DOI: https://doi.org/10.52756/ijerr.2024.v46.001 *Int. J. Exp. Res. Rev.*, Vol. 46: 01-18 (2024)

Original Article | Peer Reviewed | Open Access

International Journal of Experimental Research and Review (IJERR)
© Copyright by International Academic Publishing House (IAPH)
ISSN: 2455-4855 (Online)
www.iaph.in

A Hybrid Transfer Learning Approach Using Obesity Data for Predicting Cardiovascular Diseases Incorporating Lifestyle Factors

Krishna Modi[1]*, Ishbir Singh[2] and Yogesh Kumar[3]

[1]Department of CSE, Indus Institute of Technology and Engineering, Indus University, Ahmedabad, Gujarat, 382115, India; [2]Department of ME, Indus Institute of Technology and Engineering, Indus University, Ahmedabad, Gujarat, 382115, India; [3]Department of CSE, School of Technology, Pandit Deendayal Energy University, Gandhinagar, Gujarat, India

E-mail/Orcid Id:

KM, krishnamodi1994@gmail.com, https://orcid.org/0000-0001-9840-5919; *IB,* ishbir@rediffmail.com, https://orcid.org/0000-0002-4233-3052; *YK,* yogesh.arora10744@gmail.com, https://orcid.org/0000-0002-2879-0441

Article History:

Received: 04th Jul., 2024
Accepted: 05th Dec., 2024
Published: 30th Dec., 2024

Keywords:
Cardiovascular disease Prediction, Machine Learning, Obesity data, Prediction models, Supervised learning, Transfer learning

How to cite this Article:
Krishna Modi, Ishbir Singh and Yogesh Kumar (2024). A Hybrid Transfer Learning Approach Using Obesity Data for Predicting Cardiovascular Diseases Incorporating Lifestyle Factors. *International Journal of Experimental Research and Review, 46,* 01-18.
DOI: https://doi.org/10.52756/ijerr.2024.v46.001

Abstract: Cardiovascular Diseases (CVDs), particularly heart diseases, are becoming a significant global public health concern. This study enhances CVD detection through a novel approach that integrates obesity prediction using machine learning (ML) models. Specifically, a model trained on an obesity dataset was used to add an 'Obesity level' feature to the heart disease dataset, leveraging the relation of high obesity with increased heart disease risk. We have also calculated BMI and added as a feature in CVD dataset. We evaluated this transfer learning-based novel approach alongside eight ML models. Performance of these models was assessed using precision, recall, accuracy and F1-score metrics. Our research aims to provide healthcare practitioners with reliable tools for early disease diagnosis. Results indicate that ensemble learning methods, which combine the strengths of multiple models, significantly improve accuracy compared to other classifiers. We are able to achieve a 74% accuracy score along with 0.72 F1 score, 0.77 precision and 0.80 AUC with XGBoost classifier, followed closely by the DNN with 73.7% accuracy with 0.72 F1 score, 0.75 precision and AUC of 0.798 with our proposed model. We seek to enhance healthcare efficiency and promote public health by integrating AI-based solutions into medical practice. The findings demonstrate the potential of ML techniques and the effectiveness of incorporating obesity-related features for optimized cardiovascular disease detection.

Introduction

Cardio Vascular Diseases (CVDs) are becoming a major public health concern globally. CVDs are the leading death cause globally. CVD is responsible for 179 lakh deaths annually, which is 32 percent of worldwide losses (World Health Organization: WHO, 2021). CVDs include coronary artery disease, stroke, heart failure and hypertension. Comprehensive analysis of CVDs, its risk factors and the importance of hemorheological properties are given in Dormandy (1987). Given the worldwide burden of CVDs that have lifestyle influence, this study will therefore focus on developing a model to accurately predict CVD risk by incorporating lifestyle-related features, specifically 'Obesity level' with the transfer learning method.

CVDs are significantly influenced by lifestyle factors, positioning them as one of the major Lifestyle diseases. Lifestyle diseases arise from personal behaviour and environment. The symptoms of these diseases are developed due to improper diet, insufficient physical activities, smoking, excessive alcohol consumption or stress. Making positive lifestyle changes can greatly reduce the risk of CVDs (Rippe, 2018).

***Corresponding Author:** krishnamodi1994@gmail.com

The rise of lifestyle diseases is interrelated with modernization. With the advancement of lifestyle, people are shifting from traditional and physically demanding lifestyles 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 has undoubtedly revolutionized various aspects of daily life and reduced physical work in numerous fields. Different types of lifestyle diseases, their risk factors, symptoms, illness and comprehensive analysis of predicting lifestyle diseases with AI and ML techniques are detailed precisely in Modi et al. (2023). While these advancements brought convenience and efficiency, they have also inadvertently contributed to a lifestyle, which resulted in the rise of CVDs and other lifestyle diseases.

CVDs play a substantial role in global mortality rates as well. Addressing CVDs requires multiple approaches. Awareness programs, educating people about healthy living, encouraging physical activities, and promoting a healthy diet can reduce the burden of these diseases and can improve overall public health. In the same regard, early prediction and detection are effective preventive measures. By utilizing modern ML algorithms, a person can have personalized health assistance and assessment by accurately predicting the risk of these diseases. Hence, individuals and healthcare professionals can implement timely and appropriate preventive measures, ultimately working towards improving public health and well-being worldwide.

Machine Learning (ML) is a branch of Artificial Intelligence that enables computers to learn from data and make prediction. Current research has validated that ML techniques play a significant role in the field of medical research (Modi et al., 2024; Bhavekar et al., 2024; Sivaraman and Khanna, 2021; Krittanawong et al., 2020; Pasha et al., 2020; Waigi et al., 2020). Comparative studies of different ML and DL models for CVDs are found in Abu-Naser et al. (2023), Subramani et al. (2023) and Ahmed et al. (2023). Above research concludes that ML algorithms offer the capacity to improve disease diagnosis and prediction. Kaur et al. (2022) suggest that obesity can increase the chance of CVDs and is considered as a key factor in many lifestyle diseases such as PCOS, diabetes, cancer, CVDs, etc. Naser et al. (2024) comprehensively review ML's role in enhancing cardiovascular disease prediction and outline future trends to advance public health efforts. Kumar et al. (2023) have demonstrated the effectiveness of a CNN-based deep learning model for predicting heart disease using ECG signals, demographic data, and clinical metrics, achieving over 90% accuracy.

This paper introduces a novel approach to predicting Cardiovascular Diseases (CVDs) by integrating an 'Obesity level' feature into the CVD prediction model using a hybrid transfer learning model. Transfer learning is a technique where a model trained on one task is adapted for another related task. We have utilized a dataset available on Kaggle to develop and evaluate our approach with eight different algorithms. Our approach involves training a model on an obesity dataset and incorporating the resulting 'Obesity level' feature into the CVD dataset. This method establishes the correlation between obesity and CVD risk to enhance prediction performance. We assessed the performance of our models using precision, recall, accuracy, AUC, and F1 scores. Ultimate goal of this research is to provide an advanced tool for more effective disease detection and prediction to healthcare practitioners.

We have developed a hybrid transfer learning model, which can identify a person with the possibility of CVDs. We aim to streamline healthcare processes and improve patient outcomes by integrating AI-based solutions into medical practice.

The Organization of the Article

After the introduction, the paper is structured as follows: Background, related work, and literature are provided. The proposed methodology in detail is presented, including dataset collection, feature extraction, data pre-processing, and data splitting. This section also outlines the ML models we have implemented for prediction. An overview of the performance metrics we have used to evaluate these models is provided, ensuring a robust assessment of their predictive accuracy. Results of the proposed methodology are compared and deliberated. Finally, the paper summarizes key findings, underscoring the significance of our research and its potential implications for future studies in this field.

Background

The early prediction and diagnosis of lifestyle diseases have a significant role in recent medical research. Existing predicting models include statistical methods and various ML techniques. Alghamdi et al. (2024) have proposed an automated ML model combining an arithmetic optimization algorithm with a multilayer neural network for accurate diagnosis of cardiovascular diseases (CVDs), which achieves superior performance compared to traditional methods by selecting the most relevant features. Saputra et al. (2023) have addressed the critical challenge of predicting cardiovascular diseases (CVDs) through data mining techniques, utilizing patient datasets to assess prognosis. While SGD and ANN have given the highest classification accuracy, the clustering methods identified two clusters within the CVD patient data. These CVD prognosis models can aid in preventive measures and reduce mortality rates. Bhatt et al. (2023) predicted heart disease using ML on a dataset of 70,000 samples, and they cleaned and reduced it to 57,155 by removing outliers. They have trained and evaluated various models such as DT, MLP, RF, and XGBoost. MLP model achieved the highest accuracy. Kavitha et al. (2021) have presented a study that investigates developing and implementing a hybrid model using RF and DT to predict heart disease. By incorporating multiple models, their proposed model enhances the reliability of heart disease prediction. Their work contributes to the predictive analysis of hybrid ML approaches and their significance in disease prognosis. Maiga et al. (2019) have also performed a comparative analysis of ML algorithms on the CVD dataset available on Kaggle and concluded that RF achieves the highest classification accuracy of 73%. Khan and Mondal (2020) applied classification algorithms on 3 different datasets and compared results in detail. They found that the cross-validation method gives better results compared to other methods. Shorewala (2021) has compared kNN, Logistic Classification, and Naive Bayes with ensemble techniques - bagging, boosting, and stacking, which have achieved notable accuracy improvements. Boosting models, with 73.4% accuracy, had the highest AUC score. Despite these improvements, the study highlights a key limitation: recall scores remained relatively low, averaging 66.8% across all models.

Similarly, researchers have used ML models to classify obesity levels based on various factors, including lifestyle factors. Obesity is a condition in which excessive accumulation of body fat leads to health risks such as diabetes, CVD, joint problems etc. Ferdowsy et al. (2021) investigated obesity prediction using nine different machine-learning methods, highlighting the potential of data-driven approaches in obesity management. Gogoi (2023) has evaluated seven machine learning (ML) algorithms for obesity detection, intending to emphasize the impact of feature selection methods. This research found that the gradient boosting algorithm has achieved the highest accuracy of nearly 97%. The results highlight the importance of feature selection in enhancing ML-based obesity prediction.

A detailed study by Al-Shoaibi et al. (2024), Powell-Wiley et al. (2021), Akil and Ahmad (2011) and Carbone et al. (2019) gives the relationship between obesity and CVD. They emphasize the need for healthcare systems to prioritize obesity management as a key strategy for preventing cardiovascular disease (CVD). It stresses the importance of addressing personal health interventions and societal influences contributing to rising obesity rates. Obesity is a significant risk factor for CVD, which directly contributes to hypertension and high cholesterol level (Sarkar et al., 2021, 2022; Madhual et al., 2023; Ranganathan et al., 2024). Integrating and including obesity as a predictive feature could enhance the capability of the model to capture more complex patterns and improve model performance. This addition aligns with the existing research that undergoes the value of feature selection in improving model outcomes, as demonstrated by Alghamdi et al. (2024), Wankhede et al. (2020) and Saputra et al. (2023). Thus, including obesity in predictive modeling can bridge existing gaps and offer a more comprehensive approach to cardiovascular disease diagnosis and prognosis.

Methodology

This section outlines the detailed description of steps followed in our research to ensure a robust and accurate prediction of CVD using transfer learning and ML techniques, as displayed in Figure 5.

Data Description

Two datasets are used in this research. One is for predicting CVDs, and the other is for training the obesity model. The primary dataset was obtained from Kaggle (Cardiovascular Disease Dataset, 2019), comprising a total of 70,000 samples and covering 12 features. This dataset was preferred because it is the largest publicly available dataset for CVD and provides a range of features, including lifestyle factors. This extensive dataset was selected to ensure sufficient data volume for effectively training and evaluating the models. Secondary dataset was sourced from the UCI repository and includes 17 features capturing various lifestyle habits and physical conditions, allowing for the classification of obesity levels based on the multiclass variable in a dataset of 2,111 instances. Description of both datasets is given in Table 1.

Statistical Analysis and Data Visualization

We conducted a descriptive statistical analysis on the dataset to understand the distribution of variables. Key statistics such as median, mean, standard deviation, variance and interquartile range (IQR) are calculated for continuous variables, as shown in Table 2. We employed histograms, violin plots and boxplots to visualize distribution of these features with respect to the cardio, which are illustrated in Figure 1. Frequent count and percentage were computed for the discrete variables, as outlined in Table 3. Bar plots were used to visualize frequency distribution with respect to the cardio, depicted in Figure 2. Figure 3 shows the frequency distribution of categorical features with respect to the target attribute, providing insights into how different categories, such as gender, cholesterol levels, glucose levels, smoking status, alcohol consumption, and physical activity, are distributed among individuals with and without cardiovascular disease.

Outlier Analysis and Removal

Data quality and consistency are essential for the performance of ML models. After collecting the dataset, we performed an outlier analysis to ensure unrealistic or erroneous data points did not distort model training. Outliers can skew the model results, leading to inaccurate predictions. In this study, we used the IQR method to identify and remove outliers from the features of weight,

Table 1. Dataset description.

Dataset No	Disease	Source	Total Features	Target Attribute	Number of Samples
1.	CVD	Kaggle (Cardiovascular Disease Dataset, 2019)	12	CARDIO (Binary)	70000
2.	Obesity	UCI Archive (Estimation of Obesity Levels, n.d.)	17	Nobeyesdad (Multiclass)	2111

Table 2. Statistical analysis of continuous attributes.

	Cardio	Age (days)	Height (cm)	Weight (kg)	Ap_hi (systolic BP)	Ap_lo (Diastolic BP)
Mean	0	18882	164	71.6	120	84.3
	1	20057	164	76.8	137	109
Median	0	19005	165	70	120	80
	1	20384	165	75	130	80
Standard deviation	0	2474	8.15	13.3	104	153
	1	2316	8.27	15	191	218
Variance	0	6.12E+06	66.4	177	10723	23313
	1	5.36E+06	68.4	224	36592	47439
IQR	0	4048	11	16	10	10
	1	3512	11	19	20	10
Minimum	0	10798	59	22	-120	0
	1	14275	55	10	-150	-70
Maximum	0	23678	207	200	14020	9800
	1	23713	250	200	16020	11000

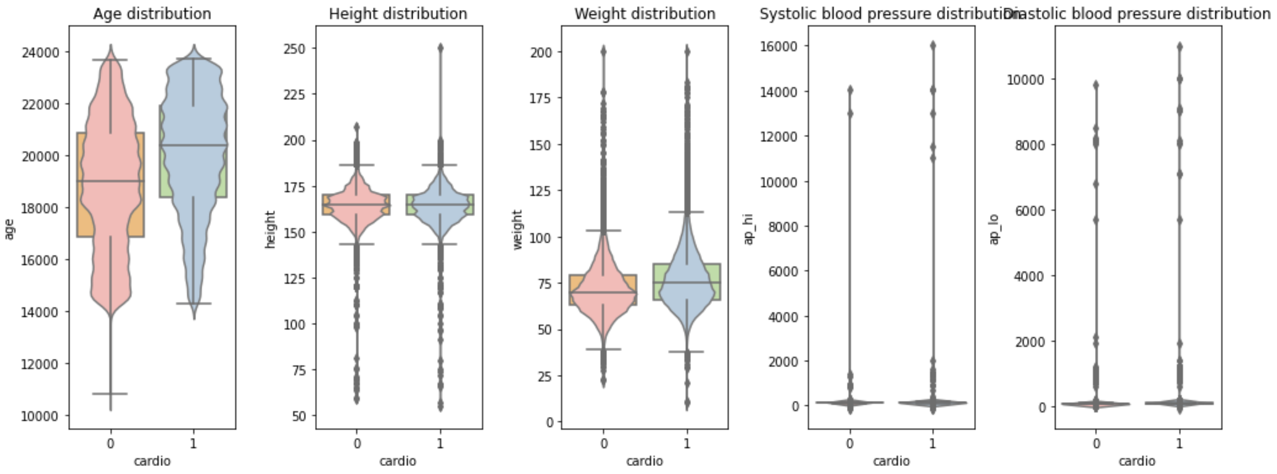


Figure 1. Box plots and violin plots of continuous features concerning the target attribute.

DOI: https://doi.org/10.52756/ijerr.2024.v46.001

height, age, and blood pressure.

Table 3. Frequency analysis of categorical attributes.

Feature	Categories	cardio	Counts	% of Total	Cumulative %
Gender	1 (Female)	0	22914	32.7 %	32.7 %
		1	22616	32.3 %	65.0 %
	2 (Male)	0	12107	17.3 %	82.3 %
		1	12363	17.7 %	100.0 %
Cholesterol (chol)	1	0	29330	41.9 %	41.9 %
		1	23055	32.9 %	74.8 %
	2	0	3799	5.4 %	80.3 %
		1	5750	8.2 %	88.5 %
	3	0	1892	2.7 %	91.2 %
		1	6174	8.8 %	100.0 %
Glucose level (gluc)	1	0	30894	44.1 %	44.1 %
		1	28585	40.8 %	85.0 %
	2	0	2112	3.0 %	88.0 %
		1	3078	4.4 %	92.4 %
	3	0	2015	2.9 %	95.3 %
		1	3316	4.7 %	100.0 %
Smoking Habit (smoke)	0	0	31781	45.4 %	45.4 %
		1	32050	45.8 %	91.2 %
	1	0	3240	4.6 %	95.8 %
		1	2929	4.2 %	100.0 %
Alcohol Consumption (alco)	0	0	33080	47.3 %	47.3 %
		1	33156	47.4 %	94.6 %
	1	0	1941	2.8 %	97.4 %
		1	1823	2.6 %	100.0 %
Physical Activity (active)	0	0	6378	9.1 %	9.1 %
		1	7361	10.5 %	19.6 %
	1	0	28643	40.9 %	60.5 %
		1	27618	39.5 %	100.0 %

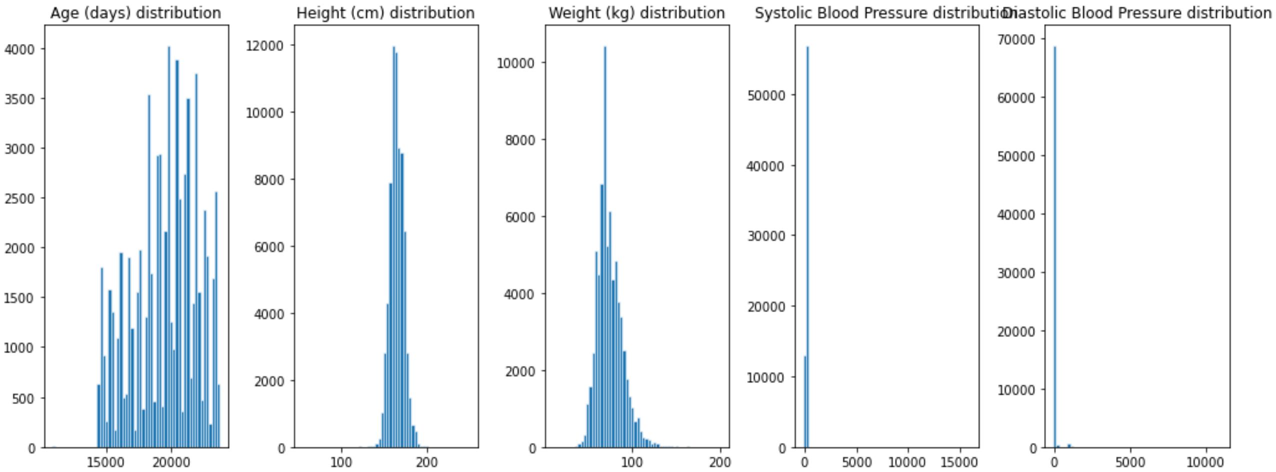


Figure 2. Frequency distribution over attributes.

IQR is a statistical method to remove outliers, in

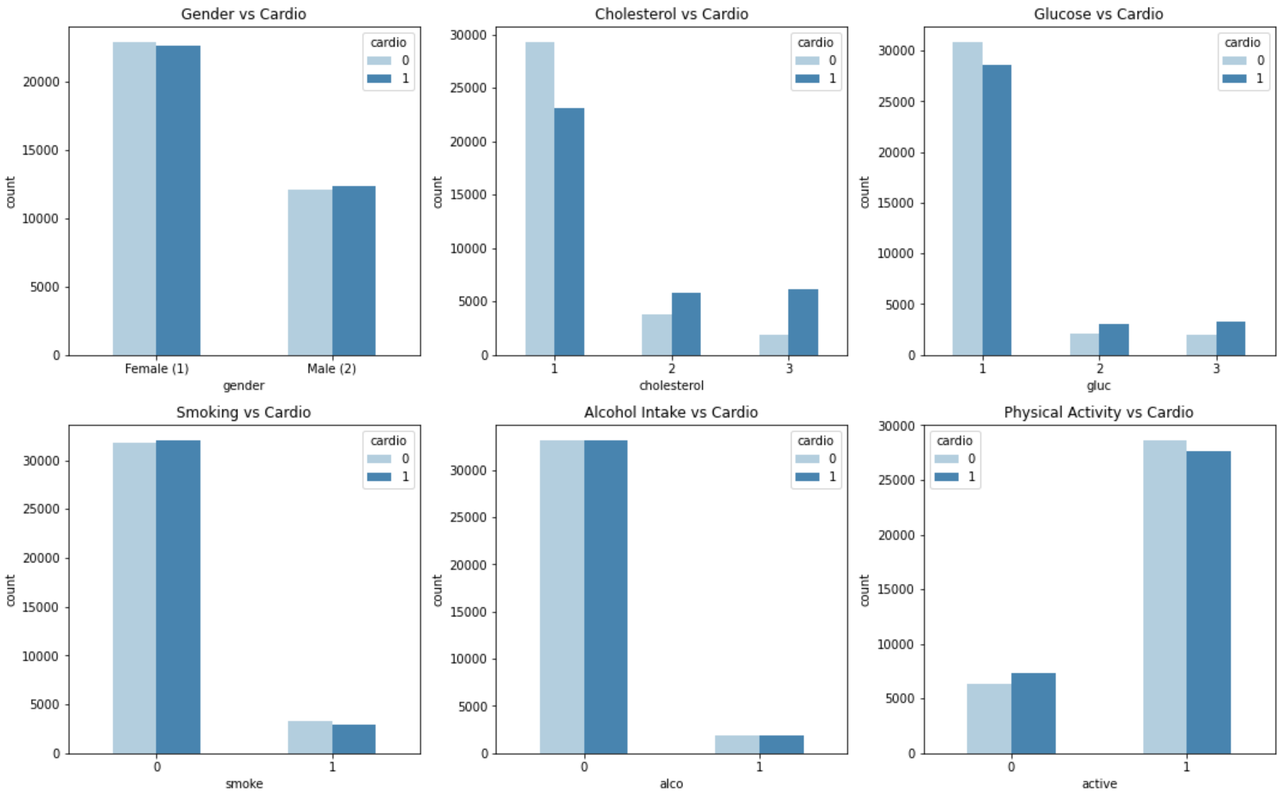


Figure 3. Frequency Distribution of categorical features with respect to the target attribute.

which the middle 50% of the data is represented by the difference between the 75th percentile (Q3) and the 25th percentile (Q1). The same is illustrated in Equation (1).

$$\mathrm{IQR} = \mathrm{Q3} - \mathrm{Q1} \qquad (1)$$

$$\mathrm{LB} = \mathrm{Q1} - (1.5 * \mathrm{IQR}) \qquad (2)$$

$$\mathrm{UB} = \mathrm{Q3} + (1.5 * \mathrm{IQR}) \qquad (3)$$

Any data points less than lower bound (LB) Equation (2) or greater than upper bound (UB) Equation (3) are classified as outliers.

Rather than removing all isolated outliers, we applied a more robust approach. Rows were flagged for removal only if they contained outliers in at least two of the selected features. This ensured that minor deviations in single features did not exclude potentially valuable data, while rows with multiple unrealistic values were removed. After the removal process, 69,150 samples remained in the dataset, providing a reliable basis for further analysis.

Additionally, when a new feature, BMI (Body Mass Index), was added, some values were exceedingly high, with a few samples having BMI values of up to **267**, which is practically impossible. We identified and removed 17 more records where the BMI exceeded 100, resulting in a final dataset with 69,133 samples.

Feature Engineering

Feature engineering is crucial because it helps ML models capture the relationship between data, improving models' predictive power. It is a critical step in building an effective model. This section outlines the feature engineering techniques applied in our study. Two additional features were engineered to enhance the dataset.

Feature creation

1. **Body Mass Index :** BMI is a well-known indicator of overall health and a strong predictor of cardiovascular risk. BMI is calculated with Equation 4.

$$\mathrm{BMI} = \frac{Weight\ (kg)}{Height\ (m^2)} \qquad (4)$$

2. **Obesity Level:** One of the most novel aspects of our approach is the integration of a feature representing obesity levels. This feature was generated using a model trained on an obesity dataset obtained separately from the UCI Repository. The obesity dataset contained features such as age, gender, height, weight, smoking habits, physical activity, and alcohol consumption, which were also present in the CVD dataset.

Feature Transformation from Obesity Dataset to CVD Dataset

To ensure consistency and compatibility between the obesity dataset and the CVD dataset, we implemented several transformations. Table 4 outlines the original features, their transformed counterparts, and the formulas used for conversion:

DOI: https://doi.org/10.52756/ijerr.2024.v46.001

Table 4. Feature Transformation.

Original Feature (Obesity dataset)	Transformed Feature (CVD Dataset)	Formula
Age (year)	Age (days)	Age(days) = Age (year) × 365
Height (m)	Height (cm)	Height(cm) = Height(m) × 100
Smoke (yes / no)	Smoke	$Smoke\ (CVD) = \begin{cases} 0\ if\ Smoke\ (Obesity) = no \\ 1\ if\ Smoke\ (Obesity) = yes \end{cases}$
FAF (Frequency of Physical Activity) (Range : 0-3)	Active (Binary)	$Active = \begin{cases} 0\ if\ FAF = 0\ or\ 1 \\ 1\ if\ FAF = 2\ or\ 3 \end{cases}$
CALC (Alcohol consumption) (No / Sometimes / Frequently / Always)	Alco (Alcohol Intake - Binary)	$Alco = \begin{cases} 0\ if\ CALC = No\ or\ Sometimes \\ 1\ if\ CALC = Frequently\ or\ Always \end{cases}$

Description of Transformations

1. **Age Conversion**: The age of individuals was originally recorded in years. To facilitate more granular analysis, this was converted to days by multiplying the age in years by 365.
2. **Height Conversion**: Height measurements in meters were transformed into centimeters, multiplying the height by 100 to provide a more commonly used unit in health assessments.
3. **Smoking Status**: The binary smoking status was derived from a yes/no response. A value of 1 indicates that the individual smokes, while 0 indicates non-smoking.
4. **Physical Activity**: FAF (Frequency of Physical Activity) was categorized into a binary format by classifying individuals who have reported frequency 0 and 1 as not active (0), and those who have reported frequency 2 and 3 were classified as active (1).
5. **Alcohol Consumption**: The consumption of alcohol was transformed into a binary variable, where individuals indicating 'No' or 'Sometimes' were classified as 0, and those indicating 'Frequently' or 'Always' were classified as 1.

Integration of Obesity Levels

We have first implemented the obesity dataset on eight different models and tested its performance. We found that XGBoost performed the best. By leveraging overlapping features from both datasets, we trained an XGBoost model to predict obesity levels, which are categorized into seven attributes. This prediction feature was then added as a new feature to the CVD dataset, enhancing its predictive power.

This transfer learning approach allowed us to apply insights from the obesity dataset to cardiovascular prediction. This approach takes advantage of the meaningful relationship between obesity and cardiovascular diseases. We subsequently converted the categorical obesity levels into a numeric format using label encoding. The ML model trained on the obesity dataset was carefully validated to ensure accurate predictions of obesity levels. Once validated, this model was applied to the CVD dataset, generating obesity-level predictions that served as an additional feature in the cardiovascular disease prediction task. The frequency chart displayed in **Figure 4** provides an insightful visualization of the distribution of categories of obesity and BMI distribution across the dataset.

Data Scaling

Data scaling is a preprocessing step to speed up training ML models. In our analysis, we employed standard scaling, which transforms the attributes to have a mean of zero and a standard deviation of one.

$$X_{scaled} = \frac{X - \mu}{\sigma} \qquad (5)$$

Where X is an original attribute, μ is the mean of the attribute and σ is the standard deviation. By applying standard scaling, we ensure that the input features are centered on zero, which helps in faster convergence during model training and enhances the overall model accuracy.

Data Splitting and Model Training

The pre-processed dataset, which included the newly created features, was split into 80% training and 20% testing sets. We employed stratified sampling for the categorical variables to maintain balanced class distribution during this split.

We evaluated and trained eight ML models on this dataset. Each model was trained on the training data, and hyper-parameters were optimized through the Grid Search method, systematically testing multiple combinations to identify the best configuration for each model. Hyper-parameters considered in this experiment are displayed in Table 5.

Additionally, the obesity dataset was validated using the same eight algorithms, with XGBoost yielding the highest performance. By utilizing the overlapping features from both datasets, we trained an XGBoost

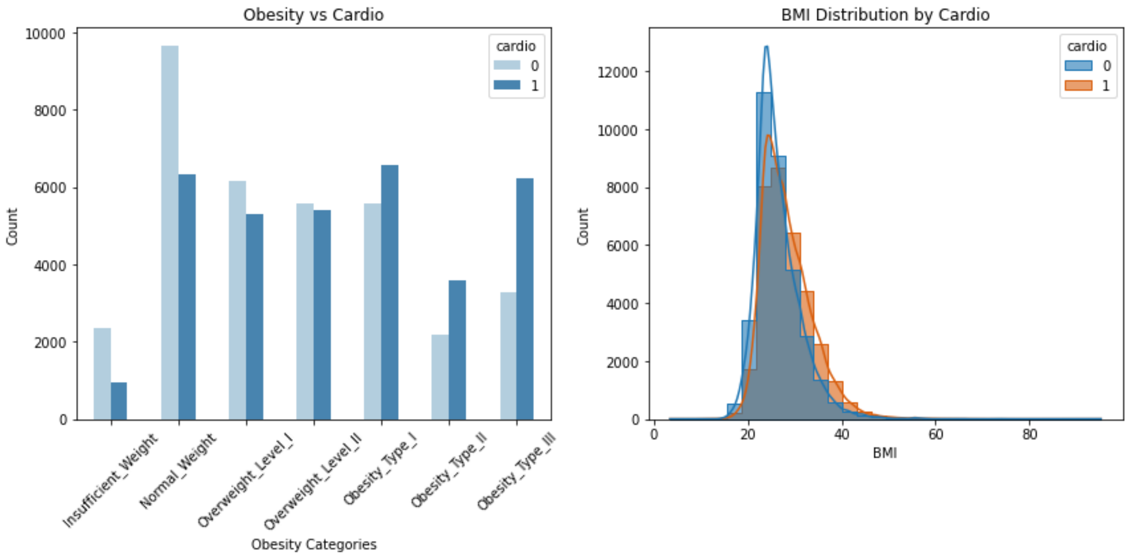


Figure 4. Distribution Analysis of Newly Incorporated Features.

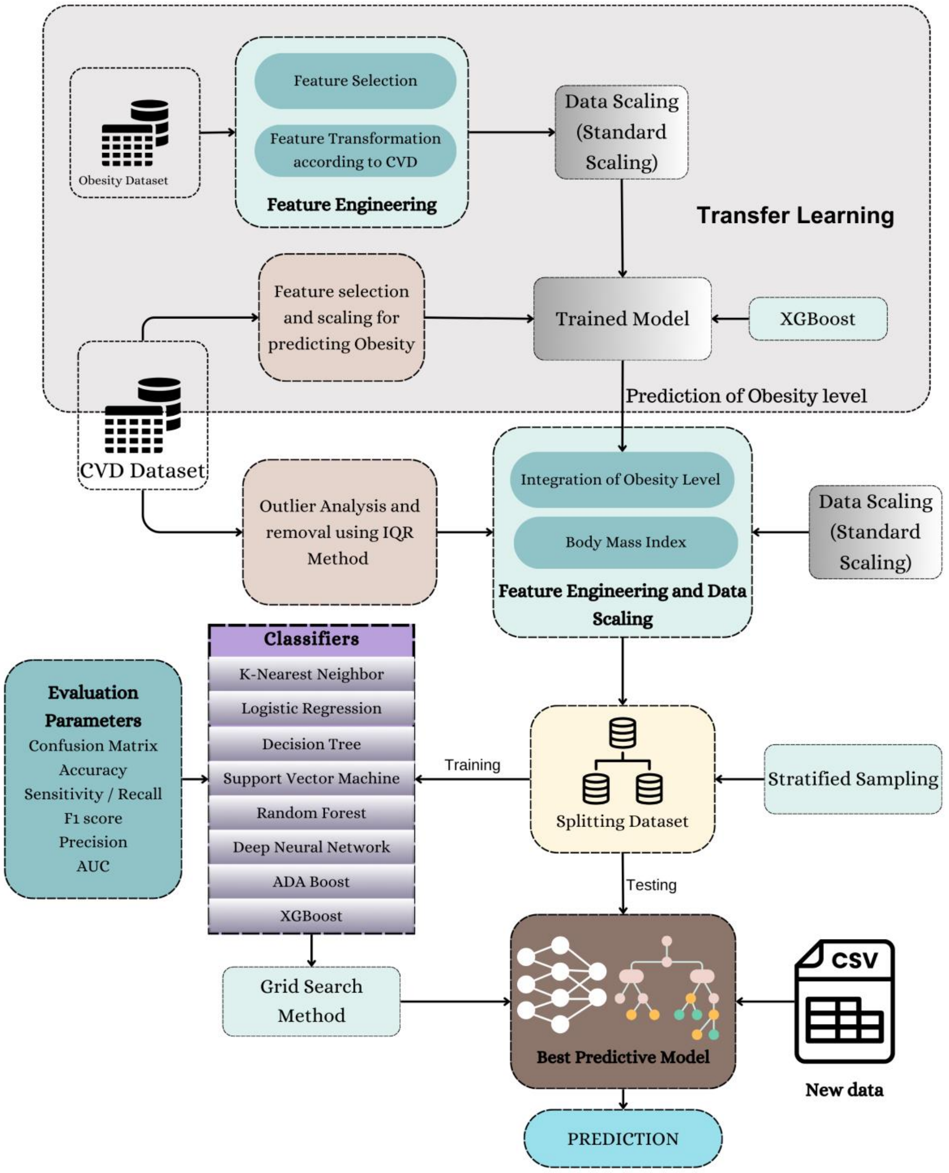


Figure 5. Proposed hybrid transfer learning approach for CVD prediction.

model to predict obesity levels and applied it to the CVD dataset. The predicted obesity levels were integrated into the CVD dataset, enhancing its predictive capabilities. This transfer-learning approach leveraged insights from the obesity dataset, allowing us to explore the significant relationship between obesity and CVD in a meaningful way.

DOI: https://doi.org/10.52756/ijerr.2024.v46.001

Table 5. Hyper parameters configuration.

Algorithm	Hyper parameters Used	Python Library
K Nearest Neighbor	K (nearest Neighbors) = {2,3,4,5,6,7} Euclidean Distance leaf_size :30	sklearn.neighbors
Logistic Regression	Regularization : L2 Regularization parameter C=1 Optimization Algorithm : lbfgs	Sklearn.linear_model
Decision Tree	Max Depth = {3,4,5,6} Splitting criteria : Gini Impurity	Sklearn.tree
Random Forest	Number of trees in forest : {100,125,150, 175} Splitting criteria : Gini Impurity	sklearn.ensemble

Support Vector Machine	Kernels : {linear, RBF, Poly, Sigmoid}
	C=1
		sklearn.svm

Deep Neural Network	Iterations : {100, 150, 200, 250} Activation Function : ReLU [F(z) = max(0, z)] Optimization Algorithm : Adam (based on stochastic gradient) Learning Rate : 0.001 Alpha = 0.001	sklearn.neural_network
ADABoost	Maximum number of estimators : {50, 75, 100, 125} Learning rate=1.0	sklearn.ensemble
XGBoost	Maximum number of estimators : {100, 125, 150, 175, 200} Maximum Depth = 3 Learning rate = 0.1 objective function = **softmax** for multiclass classification (for the dataset of obesity) **logistic** for CVD	xgboost

We have used Python 3.11 in the Spyder IDE to implement the proposed approach. The Python libraries employed in this research are mentioned in Table 5.

Model Evaluation

We evaluated the performance of these models on the test set. Several evaluation metrics were used to assess the models' effectiveness. Our proposed method, which incorporated novel feature engineering and ensemble learning, was compared to traditional machine learning approaches. The comparison demonstrated a noticeable improvement in prediction accuracy and other evaluation metrics, confirming the effectiveness of our methodology. The inclusion of the obesity level feature significantly enhanced the model's ability to predict cardiovascular diseases, particularly in cases where obesity plays a critical role.

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

Accuracy

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

$$\text{Accuracy} = \frac{\textit{Total Correct Predictions}}{\textit{All Predictions}} = \frac{N_{TN} + N_{TP}}{N_{TN} + N_{TP} + N_{FP} + N_{FN}} \quad (6)$$

Precision

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

$$\text{Precision} = \frac{\textit{Actual Predicted Positive}}{\textit{All Predicted Positive}} = \frac{N_{TP}}{N_{TP} + N_{FP}} \quad (7)$$

Recall

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

$$\text{Recall} = \frac{\textit{Actual Predicted Positive}}{\textit{All Predicted Positive}} = \frac{N_{TP}}{N_{TP} + N_{FP}} \quad (8)$$

F1 Score

The F1 Score is the harmonic mean of precision and recall. F1 Score provides a single metric that balances both precision and recall.

$$\text{F1 Score} = \frac{2}{\frac{1}{\text{precision}}+\frac{1}{\text{recall}}} \qquad (9)$$

ROC Curve (Receiver Operating Characteristics)

ROC curve plots True Positive Rate (sensitivity) and False Positive Rate at different threshold values. Value of the Area under ROC curve represents the model's performance. Higher value represents better performance.

Result and Discussion

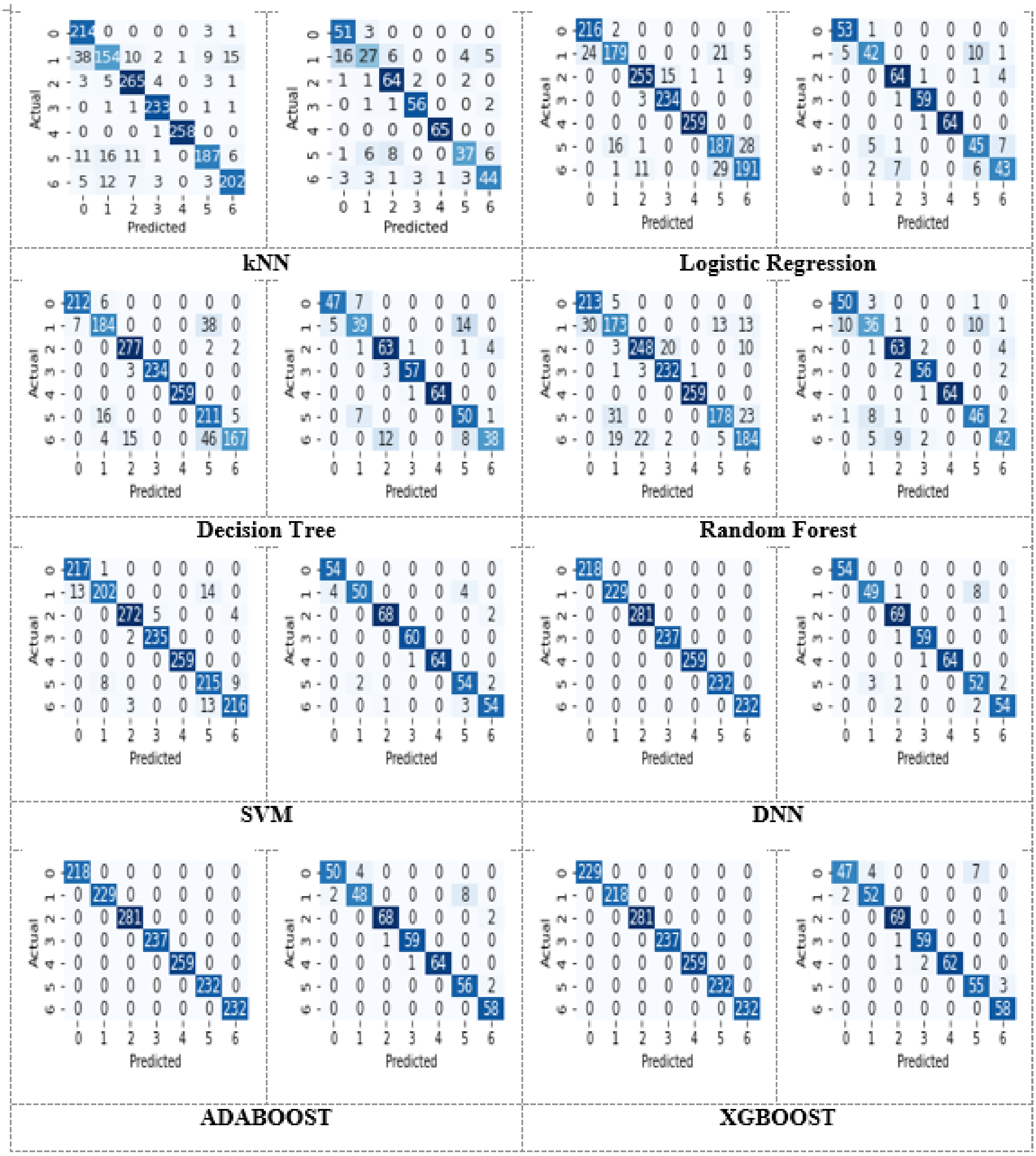
For the Obesity dataset, The ML models show excellent performance during training, with most achieving near-perfect or perfect scores, especially DNN, AdaBoost, and XGBoost, which all score 1.0 in precision, recall, F1 score, and accuracy. There is a slight drop in performance during testing, particularly with models like kNN, Random Forest, and Decision Tree, although they still perform well. SVM, AdaBoost, and XGBoost emerge as the top performers during testing, maintaining high F1 scores above 0.95. SVM and XGBoost stand out with an F1 score of 0.95, showing excellent performance and generalization, making it highly reliable for obesity prediction. DNN and AdaBoost have perfect scores during training but show some decline in testing, indicating overfitting. However, their performance in testing is still high (around 0.95), making them strong candidates for accurate prediction. Results of obesity are described in Table 6, Figure 6 and Figure 7.

The performance summary of the proposed method on ML models for cardiovascular disease prediction is provided based on training and testing evaluations using precision, recalls, F1 score, accuracy, and AUC metrics. There is no significant difference in training and testing set performance metrics, which shows that model is not overfitted. The k-Nearest Neighbors (kNN) model performed moderately. We got a training accuracy of 72.5% and a testing accuracy of 62%, with an AUC of 0.66. Logistic Regression (LR) showed improved results, with training and testing accuracies of 72% and consistent precision and recall, leading to a test AUC of 0.78. The Decision Tree model performed similarly, with a testing accuracy of 73% and a slightly lower AUC of 0.786. Random Forest also performed similarly to the decision tree, with 72.7% testing accuracy and a 0.793 AUC score. Support Vector Machine (SVM) also performed similarly to random forest, with a testing accuracy of 73% and an AUC of 0.78. Deep learning methods also fared well, with the Deep Neural Network (DNN) achieving the overall best training results, with a training accuracy of 75% and a testing accuracy of 73.7%. The testing AUC was 0.798, indicating strong predictive power. AdaBoost and XGBoost both performed competitively, with AdaBoost achieving a testing accuracy of 73% and an AUC of 0.796, while XGBoost slightly outperformed it with a testing accuracy of 74% and an AUC of 0.807, making it the top-performing model in this research.

The proposed methodology results showcased notable improvements, especially in the performance of Deep Neural Networks (DNN) and XGBoost, which now stand out as the most effective models for cardiovascular disease prediction. The DNN model, in particular, demonstrates exceptional progress, with its testing accuracy improving from 72.8% to 73.7% and AUC increasing from 0.795 to 0.798. Similarly, XGBoost also shows an impressive leap, with its testing accuracy rising from 73% to 74% and AUC climbing from 0.797 to 0.807, making it the top performer in terms of both accuracy and predictive power. These improvements solidify DNN and XGBoost's dominance over other models like kNN, which experienced a decline in performance, and Logistic Regression and Random Forest, which remained consistent but failed to match the advancements in DNN and XGBoost. The superior ability of DNN and XGBoost to adapt and optimize complex patterns in the data highlights their reliability and underscores the significance of these updated models in enhancing cardiovascular disease prediction.

Table 6. Performance of models for Obesity.

Model	Training				Testing			
	Precision	Recall	F1 score	Accuracy	Precision	Recall	F1 score	Accuracy
kNN	0.89	0.89	0.89	0.89	0.81	0.81	0.8	0.81
LR	0.90	0.90	0.90	0.90	0.87	0.87	0.87	0.87
Decision Tree	0.92	0.91	0.91	0.91	0.85	0.846	0.845	0.846
Random Forest	0.88	0.88	0.879	0.88	0.84	0.84	0.84	0.84
SVM	0.957	0.957	0.957	0.957	0.956	0.955	0.954	0.955
DNN	1	1	1	1	0.949	0.947	0.947	0.947
ADABOOST	1.0	1.0	1.0	1.0	0.953	0.952	0.952	0.952
XGBOOST	1.0	1.0	1.0	1.0	0.95	0.95	0.949	0.95



kNN

Logistic Regression

Decision Tree

Random Forest

SVM

DNN

ADABOOST

XGBOOST

Figure 6. Training and Validation Confusion Metrics for Obesity.

Table 7. Precision Comparison.

Model	Training		Testing	
	Traditional Approach	Proposed Approach	Traditional Approach	Proposed Approach
kNN	0.739	0.73	0.64	0.62
LR	0.739	0.746	0.739	0.749
DT	0.73	0.729	0.726	0.737
RF	0.75	0.746	0.74	0.75
SVM	0.768	0.74	0.76	0.748
DNN	0.77	0.768	0.75	0.757
ADABoost	0.77	0.768	0.78	0.77
XGBoost	0.759	0.76	0.77	0.769

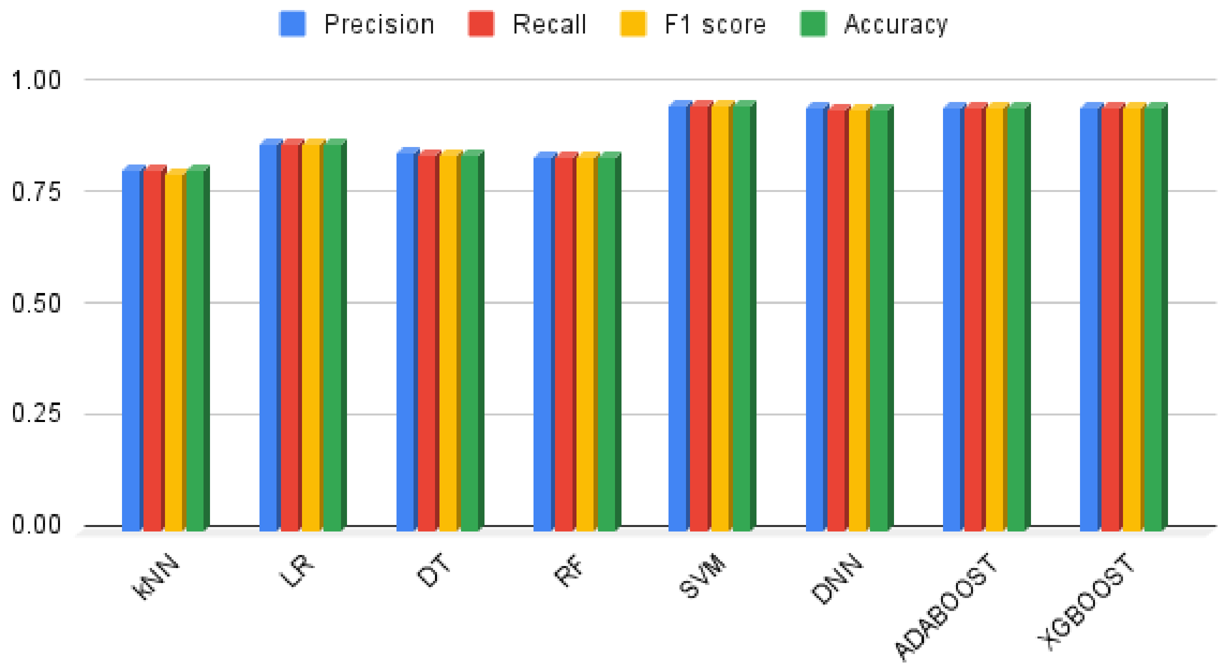


Figure 7. Graphical analyses of models on the obesity dataset.

Table 8. Recall Comparison.

Model	Training		Testing	
	Traditional Approach	Proposed Approach	Traditional Approach	Proposed Approach
kNN	0.717	0.70	0.618	0.60
LR	0.678	0.675	0.677	0.67
DT	0.72	0.72	0.72	0.71
RF	0.68	0.679	0.67	0.67
SVM	0.638	0.699	0.64	0.69
DNN	0.69	0.713	0.68	0.69
ADABoost	0.66	0.657	0.596	0.65
XGBoost	0.70	0.798	0.649	0.69

Table 9. F1 Score Comparison.

Model	Training		Testing	
	Traditional Approach	Proposed Approach	Traditional Approach	Proposed Approach
kNN	0.728	0.71	0.629	0.61
LR	0.70	0.709	0.706	0.71
DT	0.728	0.725	0.72	0.725
RF	0.71	0.71	0.70	0.71
SVM	0.69	0.72	0.70	0.72
DNN	0.73	0.739	0.715	0.72
ADABoost	0.71	0.708	0.679	0.70
XGBoost	0.729	0.73	0.707	0.728

Table 10. Accuracy Comparison.

Model	Training		Testing	
	Traditional Approach	Proposed Approach	Traditional Approach	Proposed Approach
kNN	0.732	0.725	0.636	0.62
LR	0.719	0.72	0.71	0.72
DT	0.73	0.728	0.72	0.73
RF	0.728	0.726	0.72	0.727
SVM	0.72	0.73	0.72	0.73
DNN	0.74	0.75	0.728	0.737
ADABoost	0.73	0.73	0.718	0.73
XGBoost	0.739	0.74	0.73	0.74

Table 11. AUC Comparison.

Model	Training		Testing	
	Traditional Approach	Proposed Approach	Traditional Approach	Proposed Approach
kNN	0.808	0.80	0.684	0.66
LR	0.783	0.78	0.784	0.78
DT	0.790	0.787	0.782	0.786
RF	0.798	0.794	0.789	0.793
SVM	0.789	0.79	0.787	0.78
DNN	0.819	0.823	0.795	0.798
ADABoost	0.799	0.798	0.790	0.796
XGBoost	0.808	0.815	0.797	0.807

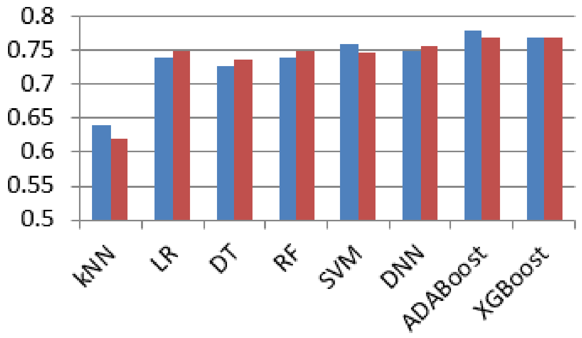


Figure 8. Comparison of Precision.

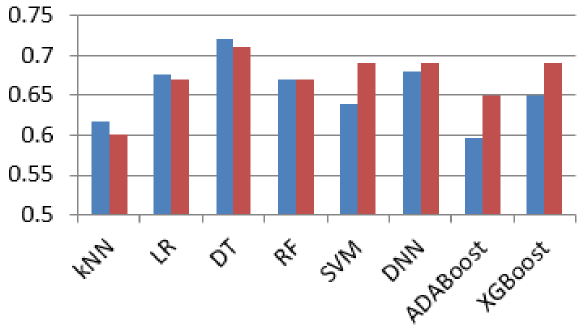


Figure 9. Comparison of Recall.

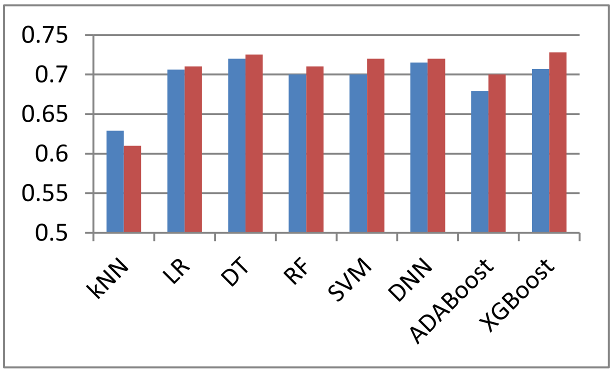


Figure 10. Comparison of F1

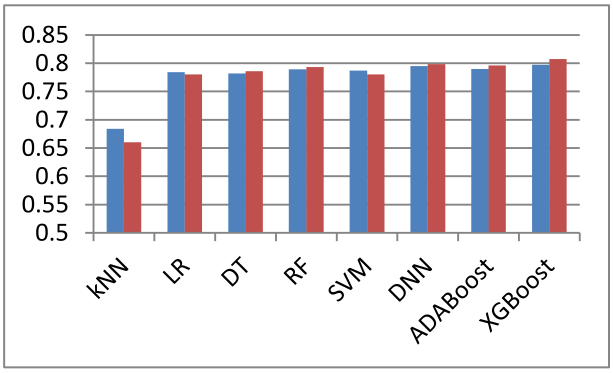


Figure 11. Comparison of AUC.

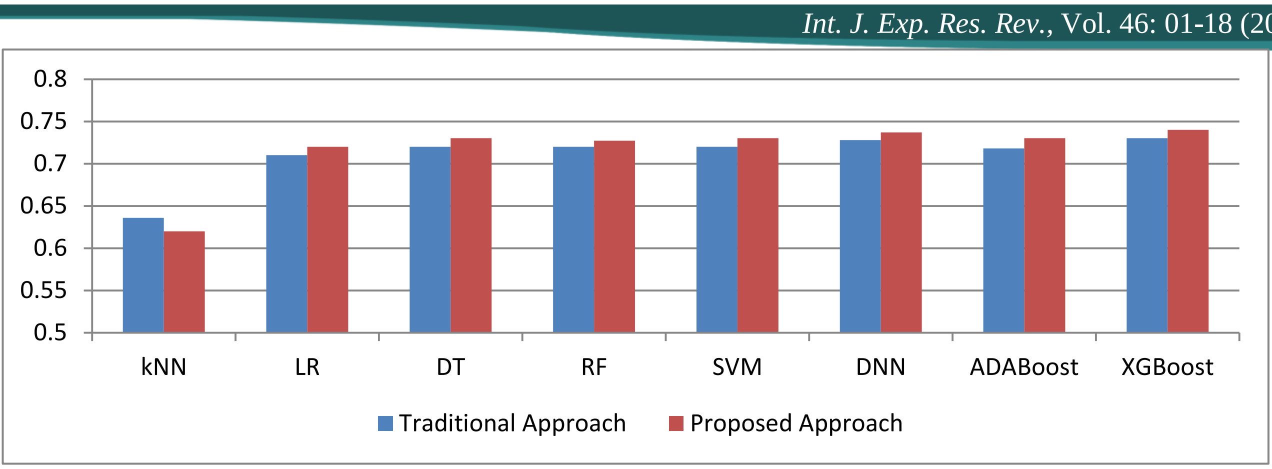
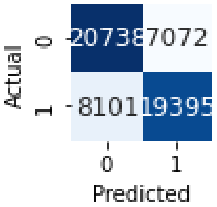
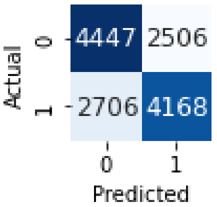
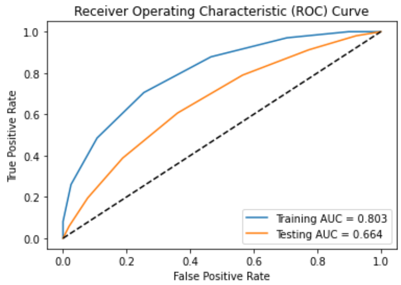
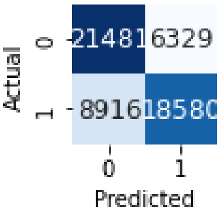
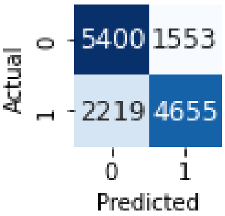
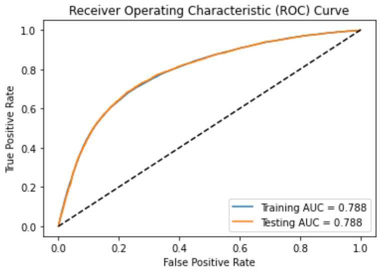
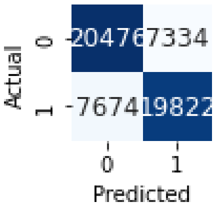
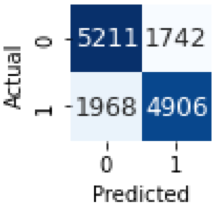
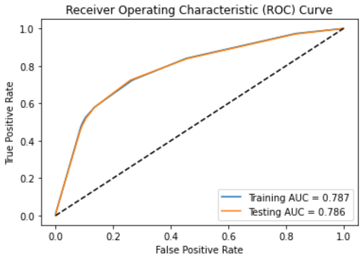
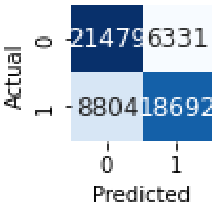
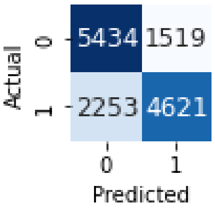
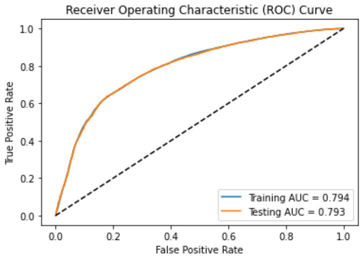
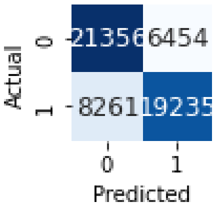
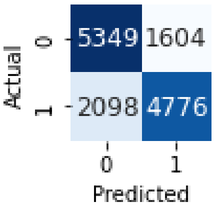
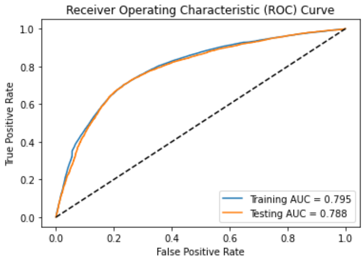
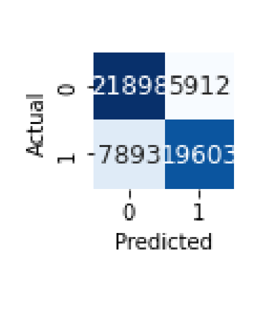
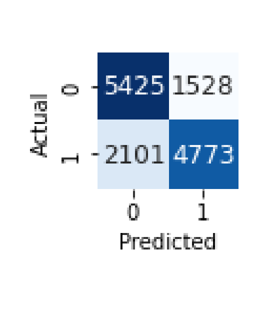
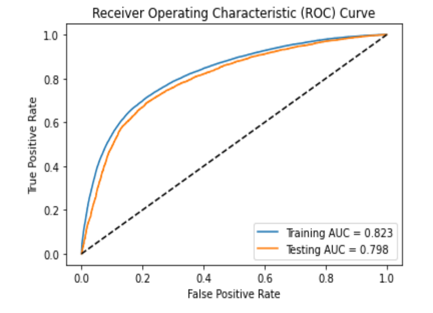
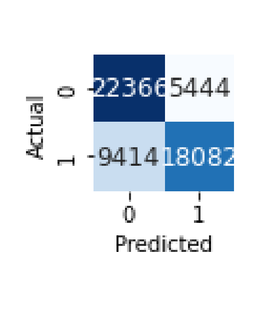
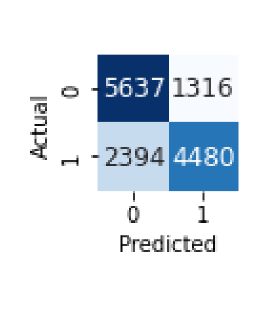
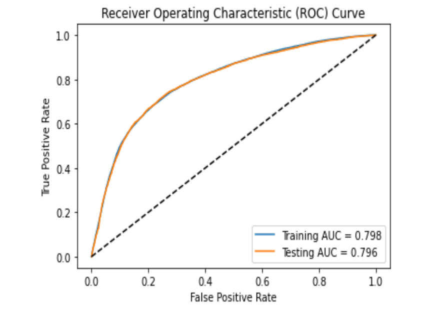
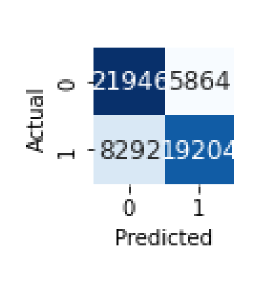
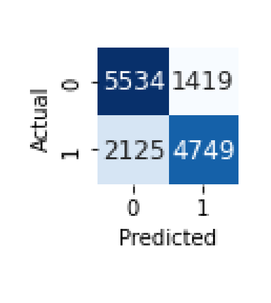
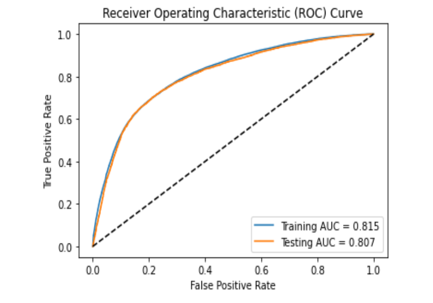


Figure 12. Comparison of Accuracy.

Table 12. Confusion Metrics and ROC Curve for CVD for the proposed approach.

Algorithm	Confusion Metrics		ROC curve
	Training	Testing	
kNN			
LR			
DT			
RF			
SVM			

DOI: https://doi.org/10.52756/ijerr.2024.v46.001

DNN			
ADABoost			
XGBoost			

Results illustrated in Figure 8 to Figure 12 clearly states that we are getting better predictive results for our proposed model 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 reduced. In addition to ensemble methods, DNN and SVM have also demonstrated their strong predictive capabilities. Using kernel functions, SVM can handle high-dimensional features and capture non-linear relationships within data.

Conclusion

This paper focuses on predicting cardiovascular diseases (CVD) using a novel transfer learning approach, aiming to improve early detection through advanced data analysis. The feature engineering process described in this paper was instrumental in preparing the dataset for ML models. We begin by reviewing existing research on CVD diagnosis and ML application in healthcare. The methodology outlines the steps involved, including data collection, feature engineering, data preprocessing and training for various models. We implemented the proposed model with several ML algorithms and evaluated them using performance metrics. The results highlight the effectiveness of our proposed model compared to others, showcasing its superior performance in predicting CVD. Specifically, XGBoost emerged as the top-performing algorithm with a testing accuracy of 74% and an AUC of 0.807, followed closely by the DNN with 73.7% accuracy and an AUC of 0.798, indicating significant improvements over other models. The study concludes by emphasizing the importance of ML in healthcare and its potential to impact the early diagnosis of cardiovascular diseases significantly. Despite these results, there are some limitations. Our dataset was limited to specific features, the inclusion of additional factors could enhance the performance model. Additionally, the transfer learning approach proved effective on our dataset. Its generalizability to other datasets may be limited without further validation. Future work is to implement the same transfer learning approach for other lifestyle diseases and test the proposed method's efficiency with other datasets, such as NHANES dataset.

Acknowledgement

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.

Conflict of Interest

The authors declare no conflict of interest.

References

Abu-Naser, S. S., Obaid, T., Abumandil, M. S. S., & Mahmoud, A. Y. (2023). Heart Disease Prediction Using a Group of Machine and Deep Learning Algorithms. *Advances on Intelligent Computing and Data Science*, pp. 81–196. https://doi.org/10.1007/978-3-031-36258-3_16

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 (IJCIM), 3*(1), 49–54. https://doi.org/10.54489/ijcim.v3i1.223

Akil, L., & Ahmad, H. A. (2011). Relationships between Obesity and Cardiovascular Diseases in Four Southern States and Colorado. *Journal of Health Care for the Poor and Underserved, 22*(4A), 61–72. https://doi.org/10.1353/hpu.2011.0166

Alghamdi, F. A., Almanaseer, H., Jaradat, G., Jaradat, A., Alsmadi, M. K., Jawarneh, S., Almurayh, A. S., Alqurni, J., & Alfagham, H. (2024). Multilayer Perceptron Neural Network with Arithmetic Optimization Algorithm-Based Feature Selection for Cardiovascular Disease Prediction. *Machine Learning and Knowledge Extraction, 6*(2), 987–1008. https://doi.org/10.3390/make6020046

Al-shoaibi, A. A. A., Li, Y., Song, Z., Hong, Y. J., Chiang, C., Nakano, Y., Hirakawa, Y., Matsunaga, M., Ota, A., Tamakoshi, K., & Yatsuya, H. (2024). Associations of overweight and obesity with the risk of cardiovascular disease according to metabolic risk factors among middle-aged Japanese workers: The Aichi Workers' cohort study. *Obesity Research &amp; Clinical Practice, 18*(2), 101–108. https://doi.org/10.1016/j.orcp.2024.02.006

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

Bhavekar, G. S., Das Goswami, A., Vasantrao, C. P., Gaikwad, A. K., Zade, A. V., & Vyawahare, H. (2024). Heart disease prediction using machine learning, deep Learning and optimization techniques-A semantic review. *Multimedia Tools and Applications, 83*(39), 86895–86922. https://doi.org/10.1007/s11042-024-19680-0

Carbone, S., Canada, J. M., Billingsley, H. E., Siddiqui, M. S., Elagizi, A., & Lavie, C. J. (2019). Obesity paradox in cardiovascular disease: where do we stand? *Vascular Health and Risk Management, 15*, 89–100. https://doi.org/10.2147/vhrm.s168946

Cardiovascular Disease dataset. (2019). Kaggle. https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset

Dormandy, J. A. (1987). Cardiovascular diseases. In Developments in cardiovascular medicine, pp. 165–194. https://doi.org/10.1007/978-94-009-4285-1_6

Estimation of Obesity Levels Based on Eating Habits and Physical Condition. (n.d.). UCI Machine Learning Repository. Retrieved March 27, 2024, from https://archive.ics.uci.edu/dataset/544/estimation+of+obesity+levels+based+on+eating+habits+and+physical+condition

Ferdowsy, F., Rahi, K. S. A., Jabiullah, Md. I., & Habib, Md. T. (2021). A machine learning approach for obesity risk prediction. *Current Research in Behavioral Sciences, 2*, 100053. https://doi.org/10.1016/j.crbeha.2021.100053

Gogoi, U. R. (2023). Importance of Feature Selection Methods in Machine Learning-Based Obesity Prediction, pp. 45–59. https://doi.org/10.1007/978-3-031-41925-6_3

Kaur, R., Kumar, R., & Gupta, M. (2022). Predicting risk of obesity and meal planning to reduce the obese in adulthood using artificial intelligence. *Endocrine, 78*(3), 458–469. https://doi.org/10.1007/s12020-022-03215-4

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), pp. 1329–1333. https://doi.org/10.1109/icict50816.2021.9358597

Khan, Md. I. H., & Mondal, M. R. H. (2020). Data-Driven Diagnosis of Heart Disease. *International Journal of Computer Applications, 176*(41), 46–54. https://doi.org/10.5120/ijca2020920549

Krittanawong, C., Virk, H. U. H., Bangalore, S., Wang, Z., Johnson, K. W., Pinotti, R., Zhang, H., Kaplin, S., Narasimhan, B., Kitai, T., Baber, U., Halperin, J. L., & Tang, W. H. W. (2020). Machine learning prediction in cardiovascular diseases: a meta-analysis. *Scientific Reports, 10*(1). https://doi.org/10.1038/s41598-020-72685-1

Kumar, L., Anitha, C., Ghodke, V. N., Nithya, N., Drave, V. A., & Farhana, A. (2023). Deep Learning Based Healthcare Method for Effective Heart Disease

Prediction. *EAI Endorsed Transactions on Pervasive Health and Technology, 9*. https://doi.org/10.4108/eetpht.9.4283

Madhual, S., Nayak, D., Dalei, S., Padhi, T., & Das, N. R. (2023). Assessment of cardiovascular risk factors in male androgenetic alopecia: A case control study in a tertiary care hospital of western Odisha. *Int. J. Exp. Res. Rev.*, *36*, 425-432. https://doi.org/10.52756/ijerr.2023.v36.037

Maiga, J., Hungilo, G. G., & Pranowo. (2019). Comparison of Machine Learning Models in Prediction of Cardiovascular Disease Using Health Record Data. *2019 International Conference on Informatics, Multimedia, Cyber and Information System* (ICIMCIS), pp. 45–48. https://doi.org/10.1109/icimcis48181.2019.8985205

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

Modi, K., Singh, I., & Kumar, Y. (2024). Predicting asthma control test score using machine learning regression models. *In CRC Press eBooks*, pp. 190–197. https://doi.org/10.1201/9781003466383-29

Naser, M. A., Majeed, A. A., Alsabah, M., Al-Shaikhli, T. R., & Kaky, K. M. (2024). A Review of Machine Learning's Role in Cardiovascular Disease Prediction: Recent Advances and Future Challenges. *Algorithms, 17*(2), 78. https://doi.org/10.3390/a17020078

Pasha, S. N., Ramesh, D., Mohmmad, S., Harshavardhan, A., & Shabana, N. (2020). Cardiovascular disease prediction using deep learning techniques. *IOP Conference Series Materials Science and Engineering, 981*(2), 022006. https://doi.org/10.1088/1757-899x/981/2/022006

Powell-Wiley, T. M., Poirier, P., Burke, L. E., Després, J., Gordon-Larsen, P., Lavie, C. J., Lear, S. A., Ndumele, C. E., Neeland, I. J., Sanders, P., & St-Onge, M. (2021). Obesity and Cardiovascular Disease: A Scientific Statement from the American Heart Association. *Circulation, 143*(21). https://doi.org/10.1161/cir.0000000000000973

Ranganathan, L., Rajasundaram, A., & Kumar, S. K. (2024). Demographic and Lifestyle Factors Influencing Cardiovascular Health Among Construction Workers: A Cross-Sectional Analysis. *International Journal of Experimental Research and Review, 42*, 312-319. https://doi.org/10.52756/ijerr.2024.v42.027

Rippe, J. M. (2018). Lifestyle Strategies for Risk Factor Reduction, Prevention, and Treatment of Cardiovascular Disease. *American Journal of Lifestyle Medicine, 13*(2), 204–212. https://doi.org/10.1177/1559827618812395

Saputra, J., Lawrencya, C., Saini, J. M., & Suharjito, S. (2023). Hyperparameter optimization for cardiovascular disease data-driven prognostic system. *Visual Computing for Industry, Biomedicine, and Art, 6*(1). https://doi.org/10.1186/s42492-023-00143-6

Sarkar, B., Biswas, P., Acharya, C.K., Jana, S.K., Nahar, N., Ghosh, S., Dasgupta, D., Ghorai, S.K., & Madhu, N.R. (2022). Obesity Epidemiology: A Serious Public Health Concern in India. *Chettinad Health City Medical Journal, 11*(1), 21-28. https://doi.org/10.24321/2278.2044.202205.

Sarkar, B., Ghorai, S. K., Jana, S. K., Dasgupta, D., Acharya, C. K., Nahar, N., Ghosh, S., & Madhu, N.R. (2021). Overweight and obesity in West Bengal: A Serious Public Health Issue. *VEETHIKA-An International Interdisciplinary Research Journal,7*(4), 9-14. https://doi.org/10.48001/veethika.2021.07.04.002

Shorewala, V. (2021). Early detection of coronary heart disease using ensemble techniques. *Informatics in Medicine Unlocked, 26*, 100655. https://doi.org/10.1016/j.imu.2021.100655

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, 18*(2), 163–182. https://doi.org/10.1007/s11761-023-00382-8

Sivaraman, K., & Khanna, V. (2021). Machine Learning Models for Prediction of Cardiovascular Diseases. *Journal of Physics Conference Series, 2040*(1), 012051. https://doi.org/10.1088/1742-6596/2040/1/012051

Subramani, S., Varshney, N., Anand, M. V., Soudagar, M. E. M., Al-keridis, L. A., Upadhyay, T. K., Alshammari, N., Saeed, M., Subramanian, K., Anbarasu, K., & Rohini, K. (2023). Cardiovascular diseases prediction by machine learning incorporation with deep learning. *Frontiers in Medicine, 10*. https://doi.org/10.3389/fmed.2023.1150933

Waigi, R., Choudhary, S., Fulzele, P., & Mishra, G. (2020). Predicting the risk of heart disease using advanced machine learning approach. European

Journal of Molecular & Clinical Medicine, 1638–1640.

Wankhede, J., Kumar, M., & Sambandam, P. (2020). Efficient heart disease prediction-based on optimal feature selection using DFCSS and classification by improved Elman-SFO. *IET Systems Biology, 14*(6), 380–390. https://doi.org/10.1049/iet-syb.2020.0041

World Health Organization (WHO). (2021). Cardiovascular diseases (CVDs). https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

How to cite this Article:

Krishna Modi, Ishbir Singh and Yogesh Kumar (2024). A Hybrid Transfer Learning Approach Using Obesity Data for Predicting Cardiovascular Diseases Incorporating Lifestyle Factors. *International Journal of Experimental Research and Review*, *46*, 01-18.
DOI : https://doi.org/10.52756/ijerr.2024.v46.001