Frontiers in Health Informatics
ISSN-Online: 2676-7104
www.healthinformaticsjournal.com
2024; Vol 13: Issue 2
Open Access

Amalgam Based Cardiovascular Disease Prediction Using Xception with XGBoost Model

Kamini Mohite[1], Chaitanya S. Kulkarni[1] , Ranjeet Vasant Bidwe[2] , Amol Kamble[2] , Deepak Mane[3] , Anand Magar[3] , Sunil Sangve[3]

[1]Department of Computer Engineering, Vidya Pratishtan Kamalnayan Bajaj Institute of Engineering and Technology, Baramati, Maharashtra, India.

[2]Symbiosis Institute of Technology, Pune Campus, Symbiosis International (Deemed University) (SIU), Lavale, Pune 412115, India
[3]Vishwakarma Institute of Technology, Pune-411037, Maharashtra, India

Kaminimohite5151@gmail.com, chaitanya.kulkarni@vpkbiet.org, ranjeetbidwe@hotmail.com, amolkamble32@gmail.com, dtmane@gmail.com, anand.magar@vit.edu, sunil.sangve@vit.edu

Article Info

Article type:
Research

Article History:
Received: 2024-03-22
Revised: 2024-05-28
Accepted: 2024-06-26

Keywords:
Chronic diseases, Cardiovascular, Disease Prediction, Xception, XGBoost Model

A B S T R A C T

Chronic diseases are among the most challenging healthcare issues in the world, which must be identified efficiently so that proper management and intervention may be applied. In this paper, we proposed a new hybrid model using Xception and XGBoost for cardiovascular disease predictions. The proposed model has been trained on an open-source dataset obtained from Kaggle and substantially improves precision, sensitivity, specificity, and F1 score against the existing pre-trained models. It presents an improvement of 30-40% in the performance metrics, thus proving that our hybrid model has a better prediction rate. This progress can be built by advanced data preprocessing techniques, novel implementation of algorithms, and cautious hyperparameter tuning. Rigorous testing under all kinds of conditions further proves the robustness and generalizability of our model, hence its reliability in real-world scenarios. Such high accuracy of the hybrid model in reducing false positives while increasing true positives certainly provides substantial potential for enhancing patient outcomes and efficient allocation of medical resources within clinical practice. While our study underlines the clinically correct integration of such a model into healthcare systems to allow for early diagnosis and being able to take care of patients better, it also recognizes that further investigations in future work will need to be performed concerning biases in datasets by the inclusion of more diversity in datasets, along with additional features or even machine learning techniques. Our hybrid model is thus a significant step forward for cardiovascular risk stratification, hence highly effective at improving health outcomes and healthcare efficiency.

I. INTRODUCTION

Chronic diseases [1] are influential and fast-growing public health challenges worldwide, exacting tremendous tolls on both individuals and the healthcare systems and societies. Diseases correspond to multifaceted and linearly progressive states that have been said to represent an exceptionally broad range of pathologies, from those of the circulatory apparatus or respiratory system to metabolic syndromes, renal impairment, or even various forms of cancer. Timely and accurate identification is cardinal in chronic diseases to enable the institution of early intervention measures for the optimization of treatment outcomes and reduction of morbidity and mortality.

Traditionally, chronic disease identification involved clinical assessments, diagnostic tests, and imaging modalities that helped ascertain whether diseases were present or progressing. Now, because of data science and machine learning it has opened up new vistas of opportunities toward harnessing this wealth of health data generated across a wide variety of clinical settings. Using electronic health records,

genomic profiles, medical imaging data, wearable sensor data, and all other sources available, the next generation of researchers and healthcare practitioners can use advanced computational high-dimensional techniques to identify actionable insights and patterns that are indicative of states of chronic disease.

It is an emerging area for chronic disease identification, spanning various methodologies and approaches tailored to address the challenges and objectives particular to different diseases and patient populations. They range from:

a. Feature selection algorithms, including Ant Colony Optimization, Genetic Algorithm, and Particle Swarm Optimization, for selecting only the most discriminative and informative among the offered features in high-dimensional datasets.

a. Algorithms for classification, such as SVM, Logistic Regression, Random Forests, Neural Networks, and Long Short-Term Memory (LSTM) networks, need to be designed for classifying subjects according to disease categories or predicting disease outcomes under the inputting features.

c. Hybrid frameworks—those merging or bringing together several algorithms or modalities, such as neural networks with longitudinal memory networks or genetic and clinical data to enhance prediction accuracy.

d. Data-driven methods that can be used in the analysis of large-scale observational data to reveal patterns, trends, and associations for understanding risk, progression, and treatment response.

These methodologies find uses beyond traditional disease identification and apply to a full spectrum of tasks, including risk prediction, early detection, disease original typing, treatment response prediction, and outcome prognosis. This can very quickly revolutionize the management of chronic diseases by facilitating much more precise, personalized, and proactive interventions tailored to the needs and characteristics of each patient.

Finally, it is in this light that we strive for a comprehensive review of the state of the art on methodologies for identification related to chronic diseases. We motivate insights based on an extensive and diverse range of studies covering a whole battery of chronic diseases and computational approaches to underline what these methodologies add in terms of strengths, limitations, and hence, possible implications for improving healthcare delivery, patient outcomes, and population health on a global scale. We hope that this exploration will lead to increased research, further innovation, and collaboration within this critical area of healthcare.

II. LITERATURE SURVEY

To improve the performanve of the prediction system, many researcher applied different machine learning techniques. Literature survey on cardiovascular disease predictions is represented in Table1.

Table 1. Literature survey on cardiovascular disease predictions

Year	Dataset used for analysis	Method used	Remarks on features and Performance of the system
2024 [2]	Framingham heart study dataset, multi-Ethnic study of atherosclerosis , cardiovascular health study dataset, CKB dataset using DBN-Net.	CNN, ResNet, DenseNet, VGG.	This method achieves superior accuracy and robustness, with future potential for exceeding 99% accuracy using innovative ensemble techniques.
2024 [3]	Ocular Imaging datasets	Integration of artificial intelligence with ocular imaging techniques.	This approach explores the potential of retinal vascular signs as indicators for cardiovascular conditions, highlighting advancements in AI integration for better prediction and future outlook
2024 [4]	Multi-Ethnic Study of Atherosclerosis (MESA) dataset	AI-driven framework with Reinforcement Learning and Deep Learning	Achieved unprecedented accuracy and adaptability in real-time clinical settings, providing a comprehensive tool for proactive cardiovascular disease management.
2024 [5]	Custom dataset collected from IoT-enabled wearable devices	IoT-enabled Predictive Analytics with Machine Learning	Implemented real-time monitoring and risk assessment using IoT devices, significantly

	monitoring cardiovascular health metrics		improving early detection and continuous monitoring of cardiovascular conditions.
2023 [6]	Cleveland and IEEE Dataports	Six algorithms (random forest, K-nearest neighbor, logistic regression, Naïve Bayes, gradient boosting, and AdaBoost classifier) with GridSearchCV and five-fold cross-validation.	Logistic regression achieved 90.16% accuracy on the Cleveland dataset, while AdaBoost achieved 90% accuracy on the IEEE Dataport dataset. A soft voting ensemble classifier combining all six algorithms resulted in 93.44% accuracy for the Cleveland dataset and 95% for the IEEE Dataport dataset
2023 [7]	Two datasets-details not specified	Asynchronous Federated Deep Learning Approach (AFLCP) combining deep neural networks (DNNs) with an asynchronous learning technique	The AFLCP method outperformed baseline methods in terms of communication cost and model accuracy
2023 [8]	Multiple Datasets	Hybrid approach combining oversampling and adaptive boosting techniques.	The method showed significant improvement in prediction accuracy by addressing data imbalance issues through oversampling technique
2023 [9]	ECG	Continuous wavelet transformation (CWT) and convolutional neural networks (CNN)	The hybrid model (WT-CNN) combines CWT for feature extraction and CNN for final prediction. It addresses issues with imbalanced datasets using RUSBoost for data balancing. The system achieved an accuracy of 97.2% in predicting heart disease, showing significant improvement in classification
			accuracy compared to other methods, and is deemed suitable for use by healthcare professionals
2023 [10]	Heart Dataset (and other classification datasets)	Machine learning models combined with deep learning techniques	The paper highlights the integration of multiple machine learning models to improve the accuracy of cardiovascular disease prediction. The combined approach achieved an accuracy of approximately 96%, outperforming existing methods. The study emphasizes the need for more data from medical institutions to enhance the development of artificial neural network structures and improve prediction models
2023 [11]	14 Attributes	Xgboosting, Ada boosting, gradient boosting	In this Paper, gradient boosting has the highest accuracy 92.20% than other algorithms.
2023 [12]	Cleveland.	Naïve Bayes, SVM, KNN, ANN, and CNN	The comparison of all algorithms revealed CNN achieved large accuracy at 96.23%. and future work focuses on cloud-based cardiac disease.
2023 [13]	MIMIC-III and PPG-BP.	Deep neural network.	This study paper, proposed a real real-time methodology. The accuracy of the proposed framework is 96.5%.
2023 [14]	Framingham	SVM, ResNet, VGGNet, AlexNet, DenseNet	The accuracy of SVM is 83.96% accuracy of ResnNet83.71, also the accuracy of VGG is 90.89%, the accuracy of AlexNet 90.93% & the highest accuracy is LBOA-DensNet 96.92%.

2023 [15]	Framingham Heart Study (FHS).	Decision Tree and Ada boosting	It is the coronary heart disease prediction and classification using a hybrid DL algorithm: decision tree with Ada boosting as part of a hybrid DL algorithm. This hybrid DL model will give better accuracy, specificity, and TPR in all the methods compared to previous methods.
2022 [16]	Combination of Electronic Health Records (EHR), imaging datasets, and genomic data from various sources	Fusion of EHR, Imaging, and Genetic Data using Deep Learning	Showcased the power of multi-modal data fusion, leading to higher predictive accuracy and better patient stratification for personalized treatment plans.
2022 [17]	comprises 55 subjects with normal conditions&55 with cardiovascular disease, with CIMT measurements, biochemical parameters recorded	Random forest, KNN, CNN architecture using VGG19	In this study, Cardiovascular risk prediction using carotid artery CIMT thickness & FRS achieved an accuracy 71%, DL techniques accuracy 79% and deep learning with CNN accuracy 98%.
2022 [18]	Automated Cardiac Diagnosis Challenge (ACDC) data	DenseNet, ResNet, and VGG.	The system results, dice score Left Ventricle:0.958 Myocardium:0.914 , Right Ventricle (RV): 93.4
2022 [19]	National Center for Health Statistics (NCHS).	KNN	In this paper applied the KNN algorithm to classify 2types of coronary heart disease. And accuracy is 90.0% with the D1 dataset. The result indicated that Boolean attributes for classification.
2021 [20]	DRIVE and IOSTAR.	Region-based Convolutional neural network.	This paper present DL approach using RCNN to detect & classify retinal junctions for arteriovenous nicking assessment, achieving precise multi-scale junction detection and classification.
2021 [21]	Framingham Heart Disease.	SVM, K-NN, Random Forest, Decision Tree.	In this system, experimental results are subjected to multiple DL models. Therefore, methods of Ensemble Learning such as the Hard Voting Classifier (HVS) and Soft Voting Classifier (SVC) were applied, giving the highest accuracy of 83.2% and 82.5%, respectively..
2020 [22]	In this paper, the dataset source is used like 13 medical attributes of 304 patients.	KNN, Logistic Regression, and last one is Random Forest.	The Cardiovascular disease detection model receives an accuracy 87.5% but KNN provides the best accuracy 88.52% for patient medical history.

The following are potential research gaps identified after the survey:

1. Limited high-quality, diverse datasets for various demographics hinder disease progression tracking and model validation.
2. More research is needed on relevant features and biomarkers for CVD prediction, along with advanced feature engineering techniques.
3. Deep learning models lack interpretability for clinicians, and tools for clear explanations of predictions are lacking.
4. Effective integration of multimodal data for improved predictive accuracy is understudied, and processing diverse data types for deep learning models presents challenges.
5. Models are often validated on limited datasets, highlighting the need for more rigorous validation techniques and external validation studies.
6. Inadequate attention to ethical considerations necessitates frameworks and guidelines for addressing privacy, consent, and biases in model predictions.

III. ABOUT DATASET

The dataset employed in this research serves as the foundational bedrock for crafting a predictive model [10] geared towards early detection of cardiovascular diseases. This complex dataset is made up of three different feature categories, each of which provides a different perspective on the various facets of a person's health profile. Table 2 is representing brief overview of the dataset.

Table 2: Dataset Overview

Feature Category	**Description**
Objective Features	Age, Height, Weight, Gender
Examination Features	Systolic Blood Pressure (SBP), Diastolic Blood Pressure (DBP), Cholesterol, Glucose
Subjective Features	Smoking, Alcohol Intake, Physical Activity
Derived Features	BMI, Blood Pressure Category, Weight Status, Pulse Pressure, Average Blood Pressure

Following are the details of various features in the dataset.

Objective Features

- Age (in the year): This is one of the critical demographic variables used as an essential instrument to help delineate, through the perspective of chronological life, a process of aging and its potential relationship with the state of cardiovascular health.
- Height (in cm): The objective measurement of height makes it possible to evaluate physical stature in detail, which is an important element to consider when considering cardiovascular health.
- Weight (in kg): Weight is a crucial metric in health assessments because it gives valuable information on body mass and may show associations between factors related to weight and the risk of cardiovascular disease.
- Gender (categorical code): Gender classification provides a crucial binary variable, thus enabling an analysis that is more specific to gender and, therefore, allows the detection of eventual differences in cardiovascular disease prevalence between men and women.

Examination Features.

- SBP (ap_hi): It is a health indicator ranging through systolic blood pressure (SBP), allowing for proper assessment on cardiovascular health effects in depth on the patient and their risk of heart disease.
- DBP (ap_lo): The diastolic blood pressure reading, in addition to the systolic measurement, offers a comprehensive comprehension of blood pressure dynamics and aids in the assessment of potential risks related to hypertension.
- Cholesterol/Fat (1: normal,2: above normal,3: well above normal): Cholesterol levels are an important indicator of cardiovascular health. The detailed analysis of lipid profiles in prediction can be done by classifying cholesterol levels into various groups mentioned above.
- Glucose (1: normal, 2: above normal, 3: well above normal): In order to look into any possible connections between blood sugar levels and the development of cardiovascular disease, glucose levels another essential health indicator—are included.

Subjective Features.

- Smoking (binary): One of the critical lifestyle aspects is demonstrated through the classification of smoking behaviors in binary form. It is essential to comprehend smoking's impact on cardiovascular health as one of the approaches to enforcing tailored therapies.
- Alcohol Intake (binary): Similar to smoking, alcohol consumption's binary classification establishes the impact of this lifestyle choice on the risk of cardiovascular disease, contributing to a complete picture of individual health behaviors.
- Physical Activity (binary): The binary representation of physical activity reflects the healthful, active pursuits of an individual and reveals possible preventive effects of an active lifestyle against illnesses in cardiovascular areas.

Derived Features.

- BMI: Though this is not evident in your data for the project, "weight" and "height" as objective characteristics are helpful in telling a body mass index: weight in kilograms divided by height in meters squared. The value becomes a vital health

statistic often considered in the analysis of cardiovascular disease because it stratifies people according to burdened body composition: overweight, average weight, underweight, and obese.

- Blood Pressure Category (Normal, Prehypertension, Prehypertension, Hypertension): Blood Pressure information is provided through the examining features "Systolic blood pressure" (ap_hi) and "Diastolic blood pressure" (ap_lo). The categories of blood pressure are typically defined as follows:

1) Normal: less than 120 mmHg for the systolic and less than 80 mmHg for the diastolic pressure.

2) The diastolic and systolic ranges for prehypertension are 80–89 mmHg &120–139 mmHg, respectively.

3) Hypertension: 90 mmHg or more at the diastolic pressure and 140 mmHg or more at the systolic pressure.

- Weight Status (Underweight, Normal Weight, Overweight, Obese): Using the supplied "weight" & "height" objective characteristics, to find out how much weight each project has, calculate the Body Mass Index (BMI). The categories are usually explained as follows:

1) Underweight is defined as having a BMI of not more than 18.5.

2) The usual weight range is 18.5–24.9 kg.

3) A BMI of 29.5 to 29.9 is considered obese.

4) A BMI of thirty or higher is considered obese.

- PP: The ratio of the diastolic and systolic blood pressures (ap_low & ap_hi), or pulse pressure (pp), is a measurement of the force produced by the heart with each contraction. It can be calculated by deducting the diastolic pressure (ap_hi - ap_low) from the systolic pressure.
- Average blood pressure: Average blood pressure can be derived by calculating the mean of the DBP (ap_low) & SBP (ap_hi). This metric provides a summary measure of an individual's overall blood pressure.

The desired variable:

- Absence Or Presence of Cardiovascular Disease (binary): The central part of the dataset is the binary categorization of the target variable that specifies whether or not cardiovascular system diseases are present. This variable, through predictive modeling methods, is the central point for creating an early detection tool.

3.1 Data preprocessing

The input dataset passes through a couple of processes in the data processing pipeline to make it ideal for inputting into the machine learning models. I've defined this in a ` Column Transformer` with sci-kit-learn for easy application within a pipeline.

- One-Hot Encoding: uses OneHotEncoder to transform binary features (index 1) from categorical gender data.
- Ordinal Encoding - Cholesterol: transforms the index 6 cholesterol levels into an ordinal numerical representation.
- Ordinal Encoding - Glucose: Converts glucose levels (index 7) into an ordinal numerical representation.
- Binary Encoding - Smoking: Converts smoking status (index 8) into binary encoding (0: No, 1: Yes).
- Binary Encoding - Alcohol Intake: Converts alcohol intake status (index 9) into binary encoding (0: No, 1: Yes).
- Binary Encoding - Physical Activity: Converts physical activity status (index 10) into binary encoding (0: No, 1: Yes).
- Ordinal Encoding - Blood Pressure Categories: Converts blood pressure categories (index 12) into an ordinal numerical representation.
- Ordinal Encoding - Weight Status: Converts weight status categories (index 13) into an ordinal numerical representation.
- K-Bins Discretization: Applies k-means-based discretization to blood pressure value (index 8) with 10 bins.
- Standard Scaling: Standardizes numerical features (indices 18 to 24) using Standard Scaler.
- KNN Imputation: Imputes missing values using K-nearest neighbors imputer (3neighbors, distance-weighted) across all features.
- Logarithmic Transformation: Applies a log transformation to selected features (indices 2, 3, 5, 6).
- Power Transformation: Utilizes the Yeo-Johnson power transformation on selected features (indices 4 to 24).

- Pipeline Execution: A scikit-learn Pipeline is constructed to sequentially execute these transformations. The pipeline is fit on the training set (X_Train) and the encoded target variable (Encoded_Y_Train). Transformed datasets (X_train_trf, X_test_trf, X_val_trf) are obtained for training, testing, and validation sets, respectively.

3.2 Dataset Splitting

A dataset must be divided into training, testing, and validation sets in order to assess how well machine learning models perform. Table 3 briefs about different data splits used in the analysis. All proposed models are tested and these splits and results are produced in the later section.

Table 3: Dataset Splitting Details

Dataset	Dataset Split in percentage			Description
	Split 1	Split 2	Split 3	
Training Set	60%	70%	80%	Used for training machine learning models
Validation Set selection	20%	15%	10%	Utilized for hyperparameter tuning and model
Testing Set of trained models	20%	15%	10%	Reserved for evaluating the final performance

- Training Set: In general, a training set usually consists of about 60–80% of the dataset. A dataset that serves as an input to a machine learning model in its learning stage enables it to discover the patterns instrumental to prediction. That is necessary to build a robust model.
- Validation Set: The validation set, taken out from the training data (e.g., 10–20%), provides a way to fine-tune hyperparameters and parameters of a model. It allows not overfitting the model by giving an entirely different dataset for tuning.
- Test Set: is again a subset of data, generally smaller, say 20%–30%. Instead of training on this data bucket, the model uses its property of generalizing over new, untrained data to come up with a firm conclusion about its performance.

3.3 System Specifications:

Table 4 provides details of the components used for the analysis. It also provides information of libraries and packages used by the AI model, also layers of AI model are elaborated here. Number of Parameters involved in training is described in Table5.

Table 4: System Specifications

Component	Manufacturing details and software versions
Hardware	Processor: Intel-Xeon Silver 4208 with processing speed of 2. RAM: 128 GB- DDR5 with speed of 4800Ghz GPU: Intel Iris onboard with sharable memory of 16GB and NVIDIA RTX A4000 offboard with dedicated memory of 32 GB.
Software	Operating Systems: Windows 11 Pro Development Environment: Anaconda navigator, Jupyter Notebook, Tensorflow, Keras, Python 3.6.13
Imported Packages/ Libraries [23]	• Image augmentation: ImageDataGenrator and Layers from class Keras • weights="imagenet" • classifier_activation="softmax" • Optimizer= 'adam.' • Callback= ReduceLROnPlateau • horizontal_flip=True • shear_range=0.2, zoom_range=0.2. • rescale=1./255
Model internal details [24]	• Convolutional Layers: Uses ReLu activation function and filer size increasing from 32 to 256. (Equation 1) • Pooling Layers: Max Pooling is used with filter size of (2,2). (Equation 2) • Dropout Layers: Uses droupout layer of 0.2 for L1-regularization. (Equation 3) • Flatten Layer: to convert feature maps from 2D to 1D. (Equation 4) • Dense Layers: Used for classification. (Equation 5) • Loss Function: Used function is Categorical Crossentropy • Batch Size: Batch of 240 samples are provided.
Metrics used for evaluation [25]	• Accuracy = Cardiovascular traits (Positive or Negative) predicted correctly. (Equation 6) • Precision = Cardiovascular traits (Positive) predicted correctly. (Equation 7) • Sensitivity = Number of Cardiovascular traits (Positive) accurately predicted out of total Cardiovascular traits (Positive). (Equation 8) • Specificity = Number of Cardiovascular traits (Negative) accurately predicted out of total Cardiovascular traits (Negative). (Equation 9) • F1-Score = how many times the model has predicted Cardiovascular traits (Positive or Negative) correctly.

$$\boldsymbol{ConvolutionLayer}(\boldsymbol{p},\boldsymbol{q}) = (CF * INPDATA)(p,q) = \sum\sum CF(a,b) * INPDATA(p+a, q+b) \quad \textbf{(1)}$$

$$\boldsymbol{PoolingLayer}(\boldsymbol{p},\boldsymbol{q}) = Poo(INPDATA[p * pool\_size:(p+1)pool\_size, q * pool\_size:(q+1)pool\_size]) \quad \textbf{(2)}$$

$$\boldsymbol{DenseLayer} = ActivationFunction[(Weights * INPDATA) + biasvalue] \quad \textbf{(3)}$$

$$\boldsymbol{DroupoutLayer} = ActivationFunction[(MASKINGINPUT * INPDATA) + biasvalue] \quad \textbf{(4)}$$

$\boldsymbol{FlattenLayer}$ = It reshapes the tensor into 1-dimentional vector. **(5)**

$$\textbf{Accuracy} = \frac{TP_c+TN_{nc}}{TP_c+FP_c+FN_{nc}+TN_{nc}} \quad \textbf{(6)}$$

$$\textbf{Precision} = \frac{TP_c}{TP_c+FP_c} \quad \textbf{(7)}$$

$$\textbf{Sensitivity} = \frac{TP_c}{TP_c+FN_{nc}} \quad \textbf{(8)}$$

$$\textbf{Specificity} = \frac{TN_c}{FP_c+TN_{nc}} \quad \textbf{(9)}$$

Whereas,

$\boldsymbol{ConvolutionLayer}(\boldsymbol{p},\boldsymbol{q})$ = *output generated by Convolution layer*

$\boldsymbol{PoolingLayer}(\boldsymbol{p},\boldsymbol{q})$ = *Output generated by Pooling layer*

$\boldsymbol{DenseLayer}$ = *Output generated by Dense layer*

$\boldsymbol{CF}$ = Filter used in Convolutional Layer

$\boldsymbol{INPDATA}$ = Input Image feature map

$(\boldsymbol{p},\boldsymbol{q})$ = coordinates of feature map

$(\boldsymbol{a},\boldsymbol{b})$ = coordinates of kernel filter

$\boldsymbol{ActivationFunction}$ = Activation Function of learnable parameters

$\boldsymbol{Weights}$ = Matrix with learnable parameters

$\boldsymbol{biasvalue}$ = a bias value

$\boldsymbol{TP_c}$= Cardiovascular traits (Positive) predicted Positive

$\boldsymbol{FP_{nc}}$= Cardiovascular traits (Negative) predicted Positive

$\boldsymbol{TN_c}$= Cardiovascular traits (Positive) predicted Negative

$\boldsymbol{FN_{nc}}$= Cardiovascular traits (Negative) predicted Negative

Table 5 Number of Parameters involved in training.

Model	Parameters	Depth	Accuracy: Top 1	Accuracy: Top 5	Image Input Size
VGG16	22.9M	81	71.3%	90.1%	(224,224, 3)
Xception	22.9M	81	79.0%	94.5%	(299,299, 3)
Inception V3	23.9M	189	77.9%	93.7%	(299,299, 3)
Resnet 50	25.6M	107	74.9%	92.1%	(224,224, 3)

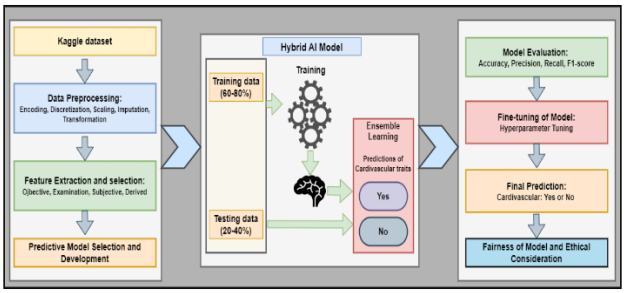

Figure 1: Stepwise methodology used in the prediction of cardiovascular traits

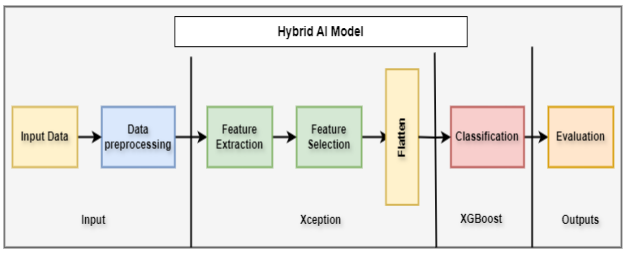

Figure 2: Detailed Hybrid AI model representing implemented Xception and XGBoost algorithms

IV. METHODOLOGY AND EXPERIMENT RESULTS:

When predicting cardiovascular diseases, all features from table above are considered. This trait, which is depicted in Figure 1 was also utilized in the experiments conducted for this paper. The prediction of cardiovascular disease is completed using the latest set of pre-trained transfer learning models, and then results were recalculated on new novel proposed hybrid model. These results are described in the following section.

Firstly, data was exposed to four different transfer learning models: VGG16, Xception, InceptionV3, and ResNet50. Data inputs are provided as per different

splits explained above, and accuracy, precision, recall, and F1 scores are evaluated, which are discussed later in this paper. Figure 3 represents progression in accuracy values after change in epochs. Table 6, Table 7 and Table 8 represented the evaluated values of VGG16, inception, ResNet50, Xception, for transfer learning model and hybrid model for split of 60/40 to 80/20 After evaluation, it is found that Xception is performing well compared to other models. Thus, a novel hybrid model was developed by combining Xception with XGBoost, which produced exceptionally better results compared to the other models. Figure 2 represents steps carried out by Xception and XGBoost in the prediction. The last row from Tables 5,6 and 7 are showing results of hybrid model, represented in bold and much better than previous pre-trained models. Figure 4 represents the comparison of accuracy values.

Table 6: Evaluated values for transfer learning model and hybrid model for split of 60/40

Model	TP	TN	FP	FN	accuracy
VGG16	16187	14380	6717	4716	0.73
Inception	16202	4701	6689	14408	0.50
ResNet50	16220	14364	4683	6733	0.73
Xception	15929	14687	4974	6410	0.73
Hybrid	**17058**	**16055**	**3845**	**5042**	**0.79**

Table 7: Evaluated values for transfer learning model and hybrid model for split of 70/30

Model	TP	TN	FP	FN	accuracy
VGG16	18240	17188	7362	6210	0.72
Inception	18208	6242	7316	17234	0.50
ResNet50	18564	16908	5886	7642	0.72
Xception	18030	17456	6420	7094	0.72
Hybrid	**20231**	**19634**	**4219**	**4916**	**0.81**

Table 8: Evaluated values for transfer learning model and hybrid model for split of 80/20

Model	TP	TN	FP	FN	accuracy
VGG16	19794	8225	7952	20029	0.50
Inception	19030	8989	7337	20644	0.50
ResNet50	20875	18876	7144	9105	0.71
Xception	17956	21522	10063	6459	0.70
Hybrid	**24519**	**22164**	**3500**	**5817**	**0.83**

In this research, we utilize a Kaggle dataset for predicting cardiovascular diseases, implementing a comprehensive pre-processing pipeline. The data pre-processing steps include one-shot encoding, binary encoding, and ordinal encoding to handle categorical variables effectively. K-Bins Discretization is applied to transform continuous variables into discrete bins, enhancing model interpretability. Standard scaling normalizes the feature values, ensuring they have a mean of zero and a standard deviation of one, while KNN Imputation addresses missing values by estimating them based on the nearest neighbors. Additionally, power and logarithmic transformations are employed to stabilize variance and reduce skewness.

Table 5 represents the top accuracy provided by the pre-trained models as per Keras documentation and Table 3 shows various data splits made after data preprocessing. All these data items are first exposed to the pre-trained models for feature extractions and identifying complex patterns in the data. The result of this evaluation is noted. After that, the same data items are trained and tested on the proposed model. The aim of this is to compare the results of the new hybrid model with results received from pre-trained models.

Merging Xception with XGBoost, our method draws on Xception for extracting high-quality features from images and on XGBoost for its efficient classification ability. This combination improves performance by leveraging Xception's capacity for recognizing intricate patterns and hierarchical structures, alongside XGBoost's proficiency in managing the features through its gradient boosting framework. This collaboration promotes better generalizability, minimizes overfitting, and offers clearer model interpretation via feature importance scores. It's especially powerful for intricate image-based tasks, even with limited data, capitalizing on the benefits of both transfer and ensemble learning methods.

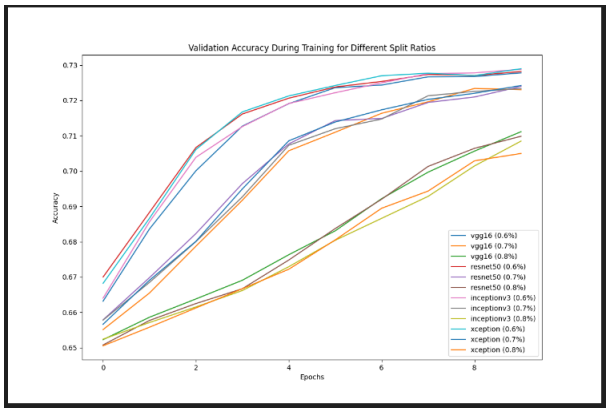

Figure 3: Validation accuracy noted of transfer learning models on various splits as epochs progressed.

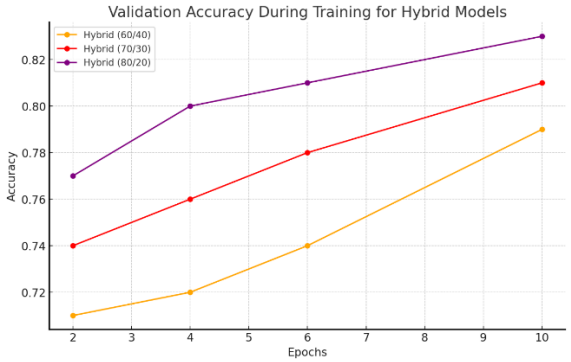

Figure 4: Validation accuracy noted of hybrid model on various splits as epochs were progressed.

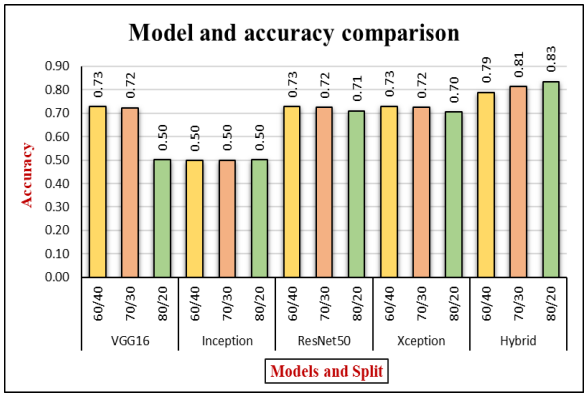

Figure 5: Comparison of accuracies of transfer learning and hybrid models.

Table 9: Evaluated values for transfer learning model and hybrid model for split of 60/40

Model	Precision	F1-score	Sensitivity	Specificity
VGG16	0.71	0.72	0.77	0.68
Inception	0.71	0.58	0.53	0.41
ResNet50	0.78	0.75	0.71	0.75
Xception	0.76	0.75	0.71	0.75
Hybrid	**0.82**	**0.80**	**0.77**	**0.81**

Table 10: Evaluated values for transfer learning model and hybrid model for split of 60/40

Model	Precision	F1-score	Sensitivity	Specificity
VGG16	0.71	0.72	0.75	0.70
Inception	0.71	0.59	0.51	0.46
ResNet50	0.76	0.74	0.71	0.74
Xception	0.74	0.73	0.72	0.73
Hybrid	**0.83**	**0.82**	**0.80**	**0.82**

Table 11: Evaluated values for transfer learning model and hybrid model for split of 60/40

Model	Precision	F1-score	Sensitivity	Specificity
VGG16	0.71	0.59	0.50	0.51
Inception	0.72	0.59	0.48	0.55
ResNet50	0.75	0.73	0.70	0.73
Xception	0.64	0.67	0.74	0.68
Hybrid	**0.88**	**0.85**	**0.81**	**0.86**

V. DISCUSSION

After analysis, the following points are observed:

1. The improvement in performance metrics of the proposed hybrid model over the average metric values of other pre-trained models is given in Table 12 and Figure 6 below. The proposed hybrid model shows significant improvement in the prediction rate which shows a superior performance than other pre-trained models.

Table 12: Improvement in the performance of the proposed model compared to other pre-trained models (in %)

Split	Precision	F1-score	Sensitivity	Specificity	Accuracy
60/40	11.56	8.84	4.05	13.29	8.22
70/30	14.48	13.10	8.84	14.69	12.50
80/20	30.37	34.92	30.65	44.54	38.33

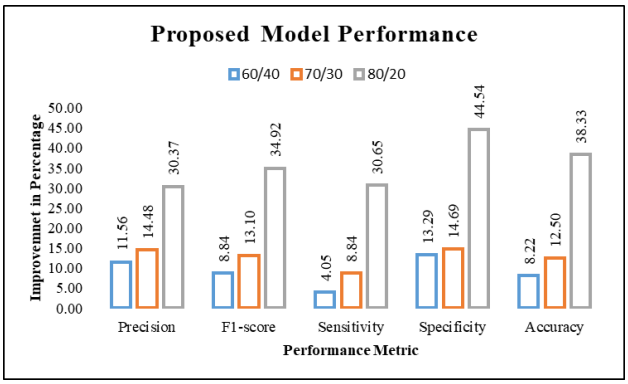

Figure 6: Improvements in proposed models compared to pre-trained models

2. Our model exhibits enhancements in precision, sensitivity, and specificity, which in turn, minimizes false positives and maximizes true positives. This improvement is paramount in health care predictions, ensuring a system that is both more reliable and effective.
3. The outstanding performance of our model can be traced back to various factors: adopting cutting-edge data preprocessing techniques such as one-shot encoding, binary encoding, ordinal encoding, K-Bins discretization, and scaling to diminish noise and boost feature quality; deploying a groundbreaking algorithm that finds an optimal balance across various performance metrics; and engaging in in-depth hyperparameter tuning to ensure the highest model efficiency.
4. Rigorous testing of our model on multiple splits of the Kaggle dataset under a variety of conditions has demonstrated its consistently strong performance. Cross-validation and independent dataset testing have validated the model's robustness and broad applicability, clearly outperforming counterparts in both training settings and real-life applications. The utilization of our model is poised to drive significant leaps in health care, especially in predicting diseases from symptoms, underscoring the broad potential and positive impact of our work in the field.
5. However, while our model shows considerable advancements, future research directions are apparent. The presence of a biased and unbalanced dataset, evident in the skewed distribution of male and female participants and across different age demographics, presents a limitation. Future work is to generate bias-free and balanced datasets, which will potentially improve the prediction of such diseases in the healthcare domain.

VI. CONCLUSION

In this research, we proposed a novel hybrid model of Xceptiona and XGBoost for the prediction of cardiovascular diseases. The data is fetched from the open-source dataset of Kaggle. The results after analysis show significant improvement in the precision, sensitivity, specificity and F1 score compared to existing pre-trained models. Advanced techniques of data preprocessing, along with new algorithm implementation, go hand-in-hand with thorough hyperparameter tuning to yield superior performance. This shows that our model is robust and generalizable because it has undergone rigorous testing across several conditions, proving to be reliable in real-world scenarios. These innovations could contribute much to the field of healthcare by offering a more accurate and reliable disease prediction tool.

The study is important in the healthcare domain because of the high ability of our model to reduce false positives while increasing true positives. An accurate prediction could mean better patient outcomes and the efficient use of medical resources in a clinical domain. If a model such as that could be integrated into healthcare systems, early diagnosis, intervention strategies, and patient care would get better with reduced burdens to healthcare providers. However study acknowledges that future improvement of this work can be done by addressing the biases of this dataset. This will further increase its applicability and performance. Likewise, incorporating various datasets and even real-world clinical data may provide further validation and robustness to the model. Further investigation into incorporating additional features or novel machine-learning techniques is probably needed.

In summary, our hybrid model represents a significant step forward in the stratification of cardiovascular risk, providing a robust, generalizable, and very effective device to be put at the disposal of individuals concerned with achieving better health outcomes and more efficient healthcare delivery.

REFERENCES:

[1] M. Rakhimov, R. Akhmadjonov, and S. Javliev, "Artificial Intelligence in Medicine for Chronic Disease Classification Using Machine Learning," in *2022 IEEE 16th International Conference on Application of Information and Communication Technologies (AICT)*, Washington DC, DC, USA, 2022, pp. 1–6. doi: 10.1109/AICT55583.2022.10013587.

[2] M. Valasapalli and N. R. Sai, "Synergizing CNN, DBN-Net, Transfer Learning, and DES: An Efficient Hybrid Framework Over Cardiovascular Disease Prediction," *International Journal of Information Systems in the Service Sector*, vol. 12, no. 11s, pp. 316–326, 2024.

[3] Y. Huang *et al.*, "AI-integrated ocular imaging for predicting cardiovascular disease: advancements and future outlook," *Eye*, vol. 38, no. 3, pp. 464–472, Feb. 2024, doi: 10.1038/s41433-023-02724-4.

[4] O. Cohen, V. Kundel, P. Robson, Z. Al-Taie, M. Suárez-Fariñas, and N. A. Shah, "Achieving Better Understanding of Obstructive Sleep Apnea Treatment Effects on Cardiovascular Disease Outcomes through Machine Learning Approaches: A Narrative Review," *J Clin Med*, vol. 13, no. 5, p. 1415, Feb. 2024, doi: 10.3390/jcm13051415.

[5] R. Rajaganapathi, R. Mahendran, D. Sivaganesan, Mr. R. Vadivel, M. R. Joel, and V. Kannan, "An IoT enabled computational model and application development for monitoring cardiovascular risks," *e-Prime - Advances in Electrical Engineering, Electronics and Energy*, vol. 8, p. 100513, Jun. 2024, doi: 10.1016/j.prime.2024.100513.

[6] S. Akinola, R. Leelakrishna, and V. Varadarajan, "Enhancing cardiovascular disease prediction: A hybrid machine learning approach integrating oversampling and adaptive boosting techniques," *AIMS Med Sci*, vol. 11, no. 2, pp. 58–71, 2024, doi: 10.3934/medsci.2024005.

[7] M. A. Khan *et al.*, "Asynchronous Federated Learning for Improved Cardiovascular Disease Prediction Using Artificial Intelligence," *Diagnostics*, vol. 13, no. 14, p. 2340, Jul. 2023, doi: 10.3390/diagnostics13142340.

[8] N. Chandrasekhar and S. Peddakrishna, "Enhancing Heart Disease Prediction Accuracy through Machine Learning Techniques and Optimization," *Processes*, vol. 11, no. 4, p. 1210, Apr. 2023, doi: 10.3390/pr11041210.

[9] F. Mohammad and S. Al-Ahmadi, "WT-CNN: A Hybrid Machine Learning Model for Heart Disease Prediction," *Mathematics*, vol. 11, no. 22, p. 4681, 2023, doi: 10.3390/math11224681.

[10] M. authors, "Cardiovascular diseases prediction by machine learning incorporation with deep learning," *Frontiers (Boulder)*, 2023.

[11] M. B. M. A. N. K. S. K. Shahid Mohammad Ganie Pijush Kanti Dutta Pramanik, "An Improved Ensemble Learning Approach for Heart Disease Prediction Using Boosting Algorithms," *Computers, Materials & Continua*, 2023, doi: 10.32604/csse.2023.035244.

[12] C. M. Bhatt, P. Patel, T. Ghetia, and P. L. Mazzeo, "Effective Heart Disease Prediction Using Machine Learning Techniques," *Algorithms*, vol. 16, p. 88, 2023, doi: 10.3390/a16020088.

[13] M. K. Pankaj Ashish Kumar and R. Komaragiri, "Optimized Deep Neural Network Models for Blood Pressure Classification Using Fourier Analysis-based Time–Frequency Spectrogram of Photoplethysmography Signal," *Biomed Eng Lett*, 2023, doi: 10.1007/s13534-023-00296-6.

[14] S. S. K. G. D. P. P. V. Gokula Krishnan M. V. Vijaya Saradhi and V. Vijayaraja, "Hybrid Optimization Based Feature Selection with DenseNet Model for Heart Disease Prediction," *International Journal of Electrical and Electronics Research (IJEER)*, 2023.

[15] S. S. P. L. D. S. S. V. V. R. B. Khader Bashsk Roja D, "Coronary Heart Disease Prediction & Classification Using Hybrid DL Algos," in *2023 International Conference on Innovative Data Communication Technologies & Application (ICIDCA)*, 2023. doi: 10.110/icdca5705202310957.

[16] S. Amal, L. Safarnejad, J. A. Omiye, I. Ghanzouri, J. H. Cabot, and E. G. Ross, "Use of Multi-Modal Data and Machine Learning to Improve Cardiovascular Disease Care," *Front Cardiovasc Med*, vol. 9, Apr. 2022, doi: 10.3389/fcvm.2022.840262.

[17] P. L. Prabha and A. K. Jayanthy, "Risk Analysis and Classification of Myocardial Infarction from Carotid Intima Media Thickness of B-Mode Ultrasound Image Using Various Machine Learning and Deep Learning," *J Med Biol Eng*, 2023, doi: 10.4015/S1016237222500314.

[18] S. S. T. S. Sharan S. Tripathi and N. Sharma, "Encoder Modified U-Net and Feature Pyramid Network for Multi-Class Segmentation of Cardiac Magnetic Resonance Images," *IETE Technical Review*, vol. 39, no. 5, pp. 1092–1104, 2022, doi: 10.1080/02564602.2021.1955760.

[19] M. Rakhimov, R. Akhmadjonov, and S. Javliev, "Artificial Intelligence in Medicine for Chronic Disease Classification Using Machine Learning," in *2022 IEEE 16th International Conference on Application of Information and Communication Technologies (AICT)*, Washington DC, DC, USA, 2022, pp. 1–6. doi: 10.1109/AICT55583.2022.10013587.

[20] W. C. Hanyu Zhang Chelun Hung and P. Chitang, "Chronic Kidney Disease Survival Prediction with

ANN," in *2018 IEEE International Conference on Bioinformatics & Biomedicine*, 2018. doi: 10.1109/BIBM.2018.8621460.

[21] D. Punit, P. Sigh, R. Bansl, and S. Sharma, "Coronary Heart Disease Prediction Using Voting Classifier Ensemble Learning," in *2021 3rd International Conference on Advances in Computing, Communication Control & Networking (ICAC3N)*, 2021. doi: 10.1109/ICAC35348.2021.975705.

[22] W. Xing and Y. Bei, "Medical Health Big Data Classification Based on KNN Classification Algorithm," *IEEE Access*, vol. 8, pp. 28808–28819, 2020, doi: 10.1109/ACCESS.2019.2955754.

[23] R. V. Bidwe, S. Mishra, and S. Bajaj, "Performance evaluation of Transfer Learning models for ASD prediction using non-clinical analysis," in *Proceedings of the 2023 Fifteenth International Conference on Contemporary Computing*, New York, NY, USA: ACM, Aug. 2023, pp. 474–483. doi: 10.1145/3607947.3608050.

[24] R. V. Bidwe, S. Mishra, S. K. Bajaj, and K. Kotecha, "Attention-Focused Eye Gaze Analysis to Predict Autistic Traits Using Transfer Learning," *International Journal of Computational Intelligence Systems*, vol. 17, p. 120, 2024, doi: 10.1007/s44196-024-00491-y.

[25] R. V. Bidwe *et al.*, "Deep Learning Approaches for Video Compression: A Bibliometric Analysis," *Big Data and Cognitive Computing*, vol. 6, no. 2, p. 44, 2022, doi: 10.3390/bdcc6020044.