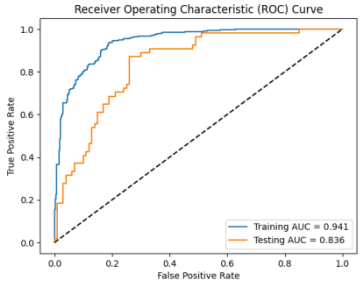
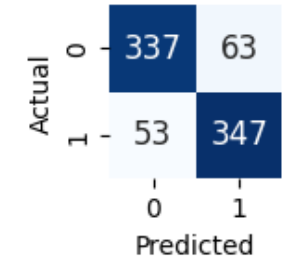
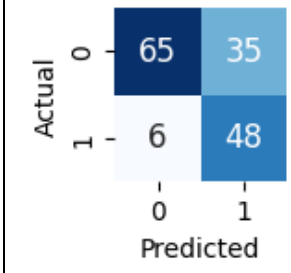
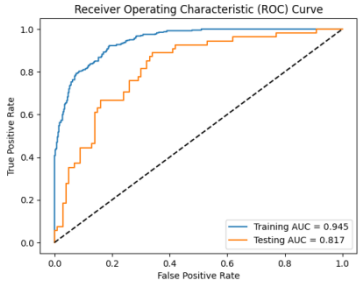
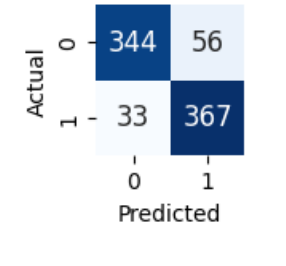
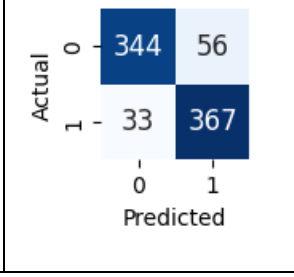
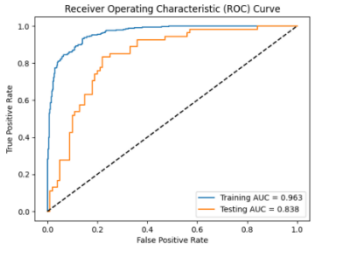
DT			
RF			
SVM			
DNN			
ADABOOST			
XGBOOST			

Table 5 Evaluation of models during training and testing phases for PIMA Diabetes

Model	Training					Testing				
	Precision	Recall	F1 score	Accuracy	AUC	Precision	Recall	F1 score	Accuracy	AUC
kNN	0.85	0.85	0.85	0.85	0.934	0.59	0.72	0.65	0.72	0.78
LR	0.77	0.73	0.75	0.76	0.85	0.61	0.83	0.7	0.75	0.82
Decision Tree	0.83	0.87	0.85	0.85	0.917	0.57	0.78	0.66	0.72	0.792
RF	0.82	0.86	0.84	0.84	0.93	0.59	0.83	0.69	0.74	0.838
SVM	0.77	0.74	0.756	0.76	0.855	0.6	0.83	0.697	0.746	0.819
DNN	0.84	0.89	0.868	0.865	0.941	0.61	0.89	0.72	0.766	0.836
ADABOOST	0.84	0.86	0.856	0.855	0.945	0.57	0.89	0.7	0.73	0.817
XGBOOST	0.87	0.92	0.89	0.89	0.96	0.58	0.92	0.71	0.74	0.83

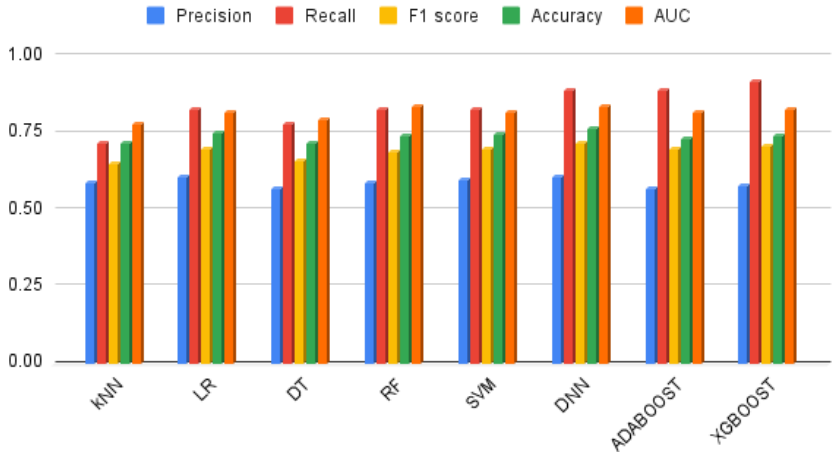
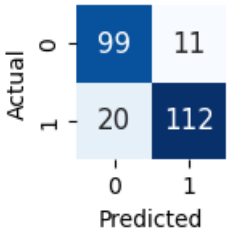
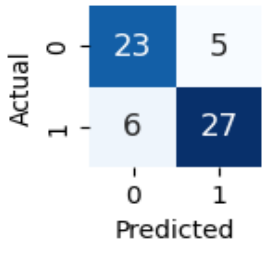
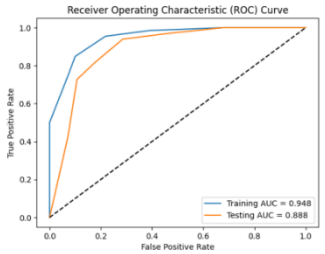
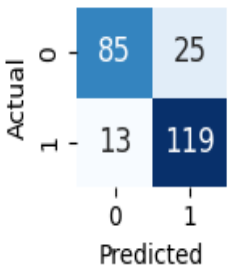
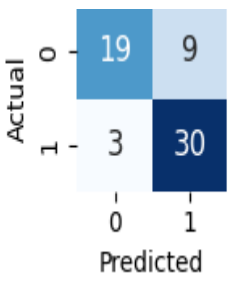
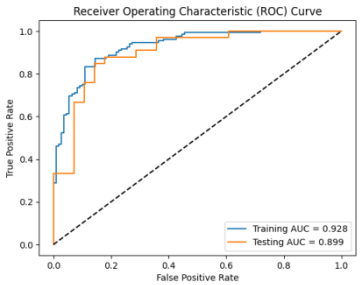
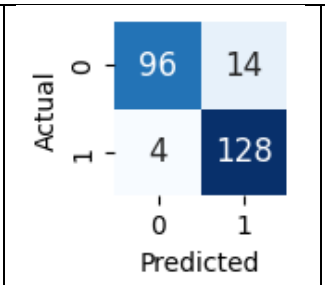
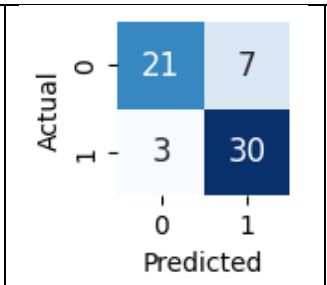
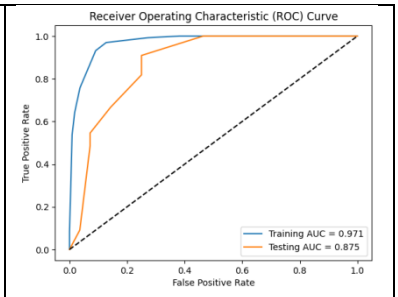
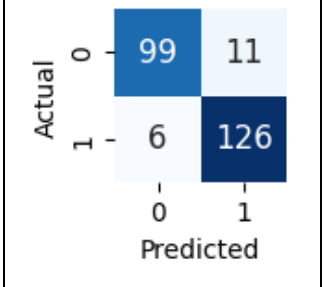
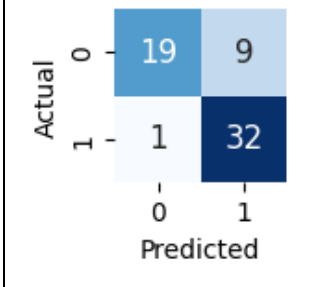
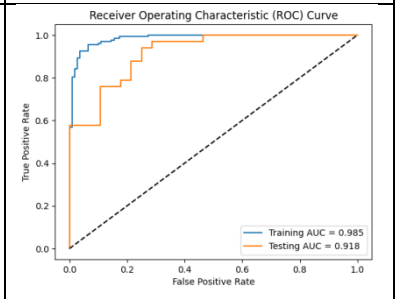
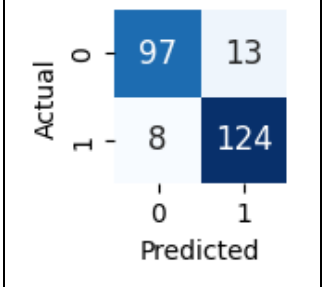
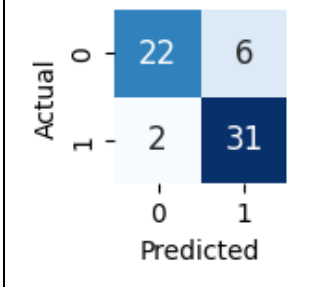
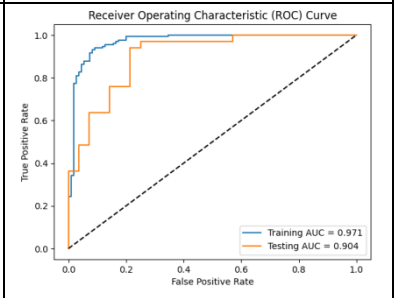
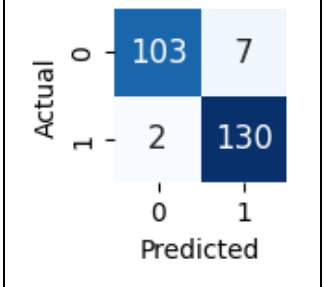
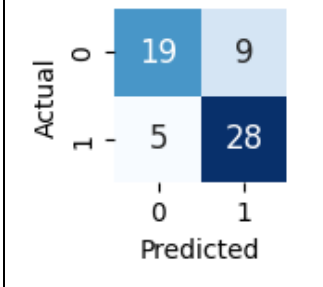
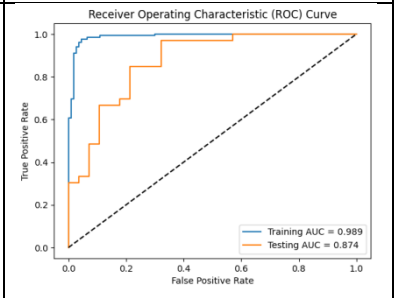
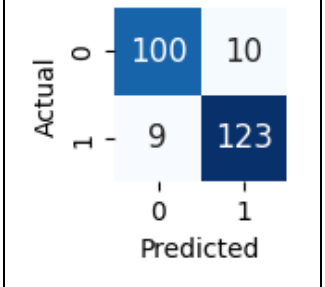
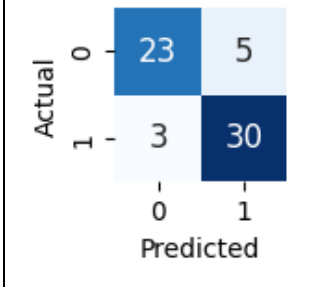
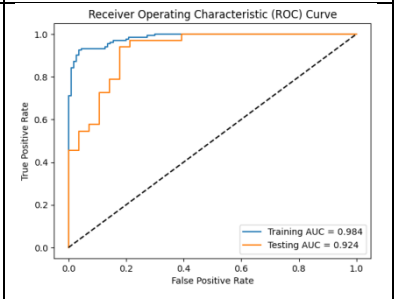


Fig. 5 Graphical analyses of models on PIMA Diabetes dataset

Table 6 Confusion Metrics and ROC Curve for Heart Disease

Algorithm	Confusion Metrics		ROC curve
	Training	Testing	
kNN			
LR			

DT			
RF			
SVM			
DNN			
ADABOOST			

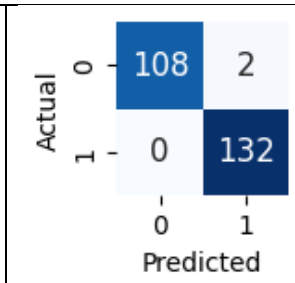
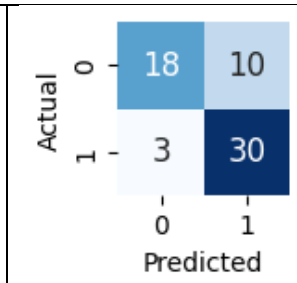
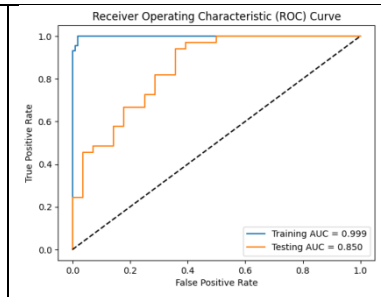
XGBOOST			

Table 7 Evaluation of models during training and validation phases for Heart Disease

Model	Training					Testing				
	Precision	Recall	F1 score	Accuracy	AUC	Precision	Recall	F1 score	Accuracy	AUC
kNN	0.91	0.848	0.878	0.87	0.948	0.84	0.818	0.83	0.819	0.888
LR	0.82	0.9	0.86	0.84	0.923	0.769	0.909	0.83	0.80	0.899
Decision Tree	0.90	0.969	0.93	0.92	0.971	0.81	0.909	0.857	0.836	0.875
RF	0.919	0.95	0.936	0.929	0.985	0.78	0.969	0.86	0.836	0.918
SVM	0.90	0.939	0.921	0.913	0.971	0.837	0.939	0.88	0.868	0.904
DNN	0.94	0.98	0.96	0.96	0.989	0.75	0.848	0.8	0.77	0.874
ADABOOST	0.92	0.93	0.92	0.92	0.984	0.85	0.909	0.88	0.86	0.924
XGBOOST	0.98	1	0.99	0.99	0.99	0.75	0.909	0.82	0.78	0.85

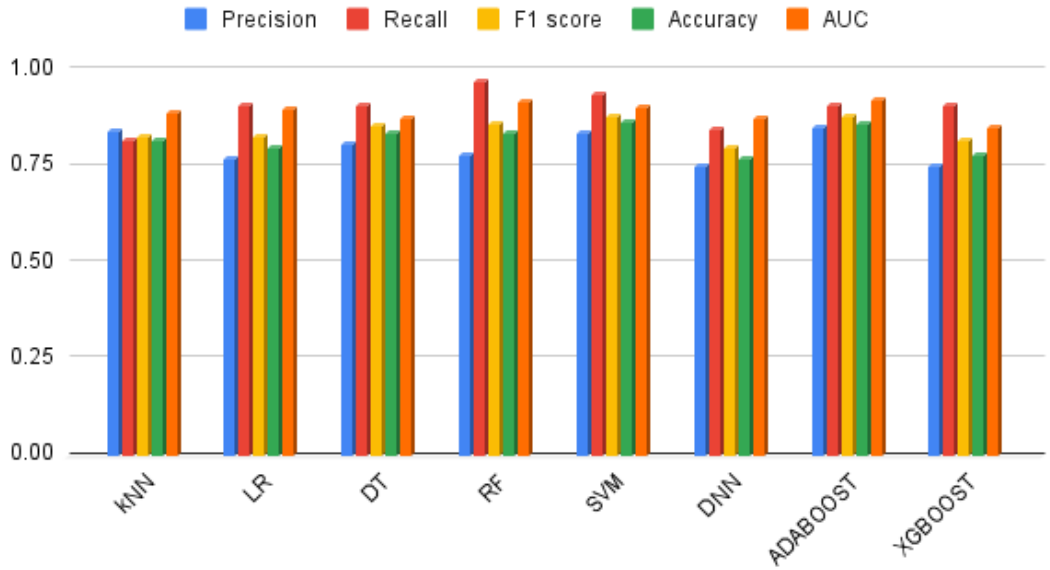


Fig. 6 Graphical analyses of models on Heart Disease dataset

VI. DISCUSSION

In PIMA Diabetes, Deep Neural Networks (DNN) and XGBoost demonstrate the best testing performance. DNN has a recall of 0.89 and an F1 score of 0.72, while XGBoost matches DNN's recall and has slightly better precision, making both models reliable for identifying diabetic patients. The AUC values for DNN (0.836) and XGBoost (0.83) confirm that these models are strong at distinguishing between positive and negative diabetes cases. Models like Random Forest and AdaBoost show good performance during training but have significantly lower test precision, indicating overfitting.

As depicted in Table 8, XGBoost and DNN show exceptionally high performance during training. However, this may indicate that models are fitting training data too closely, making the model overfit. On the other hand, Adaboost maintains the best balance between training and testing. With a precision of 0.85, recall of 0.909, and F1 score of 0.88 on testing data, along with high AUC of 0.924, it performs reliably across both phases, which ensures accurate diagnosis while minimizing false negatives. SVM is another strong performer in the testing phase with recall of 0.939 and an F1 score of 0.88, making it a good option when accuracy is critical. In disease prediction recall is especially important because it measures how many true positive diseases are correctly identified by the model. Missing positive cases could have serious consequences, so models with high recall are preferred. Adaboost and SVM offer strong recall values 0.909 and 0.939 respectively in testing phase making them particularly suitable for heart disease prediction. Logistic Regression also shows stable performance between training and testing with recall of 0.909 making it a simple but effective model for situation where interpretability is crucial.

VII. Conclusion

Our research presents prediction of lifestyle diseases, specifically diabetes and heart disease using ML models. We focused on Diabetes and heart diseases. The results of our research demonstrated the effectiveness of ensemble learning models such as Random Forest, ADABoost and XGBoost. Additionally, DNN is also proved to be highly competitive with ensemble Learning methods. These models are able to capture complex patterns within a dataset, resulting in consistently superior performance across various evaluation metrics compared to other traditional models. By incorporating multiple ML techniques, we were able to enhance the reliability of our model. Overall, our work contributes the advancement in healthcare sectors and provides insights into the application of ML in diseases diagnosis and prediction. However, several limitations exist in our research. Model performance may be impacted by dataset quality, size, and potential imbalances or missing values. Overfitting could also constrain generalization to new data if not adequately addressed. Future work should focus on refining the models by exploring more advanced techniques, such as deep learning, to further improve accuracy. Additionally, enhancing model interpretability will be crucial to provide actionable insights for healthcare practitioners. We also plan to explore the NHANES dataset for diabetes prediction using similar machine learning techniques to expand the scope of our research.

Acknowledgment

We would like to express our sincere gratitude to Indus University, Ahmedabad for providing the resources and support necessary to conduct this research. We are particularly grateful to the Indus Institute of Technology and Engineering (IITE) for their guidance and assistance throughout the study.

References

[1] Ahmed, R., Bibi, M., & Syed, S. (2023). Improving Heart Disease Prediction Accuracy Using a Hybrid Machine Learning Approach: A Comparative study of SVM and KNN Algorithms. International Journal of Computations, Information and Manufacturing, 49–54(1). https://doi.org/10.54489/ijcim.v3i1.223

[2] Begum, N. S. S., H, N. a. S. H. M., Karthikeyan, N. B., & Alanazi, N. F. Z. (2024). A Prediction of Heart Disease using IoT based ThingSpeak Basis and Deep Learning Method. Journal of Advanced Research in Applied Sciences and Engineering Technology, 47(1), 166–179. https://doi.org/10.37934/araset.47.1.166179

[3] Bhatt, C. M., Patel, P., Ghetia, T., & Mazzeo, P. L. (2023). Effective Heart Disease Prediction Using Machine Learning Techniques. Algorithms, 16(2), 88. https://doi.org/10.3390/a16020088

[4] Chang, V., Bailey, J., Xu, Q. A., & Sun, Z. (2022). Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms. Neural Computing and Applications, 35(22), 16157–16173. https://doi.org/10.1007/s00521-022-07049-z

[5] Heart Disease. (n.d.). UCI Machine Learning Repository. Retrieved March 27, 2024, from https://archive.ics.uci.edu/dataset/45/heart+disease

[6] IDF Diabetes Atlas. (n.d.). Copyright © IDF Diabetes Atlas 2024. All Rights Reserved. https://diabetesatlas.org/

[7] Jebamalar, G. B., Layola, J. A., Rajalakshmi, J., Thilagam, T., Kausthuban, M., & Nareshraj, R. (2024). Prediction and Comparison of ML Algorithm for Heart Disease. In Communications in computer and information science (pp. 303–313). https://doi.org/10.1007/978-3-031-73068-9_24

[8] Kaur, S., Bansal, K., Kumar, Y., & Changela, A. (2023). A Comprehensive Analysis of Hypertension Disease Risk-Factors, Diagnostics, and Detections Using Deep Learning-Based Approaches. Archives of Computational Methods in Engineering. https://doi.org/10.1007/s11831-023-10035-w

[9] Kavitha, M., Gnaneswar, G., Dinesh, R., Sai, Y. R., & Suraj, R. S. (2021). Heart Disease Prediction using Hybrid machine Learning Model. 2021 6th International Conference on Inventive Computation Technologies (ICICT). https://doi.org/10.1109/icict50816.2021.9358597

[10] Liang, Y., Liu, F., Yin, H., Shi, X., Chen, Y., Wang, H., Wang, Y., Bai, B., Liu, Y., Liu, Q., Wu, C., Yu, X., Ma, H., & Geng, Q. (2023). Trends in unhealthy lifestyle factors in US NHANES respondents with cardiovascular disease for the period between 1999 and 2018. Frontiers in Cardiovascular Medicine, 10. https://doi.org/10.3389/fcvm.2023.1169036

[11] Luong, A., Cheung, J., McMurtry, S., Nelson, C., Najac, T., Ortiz, P., Aronoff, S., Henderer, J., & Zhang, Y. (2024). Comparison of Machine Learning Models to a Novel Score in the Identification of Patients at Low Risk for Diabetic Retinopathy. Ophthalmology Science, 5(1), 100592. https://doi.org/10.1016/j.xops.2024.100592

[12] Mathukiya, D., Kumar, Y., & Koul, A. (2024). Automated diabetes diagnosis and risk assessment using machine learning. In CRC Press eBooks (pp. 215–222). https://doi.org/10.1201/9781003466383-32

[13] Modi, K., Singh, I., & Kumar, Y. (2023). A Comprehensive Analysis of Artificial Intelligence Techniques for the Prediction and Prognosis of Lifestyle Diseases. Archives of Computational Methods in Engineering, 30(8), 4733–4756. https://doi.org/10.1007/s11831-023-09957-2

[14] Patel, R., Sina, R. E., & Keyes, D. (2024, February 12). Lifestyle Modification for Diabetes and Heart Disease Prevention. StatPearls - NCBI Bookshelf. https://www.ncbi.nlm.nih.gov/books/NBK585052/

[15] pima-indians-diabetes.csv. (2018, February 27). Kaggle. https://www.kaggle.com/datasets/kumargh/pimaindiansdiabetescsv?resource=download

[16] Singh, J., Sandhu, J. K., & Kumar, Y. (2024). Metaheuristic-based hyperparameter optimization for multi-disease detection and diagnosis in machine learning. Service-oriented Computing and Applications. https://doi.org/10.1007/s11761-023-00382-8

[17] Singh, P., Silakari, S., & Agrawal, S. (2023). An Efficient Deep Learning Technique for Diabetes Classification and Prediction Based on Indian Diabetes Dataset. IEEE Conference Publication | IEEE Xplore. https://ieeexplore.ieee.org/document/10390518

[18] Tanim, S. A., Aurnob, A. R., Shrestha, T. E., Emon, M. R. I., Mridha, M., & Miah, M. S. U. (2024). Explainable deep learning for diabetes diagnosis with DeepNetX2. Biomedical Signal Processing and Control, 99, 106902. https://doi.org/10.1016/j.bspc.2024.106902

[19] Wee, B. F., Sivakumar, S., Lim, K. H., Wong, W. K., & Juwono, F. H. (2023). Diabetes detection based on machine learning and deep learning approaches. Multimedia Tools and Applications, 83(8), 24153–24185. https://doi.org/10.1007/s11042-023-16407-5

[20] Yasar, A. (2021). Data Classification of Early-Stage Diabetes Risk Prediction Datasets and Analysis of Algorithm Performance Using Feature Extraction Methods and Machine Learning Techniques. International Journal of Intelligent Systems and Applications in Engineering, 9(4), 273–281. https://doi.org/10.18201/ijisae.2021473767

[21] Zhu, J., Liu, H., Liu, X., Chen, C., & Shu, M. (2024). Cardiovascular disease detection based on deep learning and multi-modal data fusion. Biomedical Signal Processing and Control, 99, 106882. https://doi.org/10.1016/j.bspc.2024.106882