IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 27 September 2023, accepted 16 October 2023, date of publication 18 October 2023, date of current version 27 October 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3325705

RESEARCH ARTICLE

DASMcC: Data Augmented SMOTE Multi-Class Classifier for Prediction of Cardiovascular Diseases Using Time Series Features

NIDHI SINHA[1], (Member, IEEE), M. A. GANESH KUMAR[1], AMIT M. JOSHI[1], (Senior Member, IEEE), AND LINGA REDDY CENKERAMADDI[2], (Senior Member, IEEE)

[1]Department of Electronics and Communication Engineering, Malaviya National Institute of Technology, Jaipur 302017, India
[2]Department of Information and Communication Technology, University of Agder, 4879 Grimstad, Norway

Corresponding author: Linga Reddy Cenkeramaddi (linga.cenkeramaddi@uia.no)

This work was supported in part by the International Partnerships for Excellent Education, Research and Innovation (INTPART) Program from the Research Council of Norway through the Indo–Norwegian Collaboration in Autonomous Cyber-Physical Systems (INCAPS) Project under Grant 287918.

ABSTRACT One of the leading causes of mortality worldwide is cardiovascular disease (CVD). Electrocardiography (ECG) is a noninvasive and cost-effective tool to diagnose the heart's health. This study presents a multi-class classifier for the prediction of four different types of Cardiovascular Diseases, i.e., Myocardial Infarction, Hypertrophy, Conduction Disturbances, and ST-T abnormality using 12-lead ECG. There are four key steps involved in the presented work: data preprocessing, feature extraction, data preparation, and augmentation, and modelling for multi-class CVD classification. The sixteen-time domain augmented features are used to train the classifier. The work is divided into three parts: extracting the features from raw 12-lead ECG signals, data preparation and augmentation, and training, testing, and validating the classifier. A comparative study of the performance of five different classifiers (i.e., Random Forest (RF), K Nearest Neighbors (KNN), Gradient Boost, Adda Boost, and XG Boost has also been presented. Accuracy, precision, recall, and F1 scores are used for performance evaluation. Further, the Receiver Operating Curve (ROC) is traced, and the Area Under the Curve (AUC) is calculated to ensure the unbiased performance of the classifier. The application of the proposed classifier in the Smart Healthcare framework has also been discussed.

INDEX TERMS Cardiovascular disease (CVD), PTB-XL data, machine learning, smart healthcare, ECG, heart failure, XG boost (XGB), random forest (RF), cat boost, K nearest neighbor (KNN), gradient boost (GB).

I. INTRODUCTION

According to the World Health Organisation, cardiovascular disease is the leading cause of premature mortality worldwide. It is estimated to be 31% of global deaths, which is around 17 million every year due to CVD [1]. In 2021, cardiovascular disease (CVD) was the largest cause of death globally, with low- and middle-income (LMIC) nations accounting for 4/5 of all CVD fatalities. According to a report by the World Heart Federation (WHF) published on May 20, 2023, deaths due to Cardiovascular Disease leaped from 12.1 million in 1990 to 20.5 million in 2021 [2]. In India, cardiovascular problems are very common, especially in those who are only slightly older than 45 [3].

Detection of Cardiovascular Disease (CVD) at an early stage is challenging due to the indistinguishable symptoms [4], [5]. The Electrocardiogram, often known as an ECG or EKG, is a unique graph showing the electrical activity of the heart from one moment to the next. The ECG specifically offers a time-voltage chart of the heartbeat. Due

The associate editor coordinating the review of this manuscript and approving it for publication was Wei-Wen Hu.

© 2023 The Authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License.
For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

to the crucial information it offers, the ECG is an important part of clinical diagnosis and treatment. Electrocardiogram (ECG) or its feature features are proven to be an effective method for early-stage prediction of CVD [6]. It is the non-invasive clinical standard method to analyze the proper functioning or dysfunction of the human heart [7]. It also is a biomarker for predicting cardiovascular abnormalities [8]. ECG can be captured by various methods, such as using single-lead, three-lead, five-lead, six-lead, and 12-lead. Among these 12 leads, ECG recording remains the gold standard for clinical practices. The 12-lead ECG is preferred in conventional clinical practice, as it gives a detailed idea about the heart, with the help of which not only dysfunction of the heart can be spotted, but the location of the point where the problem is occurring can also be identified [9]. It further helps the cardiologist to navigate further diagnosis or treatment.

Noncommunicable disease (NCD) is primarily described as a noninfectious disorder that develops gradually over time and is referred to as a chronic disease [10]. CVD also comes under the category of NCD, where timely diagnosis would certainly help in taking preventive measurements [11]. In a smart healthcare system, constant monitoring and telemedicine are vital in controlling such NCDs. People must intentionally try to manage themselves daily by employing various self-care devices [12]. The noninvasive technique is useful since it eliminates the pricking procedure in the body, which aids in continuous health monitoring. Consumer electronics have greatly improved the quality of life in smart healthcare, but precision and reliability are critical aspects for many applications.

Apart from the ECG, there are other methods like Photoplethesographgy (PPG), Electronic Heart Record (EHR) data, demographic features like age, sex, and weight, and behavioral features like smoking and alcohol consumption that can also be used in CVD prediction. Albeit sometimes it is seen that to improve the quality or accuracy of prediction, sometimes ECG is combined with other features like age, sex, weight, BMI, etc. The proposed CVD detection work is very important in the Smart Healthcare and Intelligent Devices era. This can be incorporated into an Internet of Things (IoT) or Internet of Medical Things (IoMT) using cloud services, where all the signal preprocessing and feature extraction will be done within the Python environment. It would contribute a lot to improving Quality of Life (QoL) and life expectancy.

The same idea is shown in Figure 1. The recorded ECG signal can be evaluated at the node device to make a prediction, or it can be sent over the cloud for evaluation, and the prediction can be transmitted directly to the health service provider or to the patients and caretakers.

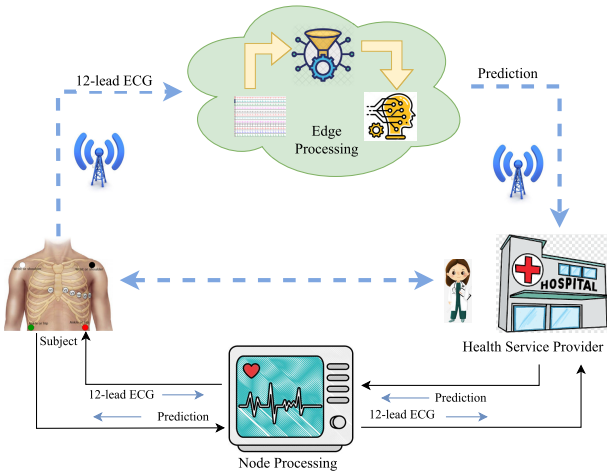


FIGURE 1. Smart healthcare framework for proposed model.

II. RELATED WORK

There is abundant work available in CVD detection using artificial intelligence, including machine learning and deep learning models. Usually, a supervised machine learning approach using ECG or its features, along with some demographic or laboratory features, is preferred for CVD classification. There is a significant amount of work in the field of unsupervised machine learning [13], [14], [15] that can be later utilized for CVD detection using different biomarkers. Gupta presented the performance of various supervised machine learning (ML) models like Decision Tree (DT), Random Forest (RF), K-Nearest Neighbor (KNN), and Support Vector Machine (SVM) to predict CVD [16]. Yang et al. [17] designed a Random Forest-based model for CVD prediction for the eastern China population. Chen et al. [18] presented the effect of consecutive moderately cold days on CVD Mortality in Shenzhen, China. Whereas Al-Absi et al. [19] have proposed a study to reveal the risk and co-morbidity associated with CVD for the Qatar population using machine learning algorithms that utilize the case-control study data. Molloy et al. [20] explored the challenges and scope of Implantable Medical Devices (IMD) and their potential for self-reporting innovative CVD healthcare technology. They also talked about the associated risk and required regulatory framework in commercialization. In addition to that, they have also listed all the IMD products that are recently approved, along with those in the pipeline. Many researchers used Cleveland data to predict the presence of CVD with different machine learning techniques, which include Logistic Regression, Random Forest, Naive Bayes, Bayes Net, Decision Tree, Support Vector Machine, K-Nearest Neighbor, and weighted KNN [21], [22], [23], [24]. Apart from the above traditional machine learning classifiers, ensemble machine learning methods were also proposed, i.e., the Hard voting ensemble method by Atallah et al. [25]. Miao et al. [26] described an Adaptive Boosting classifier to predict CVD with the help of the UCI repository heart disease database. While Basir et al. [27] has utilized memory-based learner DT-IG (Decision Tree Induction Based on Gini Index) and Ensemble of Naive Bayes and SVM to predict CVD using UCI repository data and Ricco database. Recently, Sarah et al. [28] utilized Cleveland data to compare the performance of different classifiers to predict the presence of CVD and proved that LR is best with

an accuracy of 85.25 %. Since ECG is a gold standard for level one diagnosis of CVD, PPG can be acquired easily. WH Ho et al. [29] proposed a novel method to transform the PPG signal into an ECG signal. Here the previously available research work was classified into two categories: ECG-based and Non-ECG-based. These have been discussed separately in the upcoming section and tabularized in Table 1.

A. ECG-BASED CVD DETECTION

ECG is an electrical impulse the heart's muscles produce during rhythmic contraction and relaxation. It reflects the dysfunction of the heart, like conduction disturbance, ischemia, or cardiovascular disease like arrhythmia, hypertrophy, stenosis, etc. It is an efficient and cost-effective method to evaluate the heart's health. It is also a crucial indicator for the early detection of cardiovascular disease. Considering the importance of ECG, many researchers have presented different studies related to the prediction CVDs [30], [31], [32], [33], [34], [35], [36], [37], [38], [39], [40]. Sadasivuni et al. [41] presented an ECG and Electronic Medical Record-based reservoir-computing and fusion model to predict ischemic heart disease, which is basically obstructed blood to different parts of the heart. Obayya et al. [42] designed a neural network-based DNN classifier to predict CVD, which utilizes Honey Badger Optimization for feature selection and Bayesian optimization for hyperparameter tuning. Guo et al. [43] identified the critical features in Coronary Artery Disease (CAD) and also predicted the CAD using Recursion Enhanced Random Forest with an Improved Linear Model (RERF-ILM). Mohan et al. [44] proposed a hybrid model (HRFLM), which utilizes the RF and linear methods to predict heart disease using the UCI Heart Disease dataset. Ghosh et al. [45] used the same data to extract features using LASSO and Relief techniques and then used those features to predict CVD.

B. NON-ECG BASED CVD DETECTION

In addition to the ECG, additional techniques are used to predict CVD, including photoplethysmography (PPG), data from electronic heart records (EHR), demographic factors, and behavioral variables [46], [47]. The frequency-aware Frequency attention LSTM (FA-Attn-LSTM) model is proposed by Park et al. [48] to predict CVD, which uses the frequency of the features present in Electronic Health Recor (EHR). Whereas body Mass Index (BMI) is identified as one of the important features for CVD prediction by Nikam et al. [49]. Spectral analysis-based features of Photoplethosography (PPG) signals like power density of low-frequency, high frequency, and their ratio are used to predict CVD at an early age by Simonyan et al. [50]. Qian et al. [51] has done a cohort study to predict CVD using routine physical examination indicators with the help of machine learning algorithms. Rahim et al. [52] has applied an ensemble ML Classifier of LR and KNN algorithms along with the SMOTE for data balancing to predict CVD accurately using the Framingham dataset. While Yang et al. [53] utilized the same data to an Optuna hyper-parameter tuned LightGBM classifier to predict coronary heart disease. Ghorashi et al. [54] utilized convenience sampling on UAE hospital data from 2621 entries and predicted the risk of acquiring CVD using PCA feature selection and LSTM model. They used SPSS to perform Simple LR and multiple LR for further analysis. Chicco et al. [55] used the EMR of 491 patients to analyze the CKD patients for the risk of CVD and also identified the variables that contribute most to CKD. Joo et al. [56] compared the performance of LR, DNN, Light GBM, and RF classifiers to estimate the risk of CVD in the cohort in 2 years and 10 years. They also performed the SHAP feature importance to look at which features contribute more to the risk of CVD development. Shuvo et al. [57] used a Phonocardiogram, i.e., heart sounds, to predict five different types of CVD with the help of CRNN while utilizing CNN and bi-LSTM to extract features.

C. NOVEL CONTRIBUTION

The presented work utilizes ECG-based features to predict CVD, and the following are the contributions of the proposed work:

- The proposed machine learning-based classifier utilizes only time-series features of twelve lead raw ECG signals for CVD classification.
- It is a reliable, accurate, and robust CVD detection method for real-time.
- The proposed classifier is lightweight and suitable for integration with any Internet of Things (IoT) based Smart Healthcare framework.
- The proposed model has an excellent recall, i.e. True Positive Rate (TPR) in ten-fold cross-validation, which indicates that it can effectively distinguish the four main kinds of CVD from the control group.

III. METHODS

The proposed method is a highly accurate CVD prediction model which includes five classes. Figure 2 shows the breakdowns of the functioning of the proposed model. Each step is explained in the subsections below.

A. DATA

In this study, Physionet's PTB-XL public electrocardiography data set is utilized [58], [59]. It contains 12 lead ECG recordings (i.e. (V1, V2, V3, V4, V5, V6, I, II, III, aVL, aVR, and aVF) of 21837 records that have been taken from 18885 subjects, and each recording is ten second long. The ECG data is a multi-label data set as up to two cardiologists annotated it. Later, it was aggregated as diagnostic super and subclass. The five superclasses are Conduction Disturbances (CD), ST/T change (STTC), myocardial infarction (MI), hypertrophy (HYP), and normal ECG (NORM). A brief description of each of the four different kinds of CVD is as follows: Conduction Disturbances (CD): A conduction disturbance (CD) or disorder is a condition of the heart with a

TABLE 1. Existing work related to CVD prediction.

Work related to CVD prediction using ECG or ECG features			
Reference	Data	Features	Algorithm(s)
Subbramani et al., 2023 [30]	Cleveland, Hungarian, Switzerland, Long Beach VA, and Stalog (Heart) and Irvine (UCI) datasets	GBDT selected features	RF, LR, MLP, ET, and Cat-Boost
Alqahtani et al, 2023 [31]	Kaggle repository's Cardiovascular Disease dataset	age, sex, BP, cholesterol, ST slope, heart rate, angina, no. of major vessels	RF, KNN, DT, XGB, DNN, KDNN
Modaket et-al, 2022 [32]	Cleveland, Hungarian, Switzerland, Long Beach, and Statlog	age, sex, BP, cholesterol, ST slope, heart rate, angina, no. of major vessels	Multilayer Perceptron
Likith Reddy et al, 2021 [33]	PTB-XL	Raw data	IMLE-Net (CNN)
Ibrahim Patel et al, 2021 [34]	Methods and explanation	P,T waves and QRS complex	NA
Chen, Xiehui et al, 2021 [35]	PTB-XL	Raw data	ResNet (CNN)
Smisek, Radovan et al, 2020 [36]	MIT-BIH PTB Diagnostic ECG Database	QRS complex, P wave, T wave	DT
Hu, Yusong et al, 2020 [37]	MIT-BIH	R peak, RR interval	Decision Tree
Alfaras et al, 2019 [38]	MIT-BIH	RR Interval	Random Forest
Roopa et al., 2017 [39]	Kaggle Cardiovascular Dataset	P wave, PR Interval, ST interval, T wave	SVM, DT
Sadiq et al, 2013 [40]	MIT-BIH	P, Q, R, S, T waves	SVM, DT
Work related to CVD prediction using Non-ECG me			
Xin Qian et-al, 2022 [51]	Private Data collected via questionnaire, physical and laboratory examination	Routine physical examination based features	LR, SVM, RF, Adda boost
Akkaya et-al, 2022 [46]	Survey data and Medical history	demographic features and medical history	LR, SVM, KNN, DT, NB, Ada Boost, MLP, and XGB
Simoyan et-al, 2021 [50]	Private PPG data of 620 subjects	Spectral analysis based features of PPG	statistical analysis using STATISTICA 12.0 (StatSoftInc, USA)
Rogers et al, 2019 [47]	developed a biomimetic cardiac tissue chip (CTC) model	studied the effect of pressure and volume overload	CTC can be utilized to construct highly relevant models for cardiovascular disease modeling that properly mimics hemodynamic loading and unloading.
Bashir et-al, 2014 [27]	UCI repo and Ricco database	SPECT (single proton emission computed tomography) images and demographic features	Memory based learner, DT-IG, DT-GI, Ensemble of Naive Bayes and SVM
Samiul et-al, 2021 [57]	Physionet Cinc challenge 2016 dataset	Time-invariant and temporal features	CRNN (Convolutional Recurrent Neural Network

block in conduction pathways. Conduction disorders lead to chronic heart failure. Hypertrophy (HYP): It is a condition in which the heart's muscles thicken, which reduces the heart's pumping capability. ST/T Change (STTC): A deviation in the pattern of the ST/T wave indicates different abnormalities, such as ischemia, i.e., reduced blood flow, hence reduced oxygen. This class includes ischemia in anterior and inferior leads and specific and non-specific ST changes. Myocardial Infarction (MI): It is commonly known as a heart attack or extreme medical emergency. It happens when blood in the coronary artery tends to cease. This class includes anterior, inferior, lateral, and posterior myocardial infarctions. Each superclass has some subclass (total 24) except NORM, listed in Table 2.

The raw ECG signal of the database was recorded between 1989 to 1996 using a device by *Schiller* AG. Furthermore, Physikalisch Technische Bundesanstalt (PTB) curates and transforms this information into an organized database. The dataset has 52% recording of males and 48% of a female whose age lies in the range 0 to 95 years. The distribution of data within the superclasses is shown in Table 3.

B. DATA PREPROCESSING

The original ECG recording was converted to binary with 1 V/LSB resolution and 16-bit precision. Using an on-and-off technique, the signal's starting and end spikes were eliminated and are available for download at a 100 Hz sampling rate. The data is available in two different portfolios. One folder has a raw 10-second ECG recording of each subject, while the other file has a scp statement. The scp statements have information on patient demography, i.e. age, sex, height, weight, etc., and diagnostic class and diagnostic superclass. Each ECG record is linked to a Python diagnostic superclass that is considered a label. The data is downloaded from physionet.org [60] in .mat format. They are loaded in a Python environment with the help of the WFDB toolbox [61] and

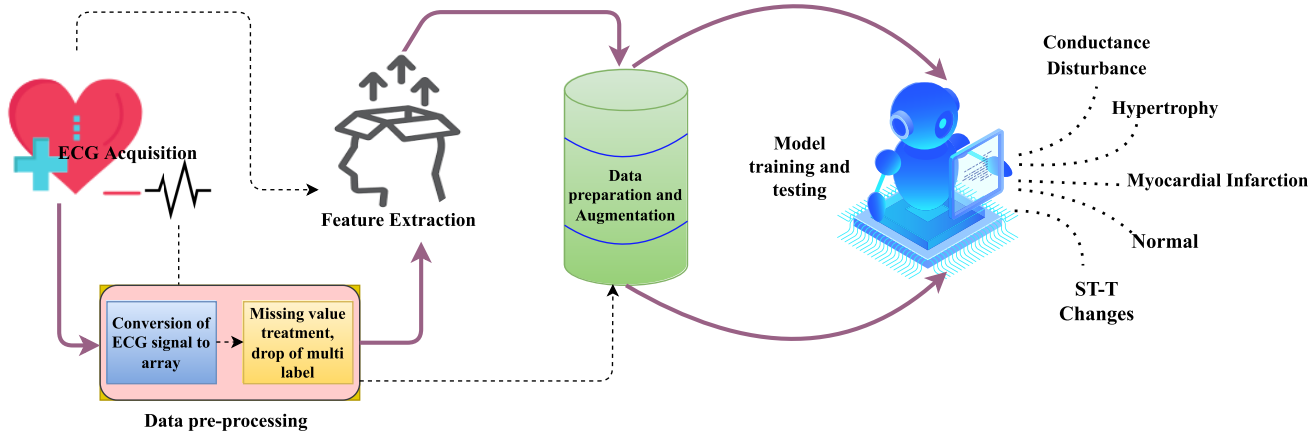


FIGURE 2. Overall process flow of the work.

TABLE 2. Diagnostic superclass and subclass of PTB-XL data.

Super-classes	Sub-classes	Abbreviation
CD	Left Anterior/Left Posterior Fascicular Block	LAFB/LPFB
	Incomplete Right Bundle Branch Block	IRBBB
	Incomplete Left Bundle Branch Block	ILBBB
	Complete Left Bundle Branch Block	CLBBB
	Complete Right Bundle Branch Block	CRBBB
	AV block	_AVB
	Non-specific intraventricular conduction disturbance (block)	IVCB
	WPW Wolff-Parkinson-White Syndrome	WPW
HYP	Left Ventricular Hypertrophy	LVH
	Right Ventricular Hypertrophy	RHV
	left atrial overload/enlargement	LAO/LAE
	Right Atrial Overload/Enlargement	RAO/RAE
	Septal Hypertrophy	SEHYP
MI	anterior myocardial infarction	AMI
	Inferior Myocardial Infarction	IMI
	Lateral Myocardial Infarction	LMI
	Posterior Myocardial Infarction	PMI
STTC	Ischemic in anterior leads	ISCA
	Ischemic in Inferior Leads	ISCI
	Ischemic (non-specific)	ISC
	ST-T changes	STTC
	Non-specific ST changes	NST
NORM	Normal ECG signal	–

TABLE 3. Distribution of diagnostic superclasses.

Super-classes	Description	Records
NORM	Normal ECG	9528
CD	Conduction Disturbance	4907
STTC	ST or T change	5250
HYP	Hypertrophy	2655
MI	Myocardial Infarction	5486

scipy. One random sample of raw ECG signal is given in Figure 3.

Later, it is converted into a three-dimensional array of size (21837,1000,12) using Numpy, which is later flattened to (21837,12000) for further processing. Label encoding is applied to the Diagnostic superclass to convert the string data type to numeric. The samples with more than one label are eliminated from the presented study, as they fall into more than one class at a time. At the same time, data with no labels

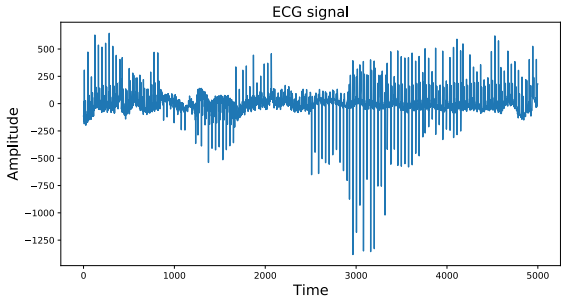


FIGURE 3. Sample of raw ECG signal.

TABLE 4. List of features.

Feature Category	Feature's Name
Peak's Positions	P-location, Q-location, R-location, S-location, T-location
Peak's amplitude	P-amplitude, Q-amplitude, R-amplitude, S-amplitude, T-amplitude
Intervals	PQ, ST, QT, PR, RR, QRS

are also dropped. Missing values in the data are filled using the forward fill ('fill') technique.

C. FEATURE EXTRACTION

The raw ECG signal has 21000 dimensions, and predicting the different cardiovascular diseases with the help of that would be complicated. Hence, time-series features are extracted to predict different CVDs. The time series features of the ECG signal are extracted using the neurokit2 toolbox [62], which can broadly be classified into three groups:

- Peak Location : P wave, Q-R-S complex, T wave
- Peak Amplitudes : P wave,Q,R,S waves,T wave
- Time intervals: PQ, ST, QT, PR, RR, QRS intervals

The explicit list of features is provided in Table 4.

The particular set of time-series features is selected as they are the basic features of the ECG and represent the electrical activity of the heart, hence the abnormality or disorder of the heart [6].

TABLE 5. Label encoding of different classes.

Acronym	CVD Class	Label
NORM	Normal	0
MI	Myocardial Infarction	1
STTC	ST/T Change	2
HYP	Hypertrophy	3
CD	Conduction Disturbances	4

To find the peak location indices, first, we need to find the R peak location; by using that R peak location, all the other peaks, i.e., P, Q, S, T, are found. Neurokit is used to delineate the ECG peaks in Python. The ecg peaks() function returns a dictionary that contains the samples where peaks are found. Further, to segment the QRS complex **ecg delineate ()** is used. One of the samples of peak detection is shown in Figure 4. After extracting the peak locations and peak amplitudes, time intervals between the peaks are extracted using the locations of the respective peaks. The graph for all the extracted intervals is shown in Figure 5.

D. DATA PREPARATION

Each subject's sixteen features are extracted as a list within the Pandas data frame. Later on, these features are expanded in the form of a three-dimensional Numpy array. For further processing, that array was flattened. The resulting array is used as input for the CVD classifier(s), where training data has 80 %of the total data while testing data has 20 %. The compilation of the data presented two significant difficulties: first, at some points, all the features for the particular subjects are NaN; second, the length of extracted features for all the ECGs is not the same, which means all the recorded signals did not have equal no. of peaks. To deal with the missing values in the extracted feature, the Forward Fill (ffill) method has been implemented at this stage. Next, the length of extracted features for each ECG signal is equated. Finally, the set of all the n features for each subject is combined with respective labels with the help of subject IDs. The labels or target variables were in string format; hence, it is converted to the numeric data type by applying label encoding. The same is given in Table 5. In the upcoming sections, the process of data augmentation and modeling are explained.

E. DATA AUGMENTATION

After preparing the data, it was observed that the number of samples in each of the five categories was different. As the ML algorithms do not address the class distribution, this class-imbalance data will adversely affect the classifier's performance. Standard machine learning approaches tend to predict merely the majority class, favoring the majority class and ignoring the minority, and miss-classifying the minority significantly compared to the huge majority. In more technical language, if our dataset's data distribution is imbalanced, the model seems more susceptible to situations where the minority class has little or no recall. Hence, to avoid this imbalance, data augmentation is performed to avoid this. The process of data augmentation includes the oversampling of the minority class in the training data. To up-sample the minority class, SMOTE (Synthetic Minority Over-sampling Technique) is used here. It is an oversampling approach for balancing the distribution of classes in the dataset. It chooses minority instances that are near the feature space. Then, it creates a line in the features space between the examples and draws a new sample at a location along that line. Simply put, the method chooses a random example from the minority class and a random neighbor using K Nearest Neighbours, and in the feature space, a synthetic example is formed by combining two instances. The process of data augmentation and feature scaling is done only on 80 % of training data, which are selected randomly in each fold of the 10-fold cross-validation process.

F. MODELING

There are plenty of machine learning algorithms for classification purposes. Although here in this presented work, five machine learning algorithms have used, which are KNN [63], RF [64], XG-Boost [65], Gradient Boost [66], and Cat-Boost [67]. These five classifiers are chosen according to the previous literature review. The reason behind using more than One ML algorithm is to achieve better and more reliable prediction results. The core difference in the algorithm of the different classifiers is discussed below. KNN calculates the distances between the data point and different classes in an n-dimensional plane where n is the number of features and assigns the data to the category that is closest to the existing categories; here, in this case, it is five. RF is an ensemble machine learning method that uses the concept of Bagging, i.e., aggregation of bootstraps to perform classification. RF is a collection of decision trees that train on the random selection of original data, and the decision is made on the majority voting. In contrast, all the other three algorithms, i.e., XG Boost, Gradient Boost, and Cat Boost, work on the concept of boosting. The idea of boosting is to improve the poor learner instead of focusing on the best. However, they are also ensemble machine learning methods. It uses the sequencing approach, which means at any moment t, the model results are weighted depending on the results of the preceding instant (i.e., t-1). The correctly predicted outcomes are given less weight, while those incorrectly classified are given more. In order to increase prediction accuracy, it assembles a group of weak learners. The prediction results for each classifier are discussed in section IV. The specific performance metrics used to evaluate the classifiers are precision, recall, F1 score, and accuracy. Further, the ROC-AUC score is calculated to ensure the unbiased performance of the classifier.

G. PROPOSED SMART HEALTHCARE FRAMEWORK FOR CVD PREDICTION

The proposed machine learning model for CVD prediction can be deployed on a central cloud server or at the edge of a network. It can also be deployed as a mobile application hosted through any of the cloud services like IaaS

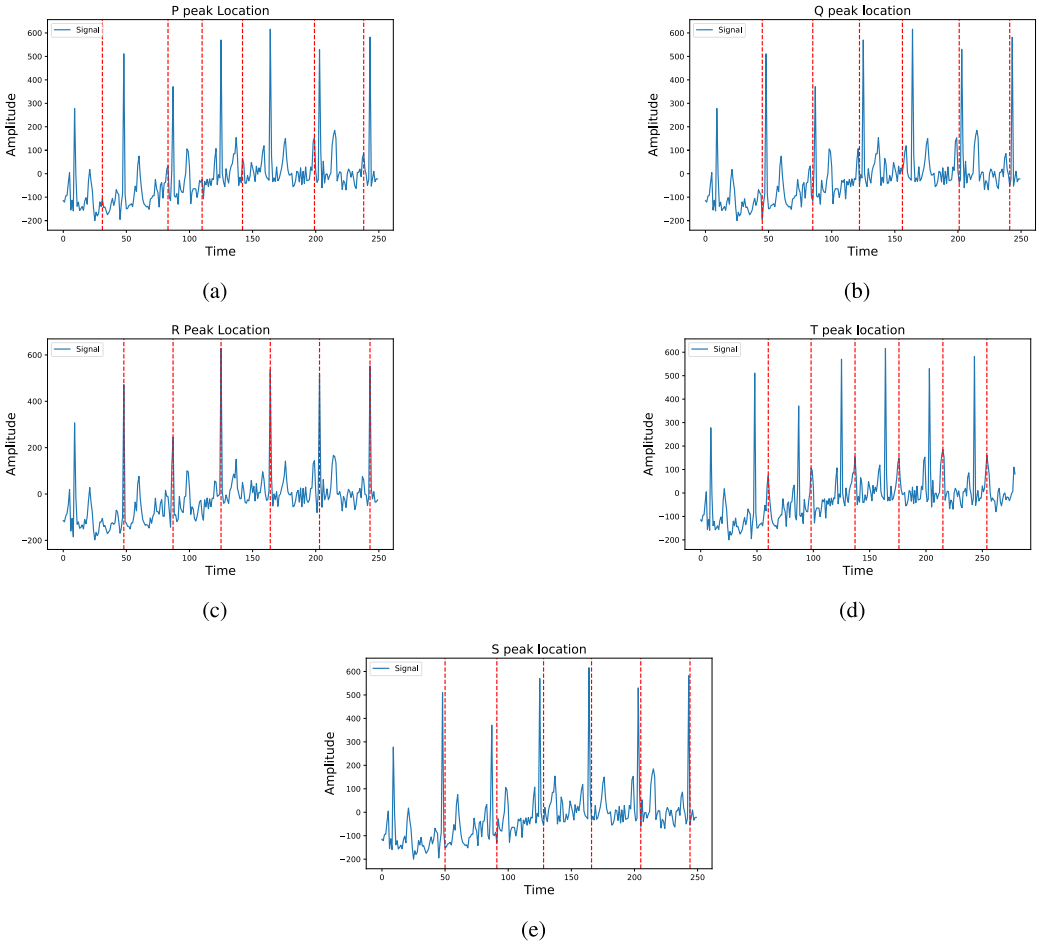


FIGURE 4. (a), (b), (c), (d) and (e) are showing the location of P, Q, R, T and S peaks respectively in a random sample of ECG signal.

(Infrastructure as a Service), PaaS (Platforms as a Service) o, or SaaS (Software a Service). A virtual computer can be launched using any cloud services, and a user can set up the Python environment and install Neurokit there in a similar manner as we do on our usual computers. Deploying the model on the edge is only recommended when the data stream size is large or very low latency is required. Various cloud platforms offer edge deployment, such as Oracle Cloud [68]. Models deployed at the edge usually have the ONNX format, which gets updated through the central cloud management server. All the packages needed to preprocess the data, like Neurokit, are installed within the Python virtual environment hosted by selected cloud services. The above-discussed framework is displayed in Figure 6. It's important to mention that the main implication of implementing such a smart healthcare framework is maintaining user data privacy as it holds personal health information.

IV. RESULT AND DISCUSSION

Each of the five classification models is trained with 80 % of the training data and tested with the remaining 20 %. Following is the list of performance measures that are used for the assessment of the classifier's performance:

TABLE 6. Performance metrics of KNN.

Class	Precision	Recall	F1 Score
NORM (0)	85	54	56
MI (1)	74	97	84
STTC (2)	77	95	85
HYP (3)	93	100	96
CD (4)	87	96	91

TABLE 7. Performance metrics of RF.

Class	Precision	Recall	F1 Score
NORM (0)	81	86	84
MI (1)	91	86	88
STTC (2)	88	89	89
HYP (3)	99	99	99
CD (4)	95	93	94

- Confusion Matrix: It is a n x n tabular representation of all the classified instances. It includes four entries: TP (True Positives): Correctly predicted CVD instance TN (True Negative): Correctly identified Normal subjects or instances FP (False Positive): Normal instances classified as any one of the CVD classes FN (False Negative): CVD instances classified as Normal

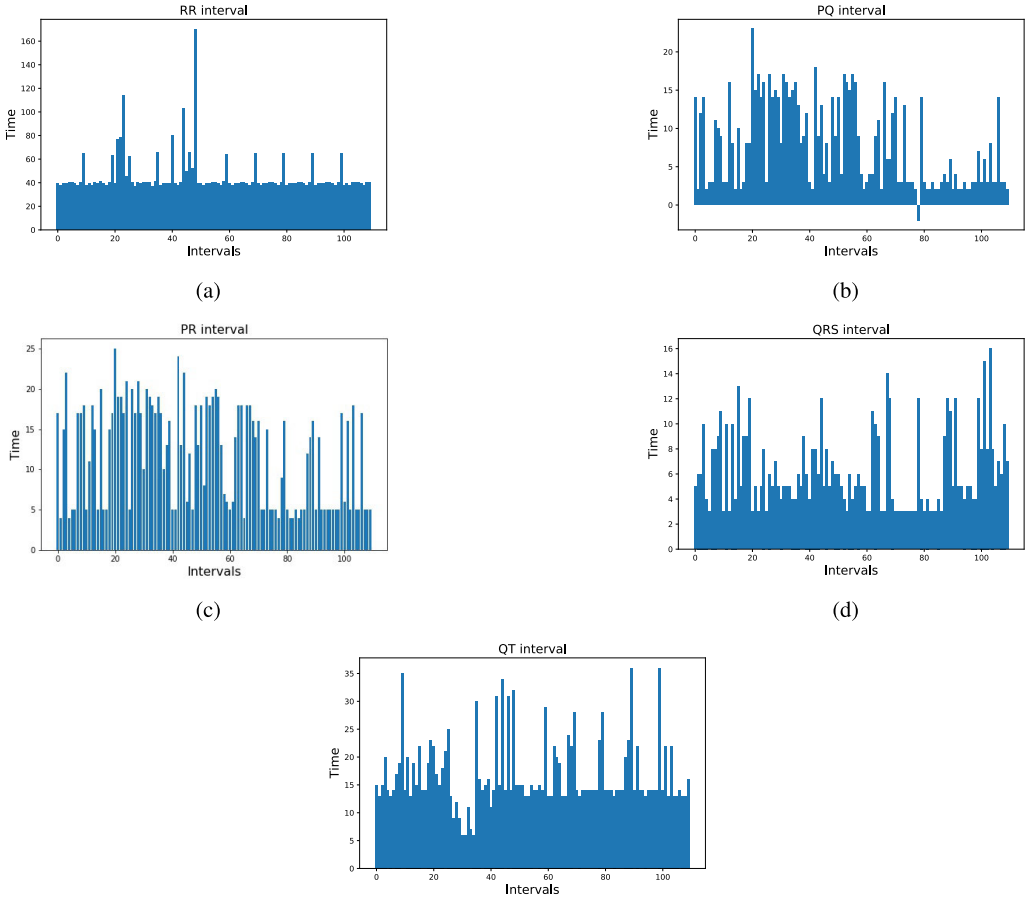


FIGURE 5. (a), (b), (c), (d), and (e) are showing the RR, PQ, PR, QRS, and QT intervals respectively in the random sample of ECG signal.

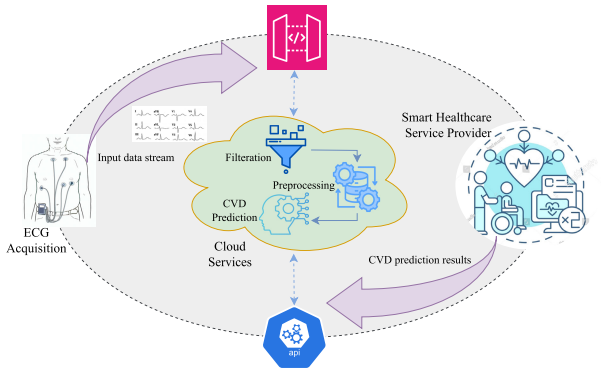


FIGURE 6. Proposed smart healthcare framework.

- Accuracy: it gives us the ratio of total correct prediction to total predictions made [69]. The formula to calculate accuracy is given in Equation 1.

$$Acc = \frac{(TP + TN)}{(TP + TN + FP + FN)} \times 100 \quad (1)$$

- Precision: It provides the ratio of True Positives to total Positive predictions, as given in equation 2.

$$Precision = \frac{TP}{(TP + FP)} \quad (2)$$

- Recall: It provides the ratio of accurately predicted CVD class cases to real CVD occurrences, and the same is written in equation 3.

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

- F1 Score: It correctly assesses precision and recall by considering their conflicting characteristics. It is calculated as; twice as much of a difference between the accuracy and recall products and their aggregate as given in equation 4.

$$F1 - Score = \frac{2 * Precision * Recall}{(Precision + Recall)} \quad (4)$$

- ROC-AUC score stands for Receiver Operating Characteristics (ROC) and Area Under Curve (AUC). False Positive Rate (FPR) and True Positive Rate (TPR) are represented by this curve. The score ranges between 0 and 1. Where one indicates the best case, and 0 demonstrates the worst. If its value lies between 0.5 to 1.0, the classifier will likely distinguish between Positive (i.e., all instances of CVD) and Negative classes (i.e., Normal means no CVD or healthy).

TABLE 8. Five-class classification matrix.

		Prediction(s)				
		0	1	2	3	4
Actual(s)	0	**cell 1**	cell 2	cell 3	cell 4	cell 5
	1	cell 6	**cell 7**	cell 8	cell 9	cell 10
	2	cell 11	cell 12	**cell 13**	cell 14	cell 15
	3	cell 16	cell 17	cell 18	**cell 19**	cell 20
	4	cell 21	cell 22	cell 23	cell 24	**cell 25**

TABLE 9. Performance metrics of XG boost.

Class	Precision	Recall	F1 Score
NORM (0)	80	94	86
MI (1)	95	87	91
STTC (2)	95	91	93
HYP (3)	99	99	99
CD (4)	98	93	96

TABLE 10. Performance metrics of gradient boost.

Class	Precision	Recall	F1 Score
NORM (0)	74	86	82
MI (1)	86	85	84
STTC (2)	89	81	87
HYP (3)	97	95	96
CD (4)	91	89	90

TABLE 11. Performance metrics of cat-boost.

Class	Precision	Recall	F1 Score
NORM (0)	76	88	81
MI (1)	87	84	83
STTC (2)	88	80	86
HYP (3)	97	97	97
CD (4)	91	89	90

The confusion matrix for the multi-class classification differs from the binary classification's confusion matrix as we do not get the TP, TN, FP, and FN directly from it. A dummy Confusion Matrix for five class classifications is shown in Table 8.

The matrix diagonal represents the TP, e.g. cell 1, 7, 13, 19, and 25. Each cell represents the TP of the corresponding class (i.e. column). For example, cell 7 represents the TP for class 1. Although, FN is the sum of all the values of the corresponding row except the TP. For example, FN for class 1 will be the sum of cell 6, cell 8, cell 9, and 10. Similarly, FP is the sum of all the values of the corresponding column except the TP, i.e., FP for class 1 will be the sum of cell 2, cell 12, cell 17, and cell 22. Whereas TN for a specific class is the sum of all the cell values except the row and column associated with that particular class, i.e. if we want to calculate the TN for class 0, then it will be the sum of cells 2,3,4,5,7,8,9,10,12,13,14,15,17,18,19,20,22,23,24 and 25, or in other words sum of all the cell corresponding to raw 0 and column 0 i.e, cell 1-5 and cell 6, 11, 16, and 21. The Confusion Matrices for all the five classifiers with 20 % of testing data are given in Figure 7.

Table 12 lists the performance for all five classifiers. The performance of XG boost is best in terms of overall accuracy, precision, and F1 score, while RF is best regarding Recall.

Looking at the performance measures of all five classifiers, it is evident that in terms of accuracy, XG Boost is best,

TABLE 12. Comparison of classifiers' Performance.

Metrics	XG Boost	RF	Cat-Boost	G-Boost	KNN
precision	92.0	91.0	89.0	85.0	83.0
Recall	90.0	92.0	86.0	87.0	82.0
F1 Score	93.0	92.0	89.0	87.0	83.0
Accuracy	93.0	91.0	89.0	88.0	82.0

TABLE 13. Overall comparison of performance with existing work.

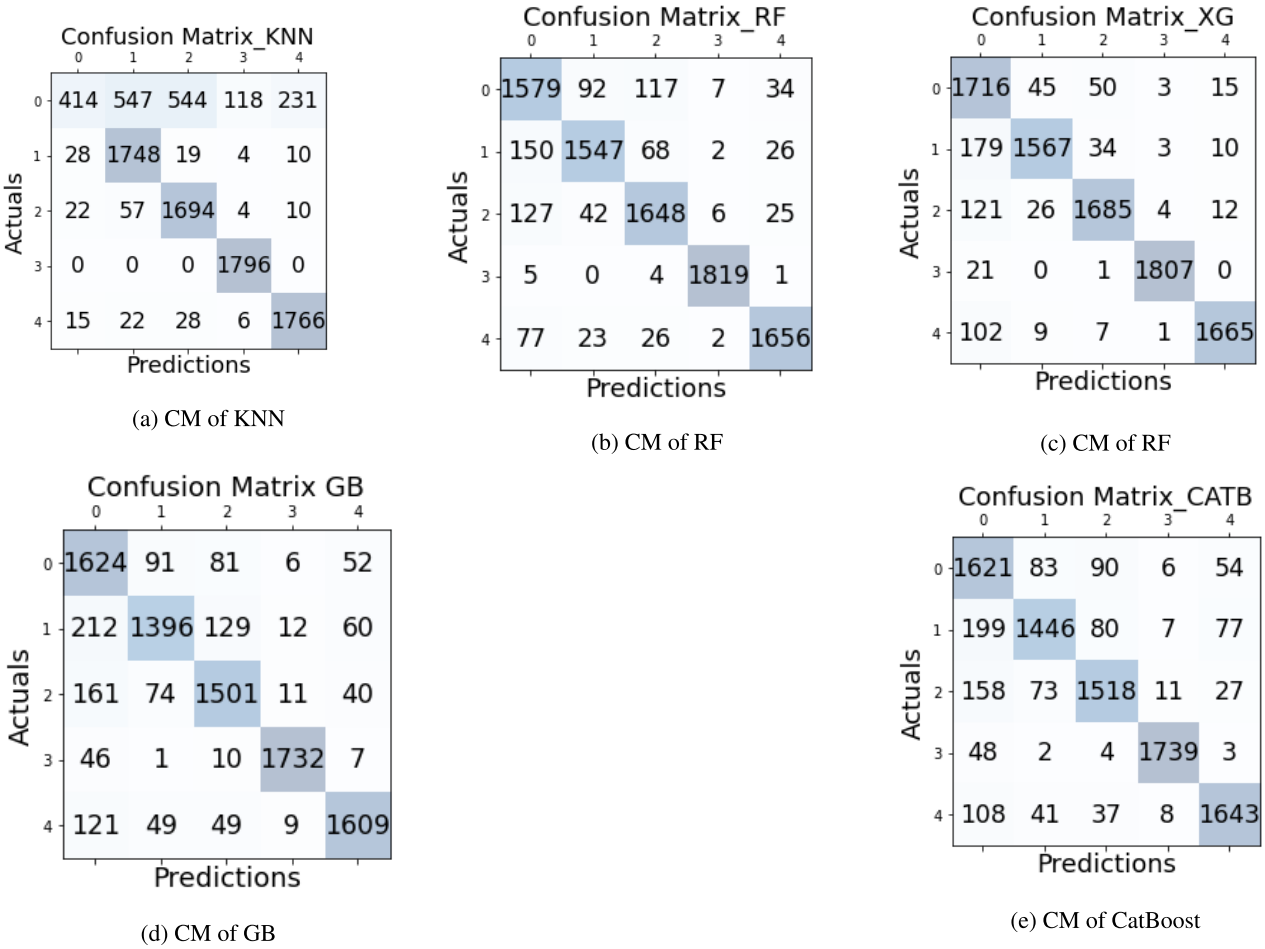
Reference	Data used	F1 Score	ROC-AUC Score	Accuracy (%)
Bashir et-al, 2014 [27]	UCI repo and Ricco database	0.79	NA	86.82
Murugesan et-al, 2018 [71]	MIT-BIH	0.78	0.91	87.25
Mausavi et-al, 2019 [72]	MIT-BIH	0.74	0.86	84.12
Rajpurkar et-al, 2019 [73]	MIT-BIH	0.79	0.92	87.92
Reddy et-al, 2021 [33]	PTB-XL	0.82	0.91	88.85
Smigiel et-al, 2021 [74]	PTB-XL	0.61	0.87	87.7
Nguyen et-al, 2021 [21]	Cleveland	0.85	NA	83.5
Modaket et-al, 2022 [32]	Cleveland, Hungarian, Switzerland, Long Beach, and Statlog	0.87	NA	87.70
Akkaya et-al, 2022 [46]	Survey data and medical history	NA	0.90	84.63
Alqahtani et-al, 2023 [31]	Kaggle Cardiovascular Disease dataset	0.88	0.92	88.65
Presented work	PTB-XL	**0.96**	**0.94**	**93.0**

While if we look at the Recall Random Forest, it is the best. It indicates that the Boosting-based technique works best in terms of overall accuracy. However, for recall, the bagging technique works best. Ten-fold cross-validation is used here for validation purposes. The mean of 10-fold cross-validation for all five classifiers' performance is listed in Table 6, 7, 9, 10, and 11. Apart from the accuracy, precision, and recall, ROC-AUC is also a crucial parameter that indicates the classifier's unbiased. In other words, it is also known as a degree or measure of separability. It tells how much the model is capable of distinguishing between classes The class-wise ROC-AUC for XG Boost is given in Figure 8.

The value of the ROC-AUC score lies between 0.0 to 1.0 [70]. A value between 0.5 to 1.0 is supposed to be good for the classifier. Here, as we can see, the AUC-ROC score is above 0.90. It is the best for 'Hypertrophy,' i.e., 1.00, followed by 'Conduction Disturbance,' for which it is 0.99. For Myocardial Infarction, ST-T abnormality, and Normal class, it is 0.98.

Further, the comparison between the presented work and a few of the existing literature in terms of overall accuracy in this field is made in Table 13.

It is evident from the table that the presented classifier is better not only in terms of accuracy but also in terms



(a) CM of KNN

(b) CM of RF

(c) CM of RF

(d) CM of GB

(e) CM of CatBoost

FIGURE 7. Confusion Matrices of all the five classifiers.

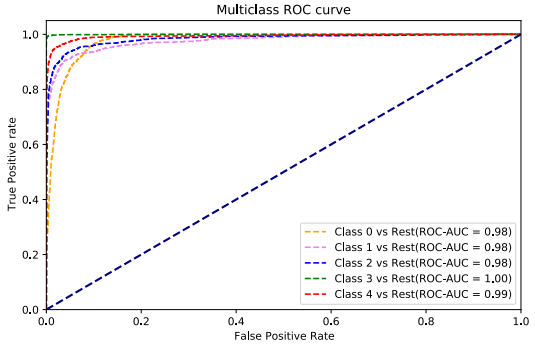


FIGURE 8. ROC and AUC score for XG boost.

of ROC-AUC score and F1 score. The higher value of the ROC-AUC score validates the unbiased classification of the classifier. Although the comparison is not made in terms of Recall, as the value of Recall was not available for most of the previous work, the presented classifier has an excellent value of Recall which is 0.90.

V. CONCLUSION AND FUTURE SCOPE

The paper presented a 12-lead ECG-based multi-class CVD classifier that predicts four main types of CVD (i.e., conduction disorders, hypertrophy, myocardial infarction, and ST/T change. It utilizes time-series features extracted from raw ECG signals. The proposed classifier is accurate, robust, and efficient. It utilizes only 16 Time domain features to classify all five categories. Since the raw signal was of 1000*12 dimensions are reduced to sixteen, the proposed classifier is very light and fast. Further, data augmentation followed by 10-fold cross-validation is performed to ensure that the classifier's performance is not biased. The ROC-AUC score indicates that the proposed classifier can distinguish between all five classes, including four main kinds of CVDs (i.e., Myocardial Infarction, Hypertrophy, Conduction Disturbances, and ST-T abnormality), against the control group. Here, a comparative analysis of all five classifiers' performances has also been done, and it is concluded that ensemble classifier XG Boost is the best in terms of overall performance. The proposed classifier's accuracy, precision, recall, and F1 score are 93.0%, 92.0%, 90.0%, and 93.0%, respectively. The paper also describes the role of the proposed classifier in smart Healthcare. The lightweight, accurate, and robust technology is crucial for an effective IoT-based Smart Healthcare framework. The potential limitation of the proposed work is that the subject or user needs to

wear a device like a Holter monitor to record the 12-lead ECG for continuous monitoring of CVD. In the future, the usability of single-lead [75] and three-lead for CVD prediction can be explored. The machine learning models or framework to predict CVD with the help of 1-lead or 3-lead ECG instead of 12-lead ECG can be developed [76], [77]. Several wearable devices in the market capture either 3-lead or 1-lead ECG, which can be useful for designing CVD prediction systems with machine learning models. It is also required to design a CVD protection system through raw signals. If we can achieve comparable performance, then this technology will be more accessible to common people and can bring change in CVD management and improve quality of life.

The main abbreviations used in this manuscript are given below:

- **AUC:** Area Under Curve
- **BMI:** Body Mass Index
- **CAD:** Coronary Artery Disease
- **CKD:** Chronic Kidney Disease
- **CVD:** Cardiovascular Diseases
- **CD:** Conduction Disturbances
- **CNN:** Convolutional Neural Network
- **CRNN:** Convolutional Recurrent Neural Network
- **LSTM:** Long short-term memory
- **DNN:** Deep Neural Network
- **ECG:** Electrocardiogram
- **EMR:** Electronic Medical Record
- **GBDT:** Gradient Boosted Decision Tree
- **HYP:** Hypertrophy
- **IMD:** Implantable Medical Device
- **KNHSC:** Korean Health Insurance Service national health Sample Cohort data
- **MI:** Myocardial Infarction
- **ML:** Machine Learning
- **MLP:** Multi-Layer Perceptron
- **NB:** Naive Bayes
- **PCA:** Principal Component Analysis
- **ROC:** Receiver Operating Curve
- **SMOTE:** Synthetic Minority Over-Sampling
- **SPSS:** Statistical Package for Social Sciences
- **STTC:** ST-T Changes
- **WHO:** World Health Organisation

REFERENCES

[1] WHO. (2023). *Cardiovascular Diseases*. Accessed: Jun. 28, 2023. [Online]. Available: https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1/

[2] WHF. (2023). *Deaths From Cardiovascular Disease*. Accessed: Sep. 21, 2023. [Online]. Available: https://world-heart-federation.org/news/deaths-from-cardiovascular-disease-surged-60-globally-over-the-last-30-years-report/

[3] M. Bhatia, P. Dixit, M. Kumar, and L. K. Dwivedi, "Impending epidemic of cardiovascular diseases among lower socioeconomic groups in India," *Lancet Healthy Longevity*, vol. 2, no. 6, pp. e314–e315, Jun. 2021.

[4] N. Sinha, T. Jangid, A. M. Joshi, and S. P. Mohanty, "iCARDO: A machine learning based smart healthcare framework for cardiovascular disease prediction," 2022, *arXiv:2212.08022*.

[5] N. Sinha, A. M. Joshi, and S. P. Mohanty, "iCardo 2.0: A smart healthcare framework for cardiovascular disease accurate prediction by using T-wave morphology of ECG," in *Proc. IEEE Int. Symp. Smart Electron. Syst. (iSES)*, Dec. 2022, pp. 343–348.

[6] A. Goldberger, *Goldberger's Clinical Electrocardiography*. Amsterdam, The Netherlands: Elsevier, 2018.

[7] M. B. Abubaker and B. Babayiğit, "Detection of cardiovascular diseases in ECG images using machine learning and deep learning methods," *IEEE Trans. Artif. Intell.*, vol. 4, no. 2, pp. 373–382, Apr. 2023.

[8] N. Sinha and A. M. Joshi, "Predicting the presence of left ventricular hypertrophy using este's criteria with support vector machine," Res. Square, Malaviya Nat. Inst. Technol. (MNIT), Jaipur, India, 2022.

[9] L.-H. Wang, Y.-T. Yu, W. Liu, L. Xu, C.-X. Xie, T. Yang, I.-C. Kuo, X.-K. Wang, J. Gao, P.-C. Huang, S.-L. Chen, W.-Y. Chiang, and P. A. R. Abu, "Three-heartbeat multilead ECG recognition method for arrhythmia classification," *IEEE Access*, vol. 10, pp. 44046–44061, 2022.

[10] A. M. Joshi, P. Jain, and S. P. Mohanty, "IGLU 3.0: A secure noninvasive glucometer and automatic insulin delivery system in IoMT," *IEEE Trans. Consum. Electron.*, vol. 68, no. 1, pp. 14–22, Feb. 2022.

[11] B. Zhang, L. Zhu, Z. Pei, Q. Zhai, J. Zhu, X. Zhong, J. Yi, and T. Liu, "A framework for remote interaction and management of home care elderly adults," *IEEE Sensors J.*, vol. 22, no. 11, pp. 11034–11044, Jun. 2022.

[12] P. Jain, A. M. Joshi, N. Agrawal, and S. Mohanty, "iGLU 2.0: A new non-invasive, accurate serum glucometer for smart healthcare," 2020, *arXiv:2001.09182*.

[13] M. Li, P.-Y. Huang, X. Chang, J. Hu, Y. Yang, and A. Hauptmann, "Video pivoting unsupervised multi-modal machine translation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 3, pp. 3918–3932, Mar. 2023.

[14] L. Zhang, X. Chang, J. Liu, M. Luo, Z. Li, L. Yao, and A. Hauptmann, "TN-ZSTAD: Transferable network for zero-shot temporal activity detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 3, pp. 3848–3861, Mar. 2023.

[15] C. Yan, X. Chang, Z. Li, W. Guan, Z. Ge, L. Zhu, and Q. Zheng, "ZeroNAS: Differentiable generative adversarial networks search for zero-shot learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 12, pp. 9733–9740, Dec. 2022.

[16] J. Gupta, "The accuracy of supervised machine learning algorithms in predicting cardiovascular disease," in *Proc. Int. Conf. Artif. Intell. Comput. Sci. Technol. (ICAICST)*, Jun. 2021, pp. 234–239.

[17] L. Yang, H. Wu, X. Jin, P. Zheng, S. Hu, X. Xu, W. Yu, and J. Yan, "Study of cardiovascular disease prediction model based on random forest in Eastern China," *Sci. Rep.*, vol. 10, no. 1, p. 5245, Mar. 2020.

[18] R. Chen, F. Miao, J. Zheng, Y. Wu, and Y. Li, "Effects of consecutive moderately cold days on cardiovascular disease mortality in Shenzhen, China: A preliminary study," in *Proc. 40th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2018, pp. 1148–1151.

[19] H. R. H. Al-Absi, M. A. Refaee, A. U. Rehman, M. T. Islam, S. B. Belhaouari, and T. Alam, "Risk factors and comorbidities associated to cardiovascular disease in qatar: A machine learning based case-control study," *IEEE Access*, vol. 9, pp. 29929–29941, 2021.

[20] A. Molloy, K. Beaumont, A. Alyami, M. Kirimi, D. Hoare, N. Mirzai, H. Heidari, S. Mitra, S. L. Neale, and J. R. Mercer, "Challenges to the development of the next generation of self-reporting cardiovascular implantable medical devices," *IEEE Rev. Biomed. Eng.*, vol. 15, pp. 260–272, 2022.

[21] K. Nguyen, J. W. Y. Lim, K. P. Lee, T. Lin, J. Tian, T. T. T. Do, M. C. H. Chua, and B. P. Nguyen, "Heart disease classification using novel heterogeneous ensemble," in *Proc. IEEE EMBS Int. Conf. Biomed. Health Informat. (BHI)*, Jul. 2021, pp. 1–4.

[22] C. B. C. Latha and S. C. Jeeva, "Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques," *Informat. Med. Unlocked*, vol. 16, Jan. 2019, Art. no. 100203.

[23] A. P. Pawlovsky, "An ensemble based on distances for a kNN method for heart disease diagnosis," in *Proc. Int. Conf. Electron., Inf., Commun. (ICEIC)*, 2018, pp. 1–4.

[24] R. El Bialy, M. A. Salama, and O. Karam, "An ensemble model for heart disease data sets: A generalized model," in *Proc. 10th Int. Conf. Informat. Syst.*, May 2016, pp. 191–196.

[25] R. Atallah and A. Al-Mousa, "Heart disease detection using machine learning majority voting ensemble method," in *Proc. 2nd Int. Conf. New Trends Comput. Sci. (ICTCS)*, Oct. 2019, pp. 1–6.

[26] K. H. Miao, J. H. Miao, and G. J. Miao, "Diagnosing coronary heart disease using ensemble machine learning," *Int. J. Adv. Comput. Sci. Appl.*, vol. 7, no. 10, pp. 1–10, 2016.

[27] S. Bashir, U. Qamar, F. H. Khan, and M. Y. Javed, "MV5: A clinical decision support framework for heart disease prediction using majority vote based classifier ensemble," *Arabian J. Sci. Eng.*, vol. 39, no. 11, pp. 7771–7783, Nov. 2014.

[28] S. Sarah, M. K. Gourisaria, S. Khare, and H. Das, "Heart disease prediction using core machine learning techniques—A comparative study," in *Advances in Data and Information Sciences*. Singapore: Springer, 2022, pp. 247–260.

[29] W.-H. Ho, C.-T. Liao, Y. J. Chen, K.-S. Hwang, and Y. Tao, "Quickly convert photoplethysmography to electrocardiogram signals by a banded kernel ensemble learning method for heart diseases detection," *IEEE Access*, vol. 10, pp. 51079–51092, 2022.

[30] S. Subramani, N. Varshney, M. V. Anand, M. E. M. Soudagar, L. A. Al-Keridis, T. K. Upadhyay, N. Alshammari, M. Saeed, K. Subramanian, K. Anbarasu, and K. Rohini, "Cardiovascular diseases prediction by machine learning incorporation with deep learning," *Frontiers Med.*, vol. 10, Apr. 2023, Art. no. 1150933.

[31] A. Alqahtani, S. Alsubai, M. Sha, L. Vilcekova, and T. Javed, "Cardiovascular disease detection using ensemble learning," *Comput. Intell. Neurosci.*, vol. 2022, pp. 1–9, Aug. 2022.

[32] A. Tiwari, A. Chugh, and A. Sharma, "Ensemble framework for cardiovascular disease prediction," *Comput. Biol. Med.*, vol. 146, Jul. 2022, Art. no. 105624.

[33] L. Reddy, V. Talwar, S. Alle, R. S. Bapi, and U. D. Priyakumar, "IMLE-Net: An interpretable multi-level multi-channel model for ECG classification," in *Proc. IEEE Int. Conf. Syst., Man, Cybern. (SMC)*, Oct. 2021, pp. 1068–1074.

[34] I. Patel, A. Sandhya, V. S. Raja, and S. Saravanan, "Extraction of features from ECG signal," *Int. J. Current Res. Rev.*, vol. 13, no. 8, pp. 103–109, 2021.

[35] X. Chen, W. Guo, L. Zhao, W. Huang, L. Wang, A. Sun, L. Li, and F. Mo, "Corrigendum: Acute myocardial infarction detection using deep learning-enabled electrocardiograms," *Frontiers Cardiovascular Med.*, vol. 8, Aug. 2021, Art. no. 735864.

[36] R. Smisek, A. Nemcova, L. Marsanova, L. Smital, M. Vitek, and J. Kozumplik, "Cardiac pathologies detection and classification in 12-lead ECG," in *Proc. Comput. Cardiol.*, Sep. 2020, pp. 1–4.

[37] Y. Hu, Y. Zhao, J. Liu, J. Pang, C. Zhang, and P. Li, "An effective frequency-domain feature of atrial fibrillation based on time–frequency analysis," *BMC Med. Informat. Decis. Making*, vol. 20, no. 1, pp. 1–11, Dec. 2020.

[38] M. Alfaras, M. C. Soriano, and S. Ortín, "A fast machine learning model for ECG-based heartbeat classification and arrhythmia detection," *Frontiers Phys.*, vol. 7, p. 103, Jul. 2019.

[39] C. K. Roopa and B. S. Harish, "A survey on various machine learning approaches for ECG analysis," *Int. J. Comput. Appl.*, vol. 163, no. 9, pp. 25–33, Apr. 2017.

[40] A. T. Sadiq and N. H. Shukr, "Classification of cardiac arrhythmia using ID3 classifier based on wavelet transform," *Iraqi J. Sci.*, vol. 54, no. 4, pp. 1167–1175, 2013.

[41] S. Sadasivuni, V. Damodaran, I. Banerjee, and A. Sanyal, "Real-time prediction of cardiovascular diseases using reservoir-computing and fusion with electronic medical record," in *Proc. IEEE 4th Int. Conf. Artif. Intell. Circuits Syst. (AICAS)*, Jun. 2022, pp. 58–61.

[42] M. Obayya, J. M. Alsamri, M. A. Al-Hagery, A. Mohammed, and M. A. Hamza, "Automated cardiovascular disease diagnosis using honey badger optimization with modified deep learning model," *IEEE Access*, vol. 11, pp. 64272–64281, 2023.

[43] C. Guo, J. Zhang, Y. Liu, Y. Xie, Z. Han, and J. Yu, "Recursion enhanced random forest with an improved linear model (RERF-ILM) for heart disease detection on the Internet of Medical Things platform," *IEEE Access*, vol. 8, pp. 59247–59256, 2020.

[44] S. Mohan, C. Thirumalai, and G. Srivastava, "Effective heart disease prediction using hybrid machine learning techniques," *IEEE Access*, vol. 7, pp. 81542–81554, 2019.

[45] P. Ghosh, S. Azam, M. Jonkman, A. Karim, F. M. J. M. Shamrat, E. Ignatious, S. Shultana, A. R. Beeravolu, and F. De Boer, "Efficient prediction of cardiovascular disease using machine learning algorithms with relief and LASSO feature selection techniques," *IEEE Access*, vol. 9, pp. 19304–19326, 2021.

[46] B. Akkaya, E. Sener, and C. Gursu, "A comparative study of heart disease prediction using machine learning techniques," in *Proc. Int. Congr. Hum.-Comput. Interact., Optim. Robot. Appl. (HORA)*, Jun. 2022, pp. 1–8.

[47] A. J. Rogers, J. M. Miller, R. Kannappan, and P. Sethu, "Cardiac tissue chips (CTCs) for modeling cardiovascular disease," *IEEE Trans. Biomed. Eng.*, vol. 66, no. 12, pp. 3436–3443, Dec. 2019.

[48] H. D. Park, Y. Han, and J. H. Choi, "Frequency-aware attention based LSTM networks for cardiovascular disease," in *Proc. Int. Conf. Inf. Commun. Technol. Converg. (ICTC)*, Oct. 2018, pp. 1503–1505.

[49] A. Nikam, S. Bhandari, A. Mhaske, and S. Mantri, "Cardiovascular disease prediction using machine learning models," in *Proc. IEEE Pune Sect. Int. Conf. (PuneCon)*, Dec. 2020, pp. 22–27.

[50] M. A. Simonyan, Y. M. Ishbulatov, O. M. Posnenkova, V. A. Shvartz, A. S. Karavaev, V. V. Skazkina, V. I. Gridnev, R. V. Ukolov, and A. R. Kiselev, "Spectral analysis of photoplethysmography signal in patients with cardiovascular diseases and healthy subjects," in *Proc. 5th Sci. School Dyn. Complex Netw. Appl. (DCNA)*, Sep. 2021, pp. 180–182.

[51] X. Qian, Y. Li, X. Zhang, H. Guo, J. He, X. Wang, Y. Yan, J. Ma, R. Ma, and S. Guo, "A cardiovascular disease prediction model based on routine physical examination indicators using machine learning methods: A cohort study," *Frontiers Cardiovascular Med.*, vol. 9, Jun. 2022, Art. no. 854287.

[52] A. Rahim, Y. Rasheed, F. Azam, M. W. Anwar, M. A. Rahim, and A. W. Muzaffar, "An integrated machine learning framework for effective prediction of cardiovascular diseases," *IEEE Access*, vol. 9, pp. 106575–106588, 2021.

[53] H. Yang, Z. Chen, H. Yang, and M. Tian, "Predicting coronary heart disease using an improved LightGBM model: Performance analysis and comparison," *IEEE Access*, vol. 11, pp. 23366–23380, 2023.

[54] S. Ghorashi, K. Rehman, A. Riaz, H. K. Alkahtani, A. H. Samak, I. Cherrez-Ojeda, and A. Parveen, "Leveraging regression analysis to predict overlapping symptoms of cardiovascular diseases," *IEEE Access*, vol. 11, pp. 60254–60266, 2023.

[55] D. Chicco, C. A. Lovejoy, and L. Oneto, "A machine learning analysis of health records of patients with chronic kidney disease at risk of cardiovascular disease," *IEEE Access*, vol. 9, pp. 165132–165144, 2021.

[56] G. Joo, Y. Song, H. Im, and J. Park, "Clinical implication of machine learning in predicting the occurrence of cardiovascular disease using big data (nationwide cohort data in Korea)," *IEEE Access*, vol. 8, pp. 157643–157653, 2020.

[57] S. B. Shuvo, S. N. Ali, S. I. Swapnil, M. S. Al-Rakhami, and A. Gumaei, "CardioXNet: A novel lightweight deep learning framework for cardiovascular disease classification using heart sound recordings," *IEEE Access*, vol. 9, pp. 36955–36967, 2021.

[58] P. Wagner, N. Strodthoff, R.-D. Bousseljot, D. Kreiseler, F. I. Lunze, W. Samek, and T. Schaeffter, "PTB-XL, a large publicly available electrocardiography dataset," *Sci. Data*, vol. 7, no. 1, p. 154, May 2020, doi: 10.1038/s41597-020-0495-6.

[59] P. Wagner, N. Strodthoff, R. Bousseljot, W. Samek, and T. Schaeffter, "PTB-XL, a large publicly available electrocardiography dataset (version 1.0.3)," *Sci. Data*, 2022.

[60] A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley, "PhysioBank, PhysioToolkit, and PhysioNet," *Circulation*, vol. 101, no. 23, pp. e215–e220, Jun. 2000.

[61] C. Xie, L. McCullum, A. Johnson, T. Pollard, B. Gow, and B. Moody, "Waveform database software package (WFDB) for Python (version 4.1.0)," PhysioNet, India, Tech. Rep. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4627662/

[62] D. Makowski, T. Pham, Z. J. Lau, J. C. Brammer, F. Lespinasse, H. Pham, C. Schölzel, and S. A. Chen, "NeuroKit2: A Python toolbox for neurophysiological signal processing," *Behav. Res. Methods*, vol. 53, pp. 1689–1696, Feb. 2021.

[63] L. E. Peterson, "K-nearest neighbor," *Scholarpedia*, vol. 4, no. 2, p. 1883, 2009.

[64] L. Breiman, "Random forest," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[65] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 785–794.

[66] A. Natekin and A. Knoll, "Gradient boosting machines, a tutorial," *Frontiers Neurorobot.*, vol. 7, p. 21, Dec. 2013.

[67] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "CatBoost: Unbiased boosting with categorical features," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 1–11.
[68] Oracle. (2023). *Deploy a Machine Learning Model Close to the Network Edge*. Accessed: Sep. 21, 2023. [Online]. Available: https://docs.oracle.com/en/solutions/deploy-ml-at-edge/index.html#GUID-8EC86246-D724-4C16-8073-8CB5B2EA6719
[69] G. Sharma, A. M. Joshi, R. Gupta, and L. R. Cenkeramaddi, "DepCap: A smart healthcare framework for EEG based depression detection using time-frequency response and deep neural network," *IEEE Access*, vol. 11, pp. 52327–52338, 2023.
[70] R. Sharma, A. M. Joshi, C. Sahu, and S. J. Nanda, "Detection of false data injection in smart grid using PCA based unsupervised learning," *Electr. Eng.*, vol. 105, pp. 2383–2396, Apr. 2023.
[71] B. Murugesan, V. Ravichandran, K. Ram, S. P. Preejith, J. Joseph, S. M. Shankaranarayana, and M. Sivaprakasam, "ECGNet: Deep network for arrhythmia classification," in *Proc. IEEE Int. Symp. Med. Meas. Appl. (MeMeA)*, Jun. 2018, pp. 1–6.
[72] S. Mousavi and F. Afghah, "Inter- and intra-patient ECG heartbeat classification for arrhythmia detection: A sequence to sequence deep learning approach," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 1308–1312.
[73] A. Y. Hannun, P. Rajpurkar, M. Haghpanahi, G. H. Tison, C. Bourn, M. P. Turakhia, and A. Y. Ng, "Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network," *Nature Med.*, vol. 25, no. 1, pp. 65–69, Jan. 2019.
[74] S. Śmigiel, K. Pałczyński, and D. Ledziński, "ECG signal classification using deep learning techniques based on the PTB-XL dataset," *Entropy*, vol. 23, no. 9, p. 1121, Aug. 2021.
[75] M. P. Witvliet, E. P. Karregat, J. C. Himmelreich, J. S. de Jong, W. A. Lucassen, and R. E. Harskamp, "Usefulness, pitfalls and interpretation of handheld single-lead electrocardiograms," *J. Electrocardiol.*, vol. 66, pp. 33–37, May/Jun. 2021.
[76] K. Rajakariar, A. N. Koshy, J. K. Sajeev, S. Nair, L. Roberts, and A. W. Teh, "Accuracy of a smartwatch based single-lead electrocardiogram device in detection of atrial fibrillation," *Heart*, vol. 106, no. 9, pp. 665–670, May 2020.
[77] A. A. Hernandez, P. Bonizzi, R. Peeters, and J. Karel, "Continuous monitoring of acute myocardial infarction with a 3-lead ECG system," *Biomed. Signal Process. Control*, vol. 79, Jan. 2023, Art. no. 104041.

NIDHI SINHA (Member, IEEE) is a Research Scholar with the Electronics and Communication Department, MNIT Jaipur, Rajasthan. She is also a data scientist. She has published research articles in international journals and conferences. She is doing her research in the field of early detection of cardiovascular failure using machine learning and AI. She has experience in machine learning and artificial intelligence project deployment and in the field of solar cell design.

M. A. GANESH KUMAR with Electronics and Communication Engineering Department, MNIT, Jaipur, Rajasthan. He did his research in the field of cardiovascular disease detection using machine learning.

AMIT M. JOSHI (Senior Member, IEEE) received the M.Tech. and Ph.D. degrees from NIT, Surat, in 2009 and 2015, respectively. He has been an Assistant Professor with the Malaviya National Institute of Technology, Jaipur (MNIT Jaipur), since July 2013. He has published more than 100 research articles in excellent peer-reviewed international journals/conferences and also has published six book chapters. He has a total of 1424 Google Scholar citations, an i10 index is 41, and an H-index is 19. His research interests include biomedical signal processing, smart healthcare, VLSI DSP systems, and embedded system design. He is a member of IETE. He served as a Technical Program Committee Member for IEEE Conferences, such as iSES, ICCE, ISVLSI, and VDAT. He received the Honor of UGC Travel Fellowship, the Award of SERB DST Travel Grant, and CSIR Fellowship. He has attended well-known IEEE Conferences, such as TENCON-16, TENCON-17, ISCAS-18, and MENACOMM-19 across the world. He served as a Reviewer for technical journals, such as IEEE TRANSACTIONS, IEEE ACCESS, Springer, and Elsevier.

LINGA REDDY CENKERAMADDI (Senior Member, IEEE) received the master's degree in electrical engineering from the Indian Institute of Technology Delhi (IIT Delhi), New Delhi, India, in 2004, and the Ph.D. degree in electrical engineering from the Norwegian University of Science and Technology (NTNU), Trondheim, Norway, in 2011. He was with Texas Instruments, worked on mixed-signal circuit design, before joining the Ph.D. Program with NTNU. After finishing the Ph.D. degree, he worked on radiation imaging for an atmosphere space interaction monitor (ASIM mission to the International Space Station) with the University of Bergen, Bergen, Norway, from 2010 to 2012. He is currently a Leader of the Autonomous and Cyber-Physical Systems (ACPS) Research Group and a Professor with the University of Agder, Grimstad, Norway. He has coauthored over 160 research publications that have published in prestigious international journals and standard conferences in the research areas of the Internet of Things (IoT), cyber-physical systems, autonomous systems, robotics and automation involving advanced sensor systems, computer vision, thermal imaging, LiDAR imaging, radar imaging, wireless sensor networks, smart electronic systems, advanced machine learning techniques, connected autonomous systems, including drones/unmanned aerial vehicles (UAVs), unmanned ground vehicles (UGVs), unmanned underwater systems (UUSs), 5G- (and beyond) enabled autonomous vehicles, and socio-technical systems, such as urban transportation systems, smart agriculture, and smart cities. He is also quite active in medical imaging.

He is a member of ACM and a member of the editorial boards of various international journals and the technical program committees of several IEEE conferences. Several of his master's students won the best master's thesis award in information and communication technology (ICT). He serves as a reviewer for several reputed international conferences and IEEE journals. He is the Principal Investigator and a Co-Principal Investigator of many research grants from the Norwegian Research Council.