Letters in High Energy Physics
ISSN: 2632-2714
Volume 2024

Intrinsically Modified Physio-Biological Features Driven Heterogenous Ensemble Learning Model for Cardio-Vascular Disease Prediction

L. Hamsaveni[1], Rajesh B[3], M Vinayaka Murthy[2], Muralidhara B L[4]
1Associate Professor, DoS in Computer Science, University of Mysore, Mysuru, Karnataka, India
hamsa1367@gmail.com
2Senior Lecturer, School of Management, Mahindra University Hyderbad, Telangana, India
Rajesh.balarama@mahindrauniversity.edu.in
3Professor, School of Computer Science, REVA University, Bengaluru, Karnataka, India.
dr.m.vinayakamurthy@gmail.com
4Senior Professor,Department of Computer Science, Bangalore University, Bangalore , Karnataka, India
murali@bub.ernet.in

Abstract- Cardiovascular disease and other non-communicable illnesses have been on the rise in recent years. Despite innovations in computer-aided diagnosis (CAD) and clinical decision systems, unlike vision-based e-healthcare practices, heart-disease prediction requires learning over the different bio-physiological parameters related to the heart's health. The limitations of the datasets including class-imbalance, redundant computation and the threat of local minima and convergence, and resulting low-accuracy confine real-time significance of the at hand cardiovascular disease prediction (CDP) systems. In this paper a robust intrinsically modified bio-physiological parameters driven heterogenous ensemble learning based CVD prediction model is proposed. We focused on both feature optimization as well as computational efficacy to achieve a robust CAD solution towards CVD diagnosis. Our proposed method applies age, gender, cholesterol, protein profiles, body mass index information, stoke profile or history, electro-cardiogram information etc. from the benchmark dataset to enable a scalable CVD prediction model. To ensure semantic feature driven learning, the aforesaid features were processed for Word2Vec embedding, which was followed by resampling by using synthetic minority over-sampling technique (SMOTE) and its variants, SMOTE-Boundary Line and SMOTE-ENN which helped to alleviate any probability of class-imbalance. Subsequently, Principal Component Analysis (PCA), Cross-Correlation Analysis (CCRA) and Significant Predictor Test (SPT) methods were applied distinctly to retain the optimal feature sets. The selected feature instances were normalized by applying Min-max Scalar Normalization method. The normalized features were taught using a mixed-method ensemble learning strategy that comprised Base Classifier (RF), Decision Tree (DT), Support Vector Machine (SVM) variations, Naïve Bayes (NB), Logistic Regression (LOGR), Linear Regression (LR), Random Forest (RF), and Extra Tree Classifier (ETC) as foundational classifiers. It used the maximum voting ensemble (MVE) method to determine if each individual was CDV-Positive or CVD-Negative. The results show that the proposed method is resilient for application in real-world CDS scenarios, as it surpasses all prior state-of-the-art approaches in terms of CVD prediction accuracy (99.93%), precision (99.69%), recall (99.53%), and F-Measure (99.60%).

Keywords— *Heart Disease Prediction, Data Mining, Machine Learning, SMOTE-ENN, Significant Predictor Test, Heterogenous Ensemble Learning, Computer Aided Diagnosis.*

I. INTRODUCTION

Software innovations, decentralised computing, and affordable hardware have all seen meteoric rises in the past few years, opening up a world of possibilities for new applications that might help enterprises make better, more timely decisions. Amongst the major demands, healthcare sector has always remained the dominant due to high-pace rising global population and allied stress on at hand manual clinical decisions. The mounting stress on human resource-based clinical decisions has triggered academia-industries to achieve more effective and scalable computer aided diagnosis (CAD) and clinical decision systems (CDS) so as to cope up global demands [1]. Despite aforesaid motivations, guaranteeing optimality of an e-healthcare tool remains a challenge, especially due to symptomatic diversity, complex symptoms and limited annotated dataset [2]. In fact, medical diagnosis turns out to be decisive yet challenging task due to aforesaid challenge that becomes even more complex over data diversity and hence serving automated diagnosis becomes more trivial [1]. On the contrary, the low availability of physicians and inability to assess electronic details make clinical decisions difficult. This as a result requires automated CAD solution is inevitable [1]. Undeniably, the last few years have witnessed computer-driven information-based CDS towards cost-effective and timely diagnosis decisions and allied medical care. In conjunction with these motivations, the majority of hospitals these days use CDSs to manage patient, diagnosis details and allied information. Unfortunately, aforesaid techniques severely form humongous volume of data, which are rarely employed to inform CDS purposes [3]. This

information facilitates a large volume of esoteric information which have not been exploited significantly and has been mostly disregarded [4]. On the other hand, the majority of the at hand CDS systems or CAD systems have been applied to exploit visual details (say, vision computing) to perform diagnosis details for instance, brain tumor detection, cancer detection, diabetic retinopathy detection, etc. However, there exists numerous other healthcare challenges which demands multivariate feature's analysis to perform healthcare diagnosis such as cardiovascular disease (CVD) detection [5][6].

The matter of fact is that the noncommunicable diseases (NCDs) and resulting mortality rate is rising with an alarming pace. According to a recent study, noncommunicable diseases (NCDs) account for about 71% of all fatalities worldwide, with a shockingly high percentage (over 80%) in poor and middle income nations [7].

Contemporarily, cardiovascular disorders (CVDs) are amongst the dominant illnesses in the world [8]. A recent study by World Health Organization (WHO, 2019) revealed that heart disease has taken more than 17.9 million people, causing almost 32% of the global death [9]. A number of organizations functional in medical domain have applied data mining and pattern analysis models extensively to perform CVD prediction. Yet, ensuring optimal set of physio-biological patterns and symptoms remains challenge for accurate clinical decisions. Nearly 45 percent of all fatalities occur from cardiovascular diseases (CVDs), which include hypertension, heart disease, and stroke, according to a World Health Organization research.

Conversely, by 2030, low and middle-income nations are projected to have a prevalence of NCDs of about 50% [8][10].

Literatures also indicate that the annual mortality rate due to the CVDs can reach up to from 17.5 million in 2012 to 22.2 million in 2030 [9]. It alarms industry to design robust and accurate CAD solution for scalable heart disease detection.

Noticeably, heart disease represents an extensively used term signifying the varied conditions impacting arteries, blood vessels and other organs, resulting malfunction. Human respiratory systems around the globe have been impacted by the SARS-CoV-2 virus, according to recent studies.

As a result, people's lungs release insufficient oxygen, which can negatively affect heart health and potentially lead to heart failure [11][12].

However, heart disease is typically the outcome of atheromatous plaques, abnormal lipid metabolism, and the buildup of lipids and other liquids within the coronary arteries. This can lead to a constriction of blood vessels, which in turn can cause myocardial ischemia, oxygen shortage, or tissue death. Chest pain, chest tightness, myocardial infarction, and other symptoms are common outcomes of these occurrences [13].

The aforesaid patterns as cumulative phenomenon has been causing 12 million deaths globally [14]. The complexity involved in heart disease diagnosis and remedial have been resulting severe death and hence high mortality rate [15]. On the other hand, considering a smaller fraction of human ecosystem the medical (diagnosis and remedial) expenses involved are expected to rise 41% in the US, mounting almost $177.5 billion by 2040 [5]. Unfortunately, affording such huge cost can't be easier for the low-income countries [15] and allied households and therefore there is a need to design more efficient and robust CAD solution for heart disease detection and diagnosis [16][17].

Unlike vision-based CAD solutions, heart disease detection and prediction model require learning a large number of bio-physiological patterns pertaining to the functional aspects of the heart mechanism [18]. In this reference, numerous efforts have been made by deploying machine learning methods over the aforesaid bio-physiological parameters to perform heart disease prediction or cardiac disorder analysis [18]. However, merely applying over redundant data can't guarantee reliability of the solution [16][18]. Despite the fact that clinically assessed and specialized bio-physiological parameter's analysis can enable data mining-based CVD prediction; yet, monitoring the most recent patterns and its relevance towards human heart functionality is decisive. In addition, learning a machine learning model over the suitable feature set is equally important. It infers that a machine learning-based model can be effective only with the optimal set of data, intrinsically optimal features and improved learning environment. The depth assessment has revealed that there exists certain set of bio-physiological parameters including gender, insulin, cholesterol, lips profiles, body mass index (BMI), stroke details, fasting blood sugar, electro-cardiogram patterns (ECG) etc. which can be used as features for multivariate learning. It can achieve heart disease of CVD disease prediction; yet, as stated earlier it requires data optimality and computational efficacy. Despite several previous attempts, most of the state-of-the-art methods (see Section II) are either too inaccurate or too unreliable to reliably forecast the occurrence of heart disease. Machine learning models have been fed sparsely characterized input data in the majority of previously published methods.

Interestingly, in almost all datasets available and used the number of data-elements (say, instances)

pertaining to the normal heart functions are more in comparison to the heart malfunction. It signifies the presence of class-imbalance and hence the likelihood of skewed learning can't be ignored. Machine learning models trained on biased data are more likely to produce inaccurate predictions due to false positives and negatives. Furthermore, there are some literatures suggesting that training a machine learning model over selected high-significant features can improve accuracy [10]. However, there has been little effort to evaluate the effectiveness of various feature selection models in predicting heart disease.

It broadens the horizon for researchers to design a robust feature model which could address both class-imbalance as well as feature optimality to tune learning models for accurate heart disease prediction. The depth assessment of literatures indicate that the major state-of-arts have applied machine learning algorithm as standalone classifier, where many machine learning models have demonstrated diverse levels of performance on the same dataset. It could be challenging to generalize a solution in this situation. An ensemble learning model could be a game-changer in resolving this issue [10].

Though, in the past a few researches have applied ensemble learning methods like RF, AdaBoost, XGBoost methods which are homogenous in nature. For instance, RF ensemble applies bootstrapped Decision Tree (DT) algorithm to constitute voting-based ensemble classification. Heterogeneous ensemble approaches, which include basic classifiers from several machine learning algorithms (e.g., regression, neuro-computing, pattern learning, decision tree, etc.), can outperform homogenous ensemble methods in terms of accuracy.

Considering it as motivation, in this paper the focus is made on improving both feature as well as computational aspects. In other words, the proposed model intends to exploit improved features from the input benchmark bio-physiological inputs, while it intends to use heterogenous ensemble model to achieve improved and reliable heart attack prediction solution.

In light of the aforementioned knowledge gaps and related areas of study, this work proposes a new model for heart disease prediction that makes use of heterogeneous ensemble learning and incorporates intrinsically changed physiological features. The many benchmark datasets that include bio-physiological inputs relevant to heart disease were especially taken into account in this study.

Unlike traditional approaches where the authors have directly passed inputs to the machine learning classifier(s), to exploit latent or semantically enriched features we at first transformed input datasets into the semantic embedded feature vector by applying Word2Vec method. Here, the key motive behind the use of Word2Vec embedding method was to improve intrinsic features which could enhance overall learning efficacy. So, to prevent class imbalance, the suggested model uses SMOTE, SMOTE-BL, and SMOTE-ENN, which provide an ideal distribution of samples without resorting to hotspot generation.

The resampled data is then processed for feature selection by applying three different feature selection methods, including PCA, CCRA and Mann Whitney SPT methods. The primary goal in this case was to find the optimal combination of features and methodologies for accurate prediction of heart disease using the aforementioned repeated resampling and feature selection techniques. After we picked the perfect set of features, we mapped each data instance in the range of 0 to 1 using the Min-Max scaler normalisation method, which prevented over-fitting and convergence.

In conclusion, the suggested heterogeneous ensemble learning model was trained using the normalised data. Support vector machine (SVM) variations, decision tree (DT), Naïve Bayes (NB), Logistic Regression (LOGR), Linear Regression (LR), Random Forest (RF), Artificial Neural Network Levenberg Marquardt (ANN-LM), and Extra Tree Classifier (ETC) were all part of this model's foundation classifiers. In order to arrive at a final prediction about cardiovascular illness, the proposed model utilized the aforementioned machine learning (base) classifiers to carry out maximum voting ensemble (MVE).

In this case, generalisable performance is achieved through the usage of MVE ensemble, which guarantees higher reliability compared to the standalone classifier(s). The suggested model performs better than existing state-of-the-art models in predicting cardiac illness, with scores of 99.93% for accuracy, 99.69% for precision, 99.53% for recall, and 99.60% for F-Measure, proving its suitability for real-world CAD applications.

What follows is a breakdown of the remaining sections of this manuscript. In Section II, we cover the relevant literature; in Section III, we formulate the problems. The study questions are presented in Section IV, followed by the suggested methodology and its execution. Sections V and VI present the findings and conclusions from the simulations, correspondingly. At the end of the manuscript, you will find the references that were used.

II. Related Work

Heart disease (CVD) prediction model proposed by Parija et al. [19] is based on machine learning. In a similar vein, Shadman et al. [10] used a variety of ML models for predicting cardiac problems,

including as ANNs, Simple Logistics (SL), RF, SVM, and NB. The results showed that a heart disease prediction accuracy of 97.53% was achieved by the SVM classifier using 10-fold cross-validation.

Noticeably, to perform aforesaid heart disease prediction, the authors designed sensors for collecting the parameters like blood-pressure, temperature, humidity, and heartbeat. The dataset from the repository of University College London (UCI) was used by Durairaj et al. [20] to predict cardiac illness using artificial neural networks (ANNs) based on multilayer perceptrons (MLPs). With an accuracy of 96.30 percent, their model was the most accurate. To forecast the occurrence of coronary heart disease using UCL clinical datasets, Hisham et al. [21] implemented a number of machine learning models, including as LR, SVM, K-Nearest Neighbor (KNN), and MLP ANN. In order to enhance the accuracy of heart disease prediction, the authors used a pre-processing method based on K-Means clustering, a Genetic Algorithm (GA), and recursive feature selection. They also employed the Synthetic Minority Oversampling Technique (SMOTE) algorithm.

Yet, the highest accuracy obtained was 86.6% [21]. Nancy et al. [22] on the other hand found that ANOVA-based feature selection could achieve higher accuracy with RF classifier for CVD prediction. In order to forecast the occurrence of heart disease using the UCI heart disease dataset, the authors [23] employed various deep learning and machine learning algorithms.

To enhance prediction accuracy, they applied Isolation Forest algorithm to drop insignificant features. The depth assessment revealed the heart disease prediction accuracy of 94.2% by using deep learning method. In [24], RF classifier was applied to perform CVD prediction where it exhibited prediction accuracy of 98%. Applying the RF classifier to the Kaggle heart disease dataset, the authors [25] achieved a maximum accuracy of 86.9% in detecting CVD in a patient. Using RF ensemble learning, Yongyong et al. [26] were able to predict CVDs. The risk of cardiovascular disease (CVD) was estimated in this study using the following variables: age, BMI, TG, and DBP.

The authors [27] designed an intelligent heart disease prediction system (IHDPS) by applying machine learning-based models like NB, ANN and DT algorithms. The simulation revealed that NB algorithm exhibited accuracy of 86.12%, while ANN and DT exhibited accuracy of 85.68% and 80.40%, respectively. In [28] applied k-NN, NB, DT algorithms where it exhibited heart disease prediction accuracy of 45.67%, 52.33% and 50.00%, correspondingly. In their work on cardiac disease prediction, the authors [29] utilized J48 DT in conjunction with bagging algorithms. In their study, the authors found that using a Gain ratio decision method in conjunction with voting discretization improved the accuracy, sensitivity, and overall performance by 72.01% to 77.90% and 78.90% to 84.10% by J48 DT and bagging, respectively. The several supervised machine learning techniques used to forecast the occurrence of cardiovascular disease are detailed in [30], among them are NB, k-NN, and DT. The usage of an ANN algorithm was shown to obtain a prediction accuracy of 100% in a simulation using 10-fold cross validation. However, a CVD prediction accuracy of 99.20% was achieved by combining DT and GA feature selection. One study used artificial neural networks (ANN) to forecast the occurrence of cardiovascular disease based on thirteen bio-physiological variables, such as gender, blood pressure, cholesterol, obesity, and smoking habits [31]. Similarly, a combined method using ANN and GA for CVD classification was suggested in [32], and it achieved an impressive 89% accuracy. Likewise, in [33] DT, Classification and Regression Tree (CART) and Iterative Dichotomized 3 (ID3) were utilized. The best CVD prediction accuracy of 83.40% was achieved using the CART approach using 10-fold cross-validation. The combined Information Gain and Adaptive Neuro-Fuzzy Inference System (ANFIS) models demonstrated a 98.24% accuracy rate in the prediction of CVD in reference [34]. In a similar vein, with a total of eleven attributes including gender, age, dummy values, heart rate, chest pain, cholesterol, blood sugar, blood pressure, cardiogram, alcohol consumption, and smoking behavior as input features, the authors of [35] utilized REPTREE, NB, Bayes Net, J48, and CART for the purpose of cardiovascular disease prediction. I found it interesting that it showed the CVD prediction accuracy of J48 (99%), REPTREE (99.07%), CART (99.07%), Bayes Net (98.15%), and NB (97.22%).

Similarly, in [36] 13 features where NB exhibited CVD prediction accuracy of 85.03%, while DT exhibited an accuracy of 84.01%. in [37], the authors applied k-NN, DT, Sequential Minimal Optimization (SMO), J48 and NB classifier that in conjunction with 10-folds cross-validation exhibited CVD prediction accuracy of 82.77%, 82.77%, 83.732% and 81.81%, respectively. The machine learning algorithms including ANN, DT and NB were applied in [38] where the last (i.e., NB) exhibited the highest CVD prediction accuracy of 82.91%. In the same manner, the authors [39] applied machine learning algorithms including C5.0 DT, k-NN, SVM and ANN, where DT exhibited the prediction accuracy of 93.2%, while ANN, SVM and k-NN exhibited CVD prediction accuracy of 80.20%, 86.05% and 88.37%, correspondingly. To forecast CVD, the authors of [40] developed a hybrid approach that uses C4.5, MLP, MLR, and FURIA, an algorithm for unpredictable rule induction. When predicting CVD, the scientists used

K-Mean clustering techniques in conjunction with particle swarm optimization (PSO) and correlation-based feature subset (CFS) (feature selection).

Training their hybrid model over a total of 26 features, MLR was found exhibiting the highest CVD prediction accuracy of 88.40%. In [41], NB, DT, k-NN, Memorial network, and ID3 algorithm were applied to perform CVD prediction. Regardless of the computational drain on input features such as age, gender, cholesterol, and blood pressure, the best prediction accuracy was 80.60%.

In [42], multiple kernel learning (MKL) with ANFIS to perform CVD prediction. They could achieve the highest CVD prediction specificity and sensitivity of 98% and 99%, correspondingly on KEGG metabolic reaction network dataset. In [43] MLP-NN with backpropagation (BP) algorithm was applied to perform heart disease prediction. Optimisation of GA using support vector machines was suggested in [44]. The authors discovered that the extracted feature produced an accuracy of 88.34% and SVM an accuracy of 83.70%.

In [45] DT and ANN models were applied, where 10-fold validation with pruned data exhibited the CVD prediction accuracy of 78.14%, while each standalone method exhibited the accuracy of 77.40% and 76.66% by using ANN and C4.5, correspondingly. In [46], the authors applied ejection fraction and serum creatinine as the vital features to perform CVD prediction. In [47], GA and PSO were applied altogether to perform feature selection that in conjunction with RF performed CVD prediction. To execute CVD prediction, the writers [48] utilized a variety of machine learning techniques, including as k-NN, AdaBoost (AB), DT, and RF.

The simulation results confirmed that the applied machine learning models exhibited CVD prediction accuracy of k-NN, AB, DT, and RF algorithms which achieved an accuracy of 100%, 100%, 96.10% and 99.03%, correspondingly. With Cleveland dataset, k-NN and RF exhibited accuracy of 97.83% and 93.437%, correspondingly.

In their study, Narain et al. [49] used quantum ANN in conjunction with the Framingham risk score (FRS) to predict CVD. The results showed an impressive accuracy rate of 98.57%. The Cleveland heart disease dataset, which contains 17 features, was utilized by Shah et al. [50] for the purpose of CVD prediction. The authors used k-NN, NB, DT, and RF as s-learner algorithms; k-NN achieved a prediction accuracy of 90.8%.

Drod et al. [51] performed the significant risk variable selection to improve CVD prediction. More specifically, they applied metabolic-related fatty liver disease (MAFLD) systems with the blood biochemical analysis and subclinical atherosclerosis assessment to perform CVD prediction. Technically, they applied LR classifier, univariate feature ranking, with PCA to perform feature selection followed by classification. The authors applied hypercholesterolemia, plaque scores, and duration of diabetes as the parameters to perform CVD prediction, where the highest accuracy was obtained as 85.11%. Alotalibi [52] applied the Cleveland Clinic Foundation dataset to perform machine learning-based CVD prediction. More specially, the authors applied DT, LR, RF, NB, and SVM, where the 10-fold cross-validation resulted the highest of 93.19% CVD prediction accuracy by using DT, which was followed by SVM (92.30%). Hasan and Bao [53] focused on identifying the optimal feature selection method towards CVD prediction. Along with RF, SVM, k-NN, NB, and XGBoost, it used a Boolean process-based common "True" condition to apply various feature selection methods, such as filtering, wrapping, and embedding. Among the three methods tested here, XGBoost classifier with wrapper-based feature selection had the best CVD prediction accuracy (73.74%), while SVC came in second with 73.18% and ANN third with 73.20%.The Hybrid Random Forest with Linear Model (HRFLM) was developed by Senthilkumar et al. [54] to ensure accurate CVD prediction using thirteen distinct features.

The highest accuracy obtained was 88.7%. Ramalingam et al. [55] designed an alternating DT model with PCA, where the later enabled suitable feature selection. Despite SVM used with Ant Colony Optimization (ACO) feature selection, the allied complexity can't be ruled out. Rajdhan et al. [56] stated that the RF algorithm can yield accuracy of 90.16 % for CVD prediction over the UCI Cleveland heart disease dataset. On the contrary, LR, NB, and DT exhibited accuracy of 85.25%, 85.25%, and 81.97%, correspondingly. Khourdifi et al. [57] too applied RF, k-NN, and ANN to perform heart disease prediction. They inferred that the use of hybrid approach with PSO and ACO-based feature selection can achieve better prediction accuracy. Applying aforesaid feature selection methods, they achieved prediction accuracy of 99.65% (PSO), and 99.6% (ACO) by using RF algorithm. While Jagtap et al. [58] used SVM, LR, and NB algorithms, their scalability is limited by the highest stated accuracy of 64.4%. To predict CVD, Haq et al. [59] employed a combination of ANN, k-NN, SVM, LR, DT, RF, and NB, as well as Lasso feature selection. At its peak, SVM achieved 88% accuracy, LR 87%, and ANN 86%. For cardiovascular disease prediction, Jindal et al. [60] used k-NN, LR, and RF. They achieved an accuracy of 87.5% using a wide range of features, including age, cholesterol, fasting sugar, chest discomfort, sex, and blood pressure.

III. Problem Formulation

This is the matter of fact that the cases of cardiovascular diseases (CVD) are on rise globally where the alarming morality rate has been triggering both academia as well as industry to achieve more efficient diagnosis measure for earlier heart disease prediction solution. On the other hand, coping with the exponentially rising global population and resulting pressure on at hand healthcare infrastructures too have forced industries and government to achieve and apply varied computer-aided diagnosis solutions to make earlier diagnosis decisions and allied medication. Though, in the past varied vision-based computing models are developed towards e-healthcare purposes; however, heart disease prediction turns out to be more challenging and trivial due to the lack of direct symptoms. Moreover, the dependency or associations amongst the different indirect symptoms including biological parameters and the physiological patterns make heart disease prediction more challenging. Though, training over the different (aforesaid) bio-physiological parameters can lead a better and reliable heart disease prediction solution; yet, most of the state-of-arts data mining-based approaches are limited in certain terms such as low-accuracy, lack of the ability to address class-imbalance, vulnerability towards local minima and convergence etc. Consequently, it makes major at hand solutions confined towards real-world CVD prediction solution. To enable a robust CAD solution towards CVD prediction, guaranteeing both feature optimality as well as computational efficacy is inevitable. Realizing the fact that the majority of the at hand CVD prediction models (including both machine learning as well as deep networks) have applied local features from the input bio-physiological parameters to train a model for prediction. However, such methods often lack in the ability to address long-term dependency (say, training contextual details) over the consecutive bio-physiological patterns. Such inabilities can be better addressed by applying semantic features or the latent features obtained over the sequential bio-physiological test patterns. Considering it as motivation, in this work Word2Vec word-embedding method was applied over the different bio-physiological features encompassing age, gender, cholesterol, lipid profile, stroke history, ECG profile etc. This approach converts input sequential patterns into the equivalent embedded (numeric) matrix. It not only addresses the problem of long-term dependency but also achieves computational efficacy. In addition to the aforesaid issue, the matter that the number of instances pertaining to the normal person is relatively higher than the heart disease patients. It gives rise to the serious issue of class-imbalance and hence training a model over such skewed data can impact training efficiency and hence can show false positive or negative performance. To address this problem performing resampling can be of great significance. In this reference unlike under-sampling and over-sampling methods, which can give rise to the iterative hotspot issue, the improve methods like SMOTE or SMOTE-ENN can be the viable approach. Considering this fact, in this work SMOTE-ENN method (along with the other SMOTE variants) has been applied over the embedded matrix data. Here, the key motive was to assess and identify the optimally performing resampling method so as to perform scalable and reliable heart disease prediction. Indeed, resampling can enhance sample (distribution) optimality; but, it comes with a price: additional computation. To get around this, it's important to keep only the features that can provide better accuracy with less computational load. This work's motivation was to find the best set of features for further learning and classification using three feature selection methods: principal component analysis (PCA), clustering, and WRST significant predictor tests. The datasets that were resampled were SMOTE, SMOTE-BL, and SMOTE-ENN. To make heart disease prediction stand out, the various resampled datasets were subjected to the aforementioned feature selection approaches. Next, the features that were chosen were subjected to Min-Max normalization. This method assigned a value between 0 and 1 to each data instance, reducing the chances of over-fitting and convergence. First, the input features were processed for the aforementioned feature optimization measures. Then, they were passed on to a novel and robust heterogeneous ensemble learning classifier for two class classification. This approach differs from traditional machine learning-based CVD prediction models, which directly feed the input data to standalone machine learning classifiers. Notably, traditional approaches have used independent machine learning classifiers to learn and classify input data, which yields varying degrees of accuracy even when using the same dataset. The solution's generalizability is limited by this performance variation. The suggested model uses a HEL learning framework with SVM, DT, NB, LOGR, LR, ANN-LM, RF, and ETC as basic classifiers to conduct MVE ensemble-based prediction, thus resolving this issue. Therefore, the heart disease prediction (i.e., CVD-Positive or CVD-Negative) for the various characteristics, resampling strategies, and proposed compositions is accomplished with the suggested HEL-MVE ensemble learning framework. To guarantee better and more generalizable performance towards heart disease prediction, it is crucial to determine the best data-resampling strategy, feature selection approach, classification environment, and feature set. In order to evaluate the effectiveness of the model(s) that have been proposed, we use MATLAB to build them and then collect confusion metrics for accuracy, precision, recall, and F-measure. By comparing the suggested model to state-of-the-art methods for CVD prediction, as well as to other models, we are able to characterize its performance.

IV. Research Questions

This work establishes a set of research questions that, when answered, will lay the groundwork for a scalable and reliable model for predicting the occurrence of heart disease, in line with the overarching research objectives and methodological scopes. Here are the research questions:

RQ1*: Can the strategic amalgamation of bio-physiological parameters including genders, age, cholesterol, lipid profile, stroke history, ECG pattern, etc. enable machine learning methods performing scalable and reliable heart disease prediction (or CVD prediction) model?*

RQ2*: Can the use of use of improved SMOTE resampling (SMOTE-ENN), Wilcoxon Rank Sum Test (WRST) Significant Predictor Test, and Min-Max Normalization and proposed heterogenous ensemble learning (HEL) method be effective towards reliable CVD prediction model?*

RQ3*:* Is it possible that the HEL ensemble learning model outperforms and is more trustworthy than the conventional machine learning approaches that operate independently?

A solid, trustworthy, and extensible CVD prediction system can be built upon the results of these research questions, which can be proven through quantitative means.

V. System Model

The general approach and related sequential implementation are the main topics of this section. Here are the steps that make up the total method:

1. *Data Acquisition and Pre-processing*
2. *Semantic Feature Modelling*
3. *Feature Resampling*
4. *Feature Selection*
5. *Data Normalization*
6. *Heterogeneous Ensemble Learning based Maximum Voting Ensemble for CVD Prediction*

The detailed discussion of the sequential implementation is given as follows:

A. Dataset Acquisition and Pre-Processing

Realizing the clinical associations amongst the different bio-physiological parameters and heart disease probability, in this work we intended to exploit maximum possible features so as to enable better training and hence (CVD) prediction. In light of this, we used the Cleveland dataset, housed in the machine learning repository at the University of California, Irvine (UCI), which consisted of 303 examples across 13 distinct feature sets. Table I provides an excerpt of the data that was considered together with the linked feature significances.

Table I. Heart disease prediction dataset

SN.	Attribute Icon	Attribute name	Description
1.	Age	Age	Patient Age
2.	Sex	Gender	Males -1, Female -0.
3.	Chest pain type	Chest Pain type	Male -1, Female-0.
4.	Resting blood pressure	Rest state blood pressure (mm/Hg)	Resting blood pressure upon hospital admission, measured in mm/Hg
5.	Serum cholesterol	Serum cholesterol (fat)	Blood cholesterol level measured in mg/dL
6.	Fasting blood sugar	Fasting blood sugar (not eating)	If the blood sugar level is over 120 mg/dL, after a fast of not eating overnight, it is considered to be high (1-high). In case, it is below 120 mg/mL, it is stated to be normal (0-false).
7.	Resting ECG	Rest ECG test	An ECG test result can be categorised as follows: 0 for a normal result, 1 for the presence of ST-T wave abnormality, and 2 for left ventricular hypertrophy.
8.	Maximum heart rate	Max. heart rate achieved	Max. heart rate during exercise.
9.	Exercise angina	Exercise induced angina	Angina occurred by a workout: 0 for No; 1 for Yes.
10.	Old-peak	ST depression (ECG test)	ST depression due to exercise relative to relaxation will observe in the ECG test.

11.	ST slope	Slope (ST depression)	Maximum workout: 1-Upsloping; 2-Flat; 3-Down sloping.
12.	Ca	No. of vessels (0-3)	The number of major
13.	Thai	Thalassemia (Haemolytic disease)	Thalassemia is a blood disorder caused by abnormal haemoglobin production with a score of 3- Indicating normal production, 6-Permanent deficiency, 7- Signifying temporary impairment.
14.	BMI	Body Mass Index	It signifies body mass index presenting patient's specific body structure and mass value.
15.	Target	Heart failure class attribute	No heart disease-0, Heart disease-1.

Once obtaining aforesaid dataset, it was processed for pre-processing before executing the proposed predictive model. The proposed dataset was processed for extensive pre-processing and cleaning that makes computing easier and hence achieves reliable training and hence higher reliability. An numeric value indicating the presence of a patient's cardiac condition is signified by a target element in the aforementioned dataset. If there is no heart illness, the target score is 0, and if there is heart disease, the score is 1. We took into account the gender of the samples as a whole because, according to the research, men had a higher risk of cardiovascular disease than women.

In other words, the data element 'sex' comprised two classes: 1 and 0, signifying male and female, respectively. Chest pain (CP) is also an indicator of heart disease and failure. Considering this fact, we considered CP profile comprising four classes. The proposed model encompassed four classes of CP where two different classes represented fasting blood sugar ('fbs'). Additionally, it encompassed three different classes of resting ECG 'restecg' and two classes presenting exercise angina 'exang'. In addition, 'slope' also called ST slope comprises three classes. Additional characteristics are included, such as resting blood pressure ('trestbps'), cholesterol ('chol'), age, and oldpeak. The patients' body mass indexes were also considered. The dataset under consideration underwent processing to eliminate duplicate or missing values.

Noticeably, the missing element signifies an incomplete or repeated data-element. Such missing elements can impact overall learning efficiency and accuracy. To alleviate it missing elements were removed by performing outlier assessment. Considering limited data instances (say, sample size), to improve computational efficacy, the missing elements were substituted either by means of a user-defined constant or the average (dataset) value. Unlike traditional methods which remove aforesaid missing elements completely, our proposed model substitutes the missing elements with the average value of the dataset. Realizing the limited size of Cleveland dataset, we combined other datasets including Hungry, Switzerland and Kaggle datasets. Thus, a total of 1100 instances was prepared with 14 different features.

B. Semantic Feature Modeling

To ensure optimal feature learning while addressing long-distance dependency, the proposed model focused on exploiting semantic features. Unlike traditional deep learning methods or tokenization approaches which often exploit local features and fail in exploiting contextual features, the proposed model applies word-embedding method to perform feature modeling. To improve computational efficacy word-embedding based semantic feature modeling is performed that yields low-dimensional semantic features for further learning and classification. The dataset comprised the different bio-physiological features for the different patients representing both classes, heart disease patients and non-heart disease patient. Such feature diversity and corresponding embedding matrix output present both contextual as well as latent information to perform accurate and reliable CVD prediction. To facilitate additional learning and classification, we employed the Word2Vec word-embedding technique to produce a semantic embedded matrix in this study.

We used Gensim Word2Vec technique to convert input instances into equivalent embedding vector. We designed Word2Vec model with dual-layer neural network encompassing two hidden layers that generated semantic feature(s) with sparser feature outputs. In this approach, the input data (say, instances or tokens) was retrieved based on the window of the connecting context-window. Let, W_{i-1} , W_{i-2} , W_{i+1}, W_{i+2} be the context words retrieved from the data corpus, then the CBOW method predicts W_i which is highly related to the other data instance available within the dataset. The predicted embedding outputs were related to the target token value W_i. From a functional standpoint, the CBOW embedding

approach is comprised of two sets of word-embedding vectors, one for each data instance (here, w∈V being the feature instance) and one for each target-side, denoted as v_w,v_w^'∈R^d. We utilized embedding methods based on Gensim, wherein a data corpus input instance window represents the center token w_0 and generates appropriate context embedded vectors w_1,...,w_C. This is how the CBOW loss is calculated: (1).

$$v_c = \frac{1}{C}\sum_{j=1}^{C} v_{w_j} \tag{1}$$

$$\mathcal{L} = -\log \sigma\left(v'_{w0}\, T_{v_c}\right) - \sum_{i=1}^{k} \log \sigma\left(-v'_{n_i}\, T_{v_c}\right) \tag{2}$$

In (2) $n_1, \ldots, n_k \in V$ signifies the negative examples obtained from the noise distribution P_n over input vectors V. In (2), $\mathcal{L}$ gradient is obtained with respect to the target value v'_{w0}, negative target value v'_{n_i} and average context source (v_c)

$$\frac{\partial \mathcal{L}}{\partial v'_{w0}} = \left(\sigma\left(v'_{w0}\ T_{v_c}\right) - 1\right) v_c \tag{3}$$

$$\frac{\partial \mathcal{L}}{\partial v'_{n_i}} = \left(\sigma\left(v'_{n_i}\ T_{v_c}\right) - 1\right) v_c \tag{4}$$

$$\frac{\partial \mathcal{L}}{\partial v_c} = \left(\sigma\left(v'_{w0}\ T_{v_c}\right) - 1\right) v'_{w0} + \sum_{i=1}^{k} \left(\sigma\left(v'_{n_i}\ T_{v_c}\right) - 1\right) v'_{n_i} \tag{5}$$

So, the gradient of the predicted word vector (let's call it the context vector) was applied using the Chain-rule approach over the source context embedding (6).

$$\frac{\partial \mathcal{L}}{\partial v_{w_j}} = \frac{1}{C}\left[\left(\sigma\left(v'_{w0}\ T_{v_c}\right) - 1\right) v'_{w0} + \sum_{i=1}^{k} \left(\sigma\left(v'_{n_i}\ T_{v_c}\right) - 1\right) v'_{n_i}\right] \tag{6}$$

We normalized the context words using a context window width sampled at random from 1 to C_max for each target value in order to fix the issue of inappropriate context vector update. Using the aforementioned technique, the whole dataset was converted into an embedding matrix, which was subsequently used for resampling and feature selection.

C. Feature Resampling

In real-time ecosystems, data imbalance is still a possibility, even with uniformly distributed datasets. To rephrase, there may be a large discrepancy between the numbers of samples reflecting normal or non-heart disease data and those representing incidences of heart disease. Because of the extreme class imbalance that might result from such skewed data, training a machine learning model with such data can lead to inaccurate results.

It can confine the real-time significance or scalability of a CVD prediction model. To alleviate such issues alleviating the data imbalance problem seems to be inevitable. Data sampling has historically made use of a variety of resampling techniques, such as up-sampling, random sampling, and down-sampling, however. The authors have used up-sampling to increase the number of samples from minority classes and down-sampling to decrease the number of samples from majority classes.

On the other hand, in random sampling approaches the number of instances is randomly increased so as to reduce the disparity of the minority and majority class samples. Unfortunately, the aforesaid resampling methods often yields iterative hotspot and hence imbalanced data. The uncontrolled or improper addition of minority class in up-sampling can iteratively cause majority class to become minority and hence the challenge of class-imbalance remains the same. To alleviate this problem, in the recent years a robust method called synthetic minority class oversampling technique (SMOTE) is proposed. The SMOTE method creates synthetic samples that reflect strongly correlated instances or attributes, avoiding any impact on the original sample distribution, in contrast to the resampling procedures mentioned earlier. Using minority samples as input, this method retrieved synthetic samples that were subsequently processed with a k-Nearest Neighbor (k-NN) classifier. In order to pinpoint the most relevant or likely samples in relation to the initial instance or sample, we utilized a k-NN method that is founded on the principle of Euclidean distance. It was a vector connecting the one from the recovered k-neighbors to the one from the present samples. To obtain the final synthetic sample, the produced vector is multiplied by a random integer between 0 and 1, which is then added to the initial sample. We used SMOTE, SMOTE-BL, and SMOTE-ENN, three variants of the SMOTE sampling method, in this study.

In function, SMOTE method applied k-NN method in reference to the original sample to achieve synthesized data.

Though, SMOTE method retrieves fairly distributed samples; however, the randomness of the data, especially over the large feature space results a scene where there can be the probability of multiple instances belong to or ambiguously belong to the multiple classes. It has the potential to affect the

effectiveness of learning and, as a result, produce inaccurate results. We suggest SMOTE-BL to help with it.

Unlike traditional SMOTE resampling method, SMOTE-BL distinguishes the ambiguous data elements present at the common boundaries to label it with specific class and thus suppresses the likelihood of ambiguity. It makes SMOTE-BL more effective over the large heterogenous high-dimensional datasets. Recently, a more evolved SMOTE variant was proposed named SMOTE-ENN which applied Edited Nearest Neighbour (ENN) concept to improve sample distribution. Unlike traditional SMOTE method, where constraining class-boundaries can be difficult due to over-lapping synthetic minority samples (with the majority class), SMOTE-ENN exploits the strength of ENN to classify ambiguous data elements and thus labels them to the appropriate class. It improves accuracy over the traditional k-NNs neighbors. With SMOTE-ENN, any discrepancy between the input sample and allied k-NNs is immediately eliminated from the synthetic sample set. For better learning, it aids in making samples more consistent and significant. By maintaining a dataset with perfectly balanced instances or samples for future learning and classification, a higher value of k accomplishes rigorous cleaning. Therefore, we separately evaluated the effectiveness of SMOTE, SMOTE-BL, and SMOTE-ENN in this study for the purpose of CVD prediction.

D. Feature Selection

The truth is that resampling method(s) enhance data distribution for better learning, but also increase computation in the process. Moreover, over a large feature space training any machine learning method can be exhaustive and hence time-consuming. Additionally, over such high-dimensional feature space, the likelihood of pre-mature convergence and local minima can't be ruled out, and therefore there is the need to apply certain suitable feature selection method which could reduce insignificant or redundant features. This approach can retain optimal set of features and allied instances that consequently can not only improve overall learning but can also alleviate aforesaid issue of convergence, local minima as well as time-exhaustion. It can be vital towards at hand CVD prediction tasks. Though, in the past the authors have suggested applying PCA [], heuristic methods [], etc. towards feature selection; however, their higher reliance over the coefficient values and large iterations limits their scalability and suitability towards at hand CVD prediction task. A variety of feature selection techniques, such as the significant predictor test and cross-correlation analysis (CCRA), have been implemented in Big Data environments and the data mining area. We used three feature selection methods—principal component analysis (PCA), principal component regression analysis (CCRA), and a Mann Whitney-based significant predictor test—with this as our driving force. This section provides an overview of several feature selection methods:

1. Principle Component Analysis (PCA)

In this work, PCA method was applied over the resampled datasets (i.e., SMOTE, SMOTE-BL and SMOTE-ENN dataset) so as to retain the optimal set of the significant features having decisive impact on the CVD prediction results. In this case, we calculated the eigenvalues and principal component for each feature set and data piece based on their covariance. Using a predetermined value of 0.5 for the mean principal component, we calculated the Eigen distance for each feature instance. That is why we kept the characteristics (instances) for future learning and classification and removed the ones with a higher eigen distance.

Those feature instances with the Eigen distance smaller than 0.5 signify higher extent of relatedness or associations. And therefore, such feature instances can have higher impact on the eventual CVD prediction results. Thus, we applied PCA over the different resampled datasets which helped retaining selected set of feature sets for further learning task.

2. Cross-Correlation Analysis (CCRA)

It is a statistical method signifying the extent to which the two variables are associated. In function. CCRA measures association between the two variables representing the extent of relatedness. CCRA also represents the correlation-strengths and allied orientation. Typically, the relationship between the two instances can exist in the range of 1 to 0, where "1" signifies higher relatedness, while the values near to "0" indicates lower association. In this work, we applied Pearson correlation method (7) to calculate correlation coefficient (r).

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})\sum_{i=1}^{n}(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \quad (7)$$

In this work, those feature instances having r larger than 0.5 (i.e., $r > 0.5$) were considered significant and were retained. On the contrary, those feature instance with $r < 0.5$ were dropped from further computation. Thus, this method retained a trimmed set of features for further learning and classification.

3. Significant Predictor Test

This approach takes advantage of the attributes' association with one another to assess how important they are for the current CVD prediction job. This technique evaluates the relevance of each feature instance in relation to a hand CVD prediction task by examining the relationships between them.

More specifically, Mann Whitney method was used as feature selection technique that exploited correlation coefficient between the feature instances to measure their impact on CVD prediction results. In our work, each feature vector was assigned as an autonomous variable, while CVD likelihood was labelled as the dependent variable. Therefore, measuring the level of significance for each feature instance the instance possessing higher level of significance were retained. We fixed p also called level of significance as 0.05, and thus the feature instances with $p > 0.05$ were retained, while remaining feature instances were dropped from further learning and classification.

E. Min-Max Scaler Normalization

To alleviate any probability of over-fitting and convergence over the large non-linear heterogenous features, the input instances (say, selected symbols) were processed for Min-Max scaler normalization in which each data instance was mapped in the range of 0 to 1. We used equation (6) to perform Min-Max normalization over the input features. In (8), x_i represents the feature instances, where $x_i \in N$, which is mapped to the allied normalized value representing x_i'. The proposed Min-Max normalization method obtained the value of x_i' in the range of 0 to 1. Noticeably, in (8), $Min(X)$ and $Max(X)$ signify the minimum and the maximum values of X, correspondingly.

$$Norm(x_i) = x_i' = \frac{x_i - Min\,(X)}{Max(X) - Min\,(X)} \quad (8)$$

F. Heterogenous Ensemble Learning Based Learning and Classification

G. To improve upon the accuracy and reliability of predictions, we developed a strong HEL classification framework that utilized many machine learning techniques as base classifiers. This approach differs from conventional standalone machine learning-based CVD prediction models. It becomes dubious to generalize the performance of these models for healthcare prediction tasks like CVD prediction, given that various machine learning algorithms exhibit varying degrees of accuracy when applied to the same dataset, as mentioned before.

To alleviate this problem, in this work we designed HEL as an ensemble learning framework that embodied the different machine learning classifiers as base classifier, where each base classifier labels each data instance with respective class probability (i.e., Heart Disease Yes-1, No Heart Disease-0). Thus, exploiting aforesaid class labels by each encompassing base classifier a consensus is built by using the maximum voting ensemble (MVE) concept. The data instance (say, patient data) with higher consensus or maximum voting as 1 was eventually predicted as CVD-Positive and labelled as 1. On the contrary, a data instance with more 0 was labelled or predicted as CVD-Negative. In this manner, the use of consensus model helped achieving higher prediction accuracy and reliability towards CAD solution, especially with the generalizability.

To achieve aforesaid HEL learning and prediction, we have applied a total of 13 machine learning classifiers belonging to the different categories including support vector machine-based pattern learning, regression methods, neuro-computing and homogenous ensemble methods (i.e., RF and ETC). As stated above, these machine learning algorithms perform classification distinctly, where their prediction results were subsequently applied to make consensus towards eventual prediction. We applied the following machine learning algorithms as base classifier to perform learning and classification.

1) *Multinomial Naïve Bayes (MNB),*
2) *SVM RBF (Radial Basis Function),*
3) *SVM-Linear,*
4) *SVM-Polynomial,*
5) *Decision Tree (DT),*
6) *Logistic Regression (LOGR),*
7) *K-NN,*
8) *AdaBoost,*
9) *Gradient Boost,*
10) *Bagging (k-NN Kernel),*
11) *Bagging (MNB Kernel),*
12) *Random Forest (RF), and*
13) *Extra Tree Classifier (ETC).*

A brief of these base classifiers is given as follows:

a) Naïve Bayes (NB)

The Naïve Bayes classifier employs Bayes' rules for pattern learning and classification and is one of the most popular and extensively studied probabilistic classifiers. The "independent feature model" postulates that the linked features continue to function independently of one another and, as a result, do not impact the classification results; NB is probabilistic in character. Another tenet of this pattern-learning strategy is that two feature instances in the same class cannot possibly be connected. According to the Bayes' rule, which is provided in (9), it assigns a data instance x to the class e^*=argmax_d P(d|x) in light of the hypotheses mentioned earlier.

$$P(d|x) = \frac{P(x|d)P(d)}{P(x)} \quad (9)$$

The likelihood of data instance x is represented as P(d|x), while the probability of class c is stated in (9).

Here, $P(x)$ signifies the predictor prior probability, which is measured as per the equation (10).

$$P(x|d) = \prod_{l=1}^{m} P(x_l|d) \quad (10)$$

Though, NB algorithm has evolved with the different kernels like Gaussian, Multinomial; however, we applied NB-Multinomial (MNB) as classifier to perform two-class classification. Here, MNB performed learning over the count's frequency, signifying x_i occurrences over n trails. It applies the occurrence(s) of the binary terms so as to predict each instance as Heart Disease Positive (say, CVD positive) or CVD negative and labelled it as 1 and o, correspondingly.

b) Support Vector Machine (SVM)

Pattern recognition and classification are two areas where support vector machines (SVMs), a type of supervised machine learning model, have found widespread use. Support vector machines (SVMs) are the go-to pattern learning approach for text and picture classification issues due to its hyper-plane learning and classification capabilities. As a non-probabilistic binary classifier, SVM learns on the normalised dataset in this study. The learning and classification processes are carried out by iteratively minimising the generalisation error over the input feature space. We calculated the hyper-plane support vector, which represents a training subset that shows the boundary conditions. In keeping with the two-category classification problem at hand, the support vector model was used to construct the hyper-plane between the two classes, positive and negative, of cardiovascular disease (CVD). It applied equation (11) to perform classification.

$$Y' = w * \phi(x) + b \quad (11)$$

In (11), parameter $\phi(x)$ states a non-linear transform that emphasizes on the allocation of the appropriate weights w and bias value b to perform learning and classification. We measured classification result Y' by reducing a regression-risk parameter, defined in equation (12).

$$R_{reg}(Y') = C * \sum_{i=0}^{l} \gamma(Y_i' - Y_i) + \frac{1}{2} * \|w\|^2 \quad (12)$$

The penalty factor and cost-function are represented by the parameters C and γ, respectively, in equation (12). Following the protocol in (13), we determined the weight scores.

$$w = \sum_{j=1}^{l} (\alpha_j - \alpha_j^*)\phi(x_j) \quad (13)$$

Here, α and α^* be the non-zero values, called Lagrange relaxation. Thus, the applied SVM model results prediction output as (14).

$$Y' = \sum_{j=1}^{l} (\alpha_j - \alpha_j^*)\phi(x_j) * \phi(x) + b = \sum_{j=1}^{l} (\alpha_j - \alpha_j^*) * K(x_j, x) + b \quad (14)$$

In (14), $K(x_j, x)$ states a kernel function. In this work, SVM was applied with three different kernel functions, including SVM-Linear, SVM-Polynomial and SVM-RBF. Here, each variant performed independently to classify each input data.

c) Decision Tree (DT)

Data mining and classification jobs often make use of DT, one of the most applied association rule mining methodologies. The CART, ID3, C4.5, and C5.0 association rule mining methods have all contributed to the development of this machine learning model. Both solo classifiers and ensemble-learning methods, such random forests and additional tree classifiers, have made use of the DT algorithm. Starting at the root node, it applies an association rule with a split-condition to split the input feature instances into numerous branches, one for each node in the tree. This is how it serves its functional purpose. After that, it learns and classifies the pattern or data by applying the information gain ratio (IGR) technique over each branch. Input features can be easily divided into numerous branches, and the system will automatically obtain the other nodes that will branch off into other data.

In this manner, this approach looks like a tree structure having multiple branches. The DT algorithm resembles a binary tree possessing single root or parent node having multiple children's nodes.

Let, the left and the right child node be LC_d and RC_d, respectively. Consider, x be the input feature, while I be the noise value. Now, with the available samples in P_d, LC_d and RC_d, DT intends to optimize information gain, iteratively by using (15).

$$Information\ Gain\ (P_d x) = I(P_d) - \frac{LC_n}{P_n} I(L.C_d) - \frac{RC_n}{P_n} I(R.C_d) \quad (15)$$

In (15), I can be calculated by using any of the methods like Entropy I_H (16), Gini-Index I_G (17), and classification error I_E (18).

$$I_H(n) = -\sum_{i=1}^{c} p(c|n) \log_2 p(c|n) \quad (16)$$

$$I_G(n) = 1 - \sum_{i=1}^{c} p(c|n)^2 \quad (17)$$

$$I_E(n) = 1 - max\{p(c|n)\} \quad (18)$$

Parameters c and n in (16-18) denote the class(es) and corresponding node(s), respectively. The probability factor was determined by dividing c by n.

Once obtaining the predicted output, each sample was classified into two classes, CVD-Positive and CVD Negative, and was labelled as 1 and 0, respectively.

d) Logistic Regression

By transforming input feature sets into independent variables, logistic regression runs regressions on such sets. On the contrary, it defined feature's CVD probability as dependent variable. The proposed LOGR prediction method applied (19) as regression function.

$$\text{logit}[\pi(x)] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots \ldots + \beta_m X_m \quad (19)$$

The dependent variable is represented by logit[π(x)] in equation (19), while the independent variable is x_i. In this case, the binary outputs were translated using the logit function, which produces different values of π(x) from 0 to 1, negative infinity to positive infinity. In the previous equation. In terms of π, the CVD probability was found to be (20), while m represents the overall independent variables.

$$\pi(x) = \frac{e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots\ldots+\beta_m X_m}}{1 + e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots\ldots+\beta_m X_m}} \quad (20)$$

Thus, applying above discussed LOGR method, each data instance or sample (say, patients bio-physiological data) was classified as CVD Positive and CVD Negative, which was labelled as "1" and "0", respectively.

e) AdaBoost

It is a type of adaptive boosting ensemble learning technique having better instance-wise learning and analysis capability. In order to implement the AdaBoost algorithm, certain weak learners were formed by assigning equal weight to the associated prerequisite exams. As a result of measuring the error rate for the previously mentioned weak classifier for each cycle, AdaBoost improved the weight for the correctly categorised samples and iteratively reduced the weights for the incorrectly classified samples.

Eventually, the weak learner turned out to be the strong learner and thus classified each sample or data as CVD Positive and CVD negative and labeled them as "1" and "0", correspondingly. Unlike traditional AdaBoost method, the gradient boosting method focuses on improving weight parameters more efficiently and thus enables more accurate sample classification (say, data classification). Unlike fixed update method, gradient boosting method tunes learners adaptively and thus achieves more detailed learning to yield higher accuracy. This approach helps in suppressing any probable convergence problem. Despite increased computational cost, it performs superior over the classical AdaBoost or Boosting ensemble. With this motivation, in addition to the AdaBoost method, we applied gradient boosting ensemble as well as a base classifier.

f) Bagging (k-NN) and Bagging (MNB)

We applied bagging method with the two different kernels (also called, base classifiers). Particularly, bagging ensemble was designed with k-NN and MNB classifier as their base classifier to perform two-class classification. Thus, these bagging ensemble methods or variants classified each data as CVD Positive and CVD Negative and labelled them as 1 and 0, respectively.

g) Random Forest (RF)

When it comes to ensemble approaches that use numerous tree-based classifiers, Random Forest is a popular choice. Because it is a tree-based learning system, every tree comes up with its own best guess for the most likely class. Assume that N is the input training set. Afterwards, RF uses the input samples or data to randomly select a sample with N cases. In order to build a new tree, these samples are used as the training set. If we use M to represent the input data, we can divide the node using the optimal split applied to m. During the execution of forest growth, we fixed the value of m. This is how it grows each tree to its maximum potential. Lightweight and computationally efficient, the RF algorithm outperforms conventional classifiers thanks to its reduced parameter requirements for learning and classification. The RF method is a mathematically determined mixture of many tree-structures that are based on the aforementioned forest development mechanism (21).

$$\{h(x, \theta_k), k = 1,2, \ldots i \ldots\} \quad (21)$$

In (21), h be the classification function, while the random vector generated throughout tress is given as $\{\theta_k\}$. In this method, each tree contains a unit vote for its most probable class. Noticeably, the capability to use multiple DTs where each (say, unit) DT acts as a distinct classifier enables RF behaving as a bootstrapped learning model to perform consensus (based on the multiple DT classifiers)-based learning and classification. In this work, we used a bootstrapped subset of training samples to train each tree throughput the constituted forest, where it applies 70% sample for training, while the remaining samples are labeled as the out-of-bag samples. These out-of-bag samples are later used for inner cross-validation to perform eventual prediction. In this manner, each data sample was classified as CVD Positive and CVD Negative samples and labelled as "1" and "0", respectively.

h) Extra Tree Classifier (ETC)

This is also an ensemble learning variant which constitutes a cluster of the unpruned DTs on the basis of the traditional top-down mechanism. Unlike RF technique, ETC method contains randomization of the data samples and cut-point selection while performing node-split. ETC algorithm is capable of constituting overall randomized trees encompassing structures which are independent of the outputs of the training sample. There are two main features that set this ensemble learning model apart from competing tree-based ensemble approaches. The first is that it uses random cut-point selection to partition nodes, and the second is that it uses the entire training set to carry out tree-growth or forestation. In this case, the prediction output (i.e., CVD class) was generated by combining the classified results from the trees using the MVE approach. When compared to other machine learning models that use weaker randomisation methods, the ETC ensemble model's main concept and functional components—ensemble averaging and overall cut-point and attribute randomization—reduce variance more appropriately. Furthermore, to get better (prediction) accuracy, using the original training samples decreases the chance of the bias-probability.

Thus, applying this mechanism the proposed ETC model classified each sample or data into two-classes, CVD Positive and CVD Negative and labelled them as "1" and "0", respectively.

Thus, applying above discussed machine learning models as the base classifier(s), each data sample was classified into two classes; CVD-Positive and CVD-Negative and labelled them as "1" and "0", respectively. Since, we applied a total of 13 machine learning models in parallel, where ach model provided unit predicted output or labels, applying consensus model (also called the maximum voting ensemble (MVE)) each sample was classified as CVD-Positive or CVD-Negative. In MVE method, a data sample with more than or equal to seven "1s" was predicted as CVD-Positive, while a data sample with seven or more "0s" was predicted and labelled as CVD-Negative. Thus, applying this method, each data sample (say, patient data) was predicted as CVD-Positive or CVD-Negative. Accuracy (%), precision (%), recall (%), and F-Measure were some of the statistical measures retrieved in order to evaluate performance and dependability. What follows is a presentation of the simulation findings along with related conclusions.

VI. Results and Discussion

In this work, we developed a robust intrinsically modified Bio-Physiological Features Driven Heterogenous Ensemble Learning-based Heart Disease Prediction Model or Cardio Vascular Disease (CVD) Prediction Model. As the name indicates to design targeted CAD solution, patient's specific biological as well as physiological features including lipid profile, cholesterol, ECG pattern, stroke events, gender, etc. were taken into consideration. More specifically, UCI Cleveland dataset along with the Kaggle datasets were taken into consideration. A total of 14 features were taken into consideration encompassing the different biological as well as physiological clinical measurements. The input data was at first processed for pre-processing where outlier analysis and missing element problem was solved. Subsequently, Word2Vec word-embedding method was applied to generate the corresponding embedded matrix. It helped retrieving the latent/semantic features to perform further learning and classification. It also helped in addressing the problem of long-term dependency that eventually improved learning and classification. The embedded matrix from each feature set was processed for resampling techniques including SMOTE, SMOTE-BL and SMOTE-ENN algorithms. It helped improving overall sample distribution by supressing any likelihood of class-imbalance. The resampled data was then processed for feature selection by applying PCA, CCRA and Mann-Whitney Significant Predictor Test (SPT). Noticeably, these feature selection methods were applied distinctly over the resampled datasets so as to assess relative (performance) efficacy. Additionally, it helped in identifying the optimally performing resampling and feature selection model to achieve optimal CAD solution for CVD prediction. The selected features were then processed for Min-max normalization which mapped each data instance in the range of 0 to 1, and thus alleviated any probability of over-fitting and convergence. Unlike traditional standalone machine learning-based classification models, in this work an HEL was designed by applying machine learning of the different categories including SVM variants (SVM-Linear, SVM-RBF, SVM-Polynomial), DT, NB, LOGR, LR, ANN-LM, RF and ETC, as base classifiers. A total of 13 machine learning algorithms were applied as base classifier that in conjunction with MVE ensemble performed eventual prediction and classified each subject's class as CVD-Positive and CVD-Negative, which was labelled as "1" and "0", respectively. True positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) are the confusion metrics that we obtained. F-Measure, recall, accuracy, and precision are some of the statistical performance metrics that were derived from these inputs. Table II provides the performance characteristics together with their related statistical derivations. The MATLAB 2020b program was used for the overall model design, and the simulation was conducted on a computer system with Microsoft Windows operating systems, an Intel i5 processor, 8 GB RAM, and a frequency of 3.2 GHz.

Table II: Performance parameters

Parameter	Mathematical Expression
Accuracy	$\frac{(TN+TP)}{(TN+FN+FP+TP)}$
Precision	$\frac{TP}{(TP+FP)}$
Recall	$\frac{TP}{(TP+FN)}$
F-measure	$2.\frac{Recall.Precision}{Recall+Precision}$

Intra-Model Assessment and Inter-Model Assessment are used to characterise the overall performance. A variety of resampling techniques, feature selection algorithms, and classifiers were used in this intra-model evaluation to see how well the suggested CVD-prediction model performed. Conversely, the suggested model's relative efficacy was evaluated using inter-model assessment in comparison to other state-of-the-arts. The following sections provide a comprehensive analysis of the suggested model.

A. Intra-Model Assessment

This study compares and contrasts the performance of various resampling strategies, feature selection methods, and base classifiers, including the suggested MVE ensemble. What this means is that we tested how well various feature resampling, feature selection, and classification models works.

In this work performance characterization was done with respect to the resampling and feature selection methods. The simulation results obtained are given as follows:

1. Resampling Methods Assessment

Table III presents the performance outputs with the different resampling methods. To assess whether the use of resampling methods (i.e., SMOTE resampling methods) can improve CVD prediction accuracy, we considered both resampled data as well as the original dataset. In this manner, four different datasets including original dataset, SMOTE, SMOTE-BL and SMOTE-ENN were assessed for their respective performance. The simulation results reveal that the original dataset shows the highest accuracy of 95.21%, precision of 96.03%, recall 94.16% and F-Measure (%) of 95.10%. On the contrary, the classical SMOTE resampling method shows the accuracy of 95.77%, precision 96.66%, recall 95.55% and F-Measure of 96.10%. This result clearly indicates that the use of SMOTE resampling yields higher accuracy than the original dataset. SMOTE-BL on the other hand shows CVD prediction accuracy of 97.94%, precision 96.91%, recall 95.97% and F-measure of 96.43%. The simulation also inferred that the use of SMOTE-ENN method achieved accuracy of 99.87%, precision of 99.32%, recall of 96.88% and F-Measure of 98.02%. These results state that undeniably, unlike traditional approaches where the original data are passed to the classifiers or prediction, the use of resampling method(s) can yield superior results. The highest accuracy with the original data was found 95.21%, while SMOTE-ENN resampled data exhibited an accuracy of 99.87%, which is almost 4.7% higher than the traditional (without resampling) method. It shows that the use of SMOTE-ENN resampling can achieve the best performance towards targeted CVD prediction tasks. The relative assessment revealed that SMOTE-ENN method performs superior over SMOTE and SMOTE-BL method. It confirms the use of SMOTE-ENN efficacy towards real-time CVD prediction tasks.

Table III Feature Resampling Performance

Data	Accuracy (%)	Precision (%)	Recall (%)	F-Measure (%)
Original Data	95.21	96.03	94.16	95.10
SMOTE	95.77	96.66	95.55	96.10
SMOTE-BL	97.94	96.91	95.97	96.43
SMOTE-ENN	99.87	99.32	96.88	98.02

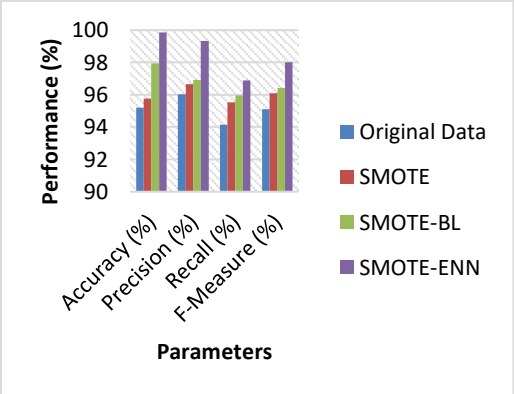

Fig. 1 Performance over the different sampling techniques

2. Feature Selection Method Assessment

To improve analytics model's accuracy feature selection methods have performed superior over the original dataset. We used PCA, CCRA, and MW-SPT (Mann Whitney Significant Predictor Test) as feature selection methods in this reference. In this case, finding the top feature for VCD prediction was our primary goal. Based on the findings of the simulation, it can be inferred that the original data, which did not undergo processing feature selection, displays an F-Measure of 95.08%, an accuracy of 95%, a precision of 95.14%, and a recall of 95.03%. In contrast, models

that are selected using principal component analysis (PCA) show recall, accuracy, precision, and F-Measure values of 96.24%, 96.76%, 95.74%, and 96.24%, respectively. In contrast, the CCRA approach demonstrated an F-Measure of 96.35%, a recall of 94.97%, a precision of 96.73%, and an accuracy of 96.67% in predicting CVD. The MW-SPT feature selection approach demonstrated an F-Measure of 99.32% and a CVD-prediction accuracy of 98.31%.

Noticeably, the simulation results indicate that the use of feature selection methods (i.e., PCA, CCRA and MW-SPT) can achieve superior results than the original data-based analytics. Interestingly, amongst the aforesaid feature selection methods the use of MW-SPT algorithm exhibits the highest accuracy of 98.31%, precision of 99.84%, recall 98.81% and F-Measure of 99.32%. Though, other feature selection methods exhibit higher (CVD-prediction) accuracy than the original feature-based model; however, amongst the all-feature selection methods applied, MW-SPT exhibited the highest CVD prediction accuracy (98.31%). The results clearly indicate that the ability to address ambiguous data elements over SMOTE makes SMOTE-ENN more effective and hence yields higher accuracy. SMOTE-ENN applies ENN as an additional machine learning approach to retain only those feature instance having high correlated-ness and significance and thus improved data quality. Consequently, it strengthens MW-SPT method that achieves superior performance towards at hand CVD prediction.

Table IV Feature Selection Method Performance

Data	Accuracy (%)	Precision (%)	Recall (%)	F-Measure (%)
Original Data	94.99	95.14	95.03	95.08
PCA	95.23	96.76	95.74	96.24
CCRA	96.67	96.73	95.97	96.35
MW-SPT	98.31	99.84	98.81	99.32

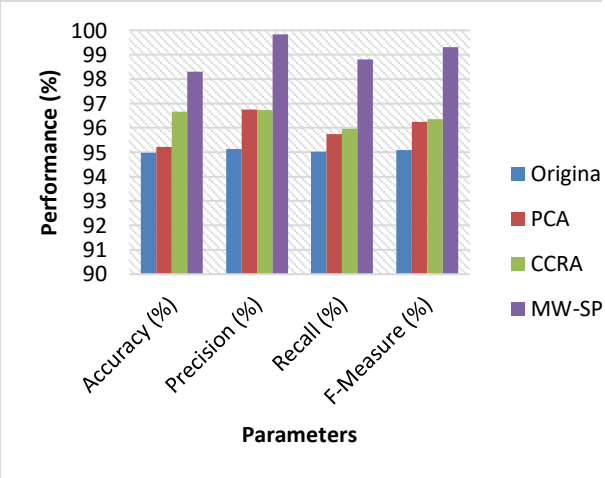

Fig. 2 Performance over the different feature selection techniques

Classification Model Assessment

Here, we compared the performance of various machine learning classifiers both as independent tools and as part of an MVE ensemble learning framework. The overall effort is made to assess whether the standalone method can perform better or the use of MVE-based HEL model can yield superior efficacy towards CVD-prediction. Furthermore, this quantification of performance can aid in comparing various machine learning methods for certain CVD-prediction jobs. According to the simulation results, the MNB machine learning model has an accuracy of 94.48% and the DT method has an accuracy of 95.01% when it comes to CVD predictions.

On the other hand, the SVM algorithms variants exhibits the highest accuracy of 94.44%, 95.34% and 96.01% by SVM-Lin, SVM-Poly and SVM-RBF, respectively. The logarithmic regression method (LOGR) exhibited the CVD-prediction accuracy of 96.66%, while ANN performed the prediction accuracy of 96.50%. The LR method exhibited CVD-prediction accuracy of 95.06%, while Bagging k-NN and Bagging-NB shows the prediction accuracy of 94.88% and 95.21%, correspondingly. AdaBoost on the other hand shows CVD-prediction accuracy of 95.95%. the other variants of ensemble learning methods including RF, ETC and the proposed MVE-HEL model exhibited the CVD-prediction accuracy of 98.81%, 99.70% and 99.93%, correspondingly. In comparison to other independent machine learning techniques, the suggested MVE-HEL method performs better in the aggregate. By comparison to the other machine learning methods, the suggested MVE-HEL method achieved better CVD-prediction accuracy (99.93%), precision (99.69%), recall (99.53%), and F-Measure (99.60%). The results show that the suggested MVE-HEL approach is reliable for CVD prediction. With an F-Measure of 0.996, the suggested CVD-prediction model is viable and scalable for use in practical CAD applications. The suggested CVD-prediction approach is robust towards real-world CAD applications, as shown by the other parameters as well.

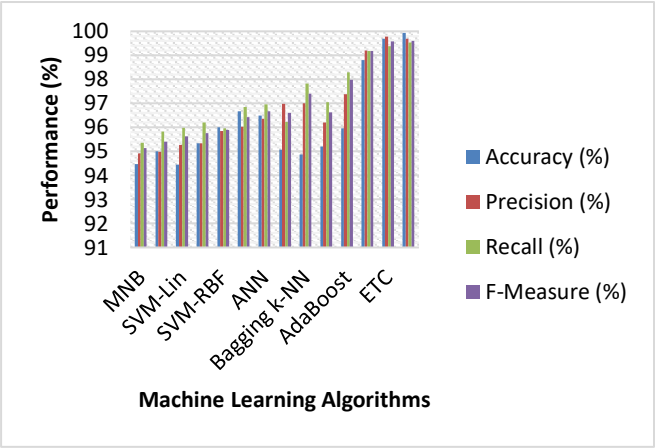

Fig. 3 Performance over the different machine learning classifiers

Table V Classification Method Performance

Machine Learning Model	Accuracy (%)	Precision (%)	Recall (%)	F-Measure (%)
MNB	94.48	94.92	95.35	95.13
DT	95.01	94.99	95.82	95.40
SVM-Lin	94.44	95.27	95.99	95.62
SVM-Poly	95.34	95.33	96.21	95.76
SVM-RBF	96.01	95.85	95.95	95.90
LOGR	96.66	96.02	96.85	96.43
ANN	96.50	96.37	96.96	96.66
LR	95.06	96.99	96.22	96.60
Bagging k-NN	94.88	97.01	97.83	97.41
Bagging-NB	95.21	96.21	97.04	96.62
AdaBoost	95.95	97.38	98.30	97.98
RF	98.81	99.20	99.18	99.18
ETC	99.70	99.78	99.38	99.57
MVE	99.93	99.69	99.53	99.60

B. Inter-Model Assessment

We conducted an inter-model evaluation to see how well the proposed CVD-prediction model performed in comparison to the other state-of-the-arts. The suggested analytics (CAD) solution was pitted against the other current CVD-prediction models in this method.

Table VI Relative performance comparison

Ref.	Year	Dataset	Classifier	Methodology	Accuracy (%)
[61]	2019	Cleveland	NB, C4.5, MLP, PART, Bagging, Boosting, Majority Voting, Stacking	Bagging and Boosting ensemble	85.48
[62]	2020	Cleveland	LR, SVM, k-NN	Feature normalization and dimensional reduction using PCA	87.00
[63]	2020	Cleveland	LR, DT and Gaussian NB (GNB)	Singular Value Deposition (SVD)-based dimensionality reduction	82.75
[64]	2022	Cleveland	SVM, NB, ConvSGLV and ensemble learning	CNN feature extraction and MVE voting for prediction	93.00
[65]	2023	Cleveland	LR, k-NN, DT, XGB, SVM, RF	GridSearchCV hyper-parameter tuning	87.91

[66]	2023	Cleveland	LR, k-NN, NB, RF, GB, AB, SVE	Soft voting ensemble	93.44
[13]	2023	Cleveland	LR, NB, k-NN, SVM, DT, RF, MLP	Feature reduction	90.00
[10]	2023	Kaggle	SVM, k-NN, RF, LOGR	IQR, CCRA	99.00
[67]	2023	Kaggle	DT, PCA	PCA feature reduction	98.00
[68]		Framingham	SVM, MLP, RF	Ensemble	97.13
Proposed	2023	Cleveland	SVM-Lin, SVM-Poly, SVM-RBF, DT, MNB, k-NN, LOGR, LR, Bagging, Boosting, AdaBoost, RF, ETC, MVE	Word2Vec embedding, SMOTE-ENN resampling, Mann-Whitney Significant Predictor Test, Min-Max normalization, MVE-HEL ensemble.	99.93

Observing the results (Table VI), it can be found that though a number of literatures have applied machine learning methods towards CVD prediction; however, the state-of-arts methods show poor accuracy than the proposed intrinsically modified or improved HEL-MVE based analytics solution for CVD prediction. Noticeably, to make justifiable and generalizable performance comparison same dataset (i.e., Cleveland) was taken into consideration. Moreover, the recent works are used to perform relative performance analysis. In [61], the different machine learning methods including NB, C4.5, MLP, PART, Bagging, Boosting, Majority Voting, Stacking were applied to perform CVD prediction. The authors applied ensemble learning approach to improve CVD-prediction accuracy, where it exhibited the highest prediction accuracy of 85.48%. On the other hand, the authors [62] applied LR, SVM, k-NN machine learning model, where the input features were normalized followed by PCA feature reduction. It achieved the CVD-prediction accuracy of 87%, which is almost 12.7% lower than the proposed CVD-prediction model. With a maximum accuracy of 82.75%—nearly 12% lower than the suggested CVD-prediction model—the authors of [63] used LR, DT, and the Gaussian NB (GNB) approach in combination with a dimensionality reduction method based on Singular Value Deposition (SVD).

The authors [64] applied the different machine learning model including SVM, NB, ConvSGLV and ensemble learning where CNN (convolutional neural network) was applied to perform feature extraction where the extracted features were processed for MVE consensus-based prediction towards CVD prediction. The highest prediction accuracy obtained was 93%. The authors [65] applied LR, k-NN, DT, XGB, SVM, RF machine learning models with GridSearchCV hyper-parameter tuning to perform heart disease prediction. The highest prediction accuracy obtained was 87.91%. On the other hand, in [66] applied the different machine learning methods including LR, k-NN, NB, RF, GB, AB, SVE algorithms. The authors applied soft-voting ensemble to perform CVD prediction, where the highest CVD-prediction results exhibited the prediction accuracy of 93.44%. Interestingy, unlike aforesaid state-of-arts, we performed both data optimization or feature optimization as well as computational enhancement to achieve higher prediction accuracy. More specifically, we performed Word2Vec embedding, SMOTE-ENN resampling, Mann-Whitney Significant Predictor Test, Min-Max normalization, MVE-HEL ensemble, which was designed by applying SVM-Lin, SVM-Poly, SVM-RBF, DT, MNB, k-NN, LOGR, LR, Bagging, Boosting, AdaBoost, RF, ETC as base classifier to constitute MVE-HEL to perform eventual prediction. Outperforming all existing state-of-the-art models, our suggested model demonstrated an unprecedented level of accuracy at 99.93%. It proves that the suggested model is stable enough for real-time CVD prediction. The authors of [13] used a variety of ML techniques, such as LR, NB, k-NN, SVM, DT, RF, and MLP. Their thorough evaluation showed that RF ensembles could achieve a maximum CVD-prediction accuracy of 90%; however, this was still lower than the suggested model's prediction accuracy of 99.7 percent.

Their highest F-Measure performance obtained was 90.91%, which still falls below the proposed model’s output (F-Measure 99.60%). Similarly, in [10], the authors applied SVM, k-NN, LOGR and RF algorithms to perform CVD prediction over Kaggle dataset. The authors claimed that the use of RF ensemble learning model achieves the highest prediction accuracy of 99%. To improve feature the authors applied inter-quartile range (IQR) method, correlation and significant assessment altogether. With PCA feature selection, the authors [67] claimed to have achieved prediction accuracy of 98%. Yet, it failed in addressing numerous challenges including data-imbalance, over-fitting and convergence. RF ensemble was applied in [68], where it exhibited an accuracy of almost 98%. Despite such efforts, the relative performance characterization confirms that the proposed approach exhibits superior over the state-of-arts and hence can serve a potential, reliable and scalable CAD solution for CVD prediction.

VII. CONCLUSION

In sync with the high pace rising cases of cardiovascular diseases and resulting mortality rate, academia-industries have been making efforts to achieve data mining-based automated heart disease prediction solution. To achieve it, the use of the different bio-physiological parameters including age, gender, cholesterol, insulin, lipids profile, stroke history etc. information has been applied extensively in the past. However, the lack of sufficient and suitable datasets, likelihood of class-imbalance and other computational complexities including local minima and convergence limit the reliability of the at hand solution. Additionally, the majority of the existing methods are limited due to low accuracy that confines its scalability towards real-time CAD solutions or CAD prediction solutions. Considering it as motivation this research focused on designing a robust intrinsically modified bio-physiological parameters driven heterogenous ensemble learning based heart disease (say, CVD) prediction model. As the name indicates the proposed CAD model emphasized on improving both feature as well as computational aspects so as to enable scalable heart disease prediction model. In this work, at first benchwork dataset encompassing age, gender, cholesterol, protein profiles, BMI, stoke profile or history, ECG information etc. to design a robust heart disease prediction. Following data collection, outlier analysis was performed on a set of thirteen characteristics. Word2Vec embedding was then applied to the input data in order to improve the features. This method converted the aforementioned features into an equivalent embedded matrix. Next, the embedded matrix was subjected to resampling processing utilising the SMOTE, SMOTE-BL, and SMOTE-ENN algorithms, all of which contributed to reducing the likelihood of class imbalance. After resampling, the features were passed via PCA, CCRA, and a significant predictor test based on the Wilcoxon Rank Sum Test in order to pick the features. The chosen feature was subsequently subjected to Min-Max normalisation, which aimed to reduce the chances of over-fitting and convergence by mapping input features between 0 and 1. Lastly, instead of using a single machine classifier as a basis for learning and prediction like in the past, the suggested model uses a heterogeneous ensemble learning approach. This includes base classifiers such as SVM, DT, NB, LOGR, LR, ANN-LM, RF, and ETC. In order to solve the two-class classification problem of consensus-based CVD prediction, the suggested model used the maximum voting ensemble (MVE). It assigned a value of 0 for "normal person" and a value of 1 for "CVD probable person" for every case.

The depth performance analysis revealed that the proposed model exhibited the highest performance in conjunction with SMOTE-ENN resampling, WRST significant predictor test-based feature selection, Min-Max normalization and the proposed heterogenous MVE ensemble model. The performance characterisation of the proposed model for predicting heart disease shows that it is robust for real-world CAD or CDS applications, with a prediction accuracy of 99.93%, precision of 99.69%, recall of 99.53%, and F-Measure of 99.60%—all higher than the other state-of-the-art models.

REFERENCE

[1] S. Mishra, P. K. Mallick, H. K. Tripathy, A. K. Bhoi, and A. González-Briones, "Performance evaluation of a proposed machine learning model for chronic disease datasets using an integrated attribute evaluator and an improved decision tree classifier," Appl. Sci., vol. 10, no. 22, p. 8137, 2020.

[2] J. Ahamed, A. M. Koli, K. Ahmad, M. A. Jamal, and B. B. Gupta, "CDPS-IoT: Cardiovascular Disease Prediction System Based on IoT Using Machine Learning", CDPS-IoT: Cardiovascular Disease Prediction System Based on IoT Using Machine Learning, 2021, pp. 78-86. [3] Bhatia, M.; Sood, S.K. Game Theoretic Decision Making in IoT-Assisted Activity Monitoring of Defence Personnel. Multimed. Tools Appl. 2017, 76, 21911–21935.

[4] Simpao, A.F.; Ahumada, L.M.; Gálvez, J.A.; Rehman, M.A. A Review of Analytics and Clinical Informatics in Health Care. J. Med. Syst. 2014, 38, 45.

[5] cardiovascular diseases. Available online: https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1 (accessed on 14 June 2022).

[6] D. S. AbdElminaam, N. Mohamed, H. Wael, A. Khaled, and A. Moataz, "MLHeartDisPrediction: Heart Disease Prediction using Machine Learning", Journal of Computing and Communication Vol.2, No.1 , PP. 50-65 , 2023.

[7] Juma PA, et al. non-communicable disease prevention policy process in five African countries authors. BMC Publ Health 2018;18(Suppl 1).

[8] World Health Organization. Global status report on noncommunicable diseases 2014. 2014.

[9] WHO. Western Pasific regional action plan of concummunicable diseases. World Heal. Organ.; 2014.

[10] K. Sumwiza, C. Twizere, G. Rushingabigwi, P. Bakunzibake, P. Bamurigire, "Enhanced cardiovascular disease prediction model using random forest algorithm", Informatics in Medicine Unlocked (41), 2013, 101316, pp. 1-9.

[11] Mijwil MM., Al-Mistarehi AH., Aggarwal K: The effectiveness of utilising modern artificial

intelligence techniques and initiatives to combat COVID-19 in South Korea: a narrative review. Asian J. Appl. Sci. 9(5) (2021). (ISSN: 2321-0893)

[12] Madjid, M., Safavi-Naeini, P., Solomon, S.D., Vardeny, O.: Potential effects of coronaviruses on the cardiovascular system: a review. JAMA Cardiol. 5(7), 831–840 (2020)

[13] M. I. Hossain, M. H. Maruf, M. A. R. Khan, F. S. Prity, S. Fatema, M. S. Ejaz, M. A. S. Khan, "Heart disease prediction using distinct artificial intelligence techniques: performance analysis and comparison", Iran Journal of Computer Science (2023) 6:397–417.

[14] Soni, J., Ansari, U., Sharma, D., Soni, S.: Predictive data mining for medical diagnosis: an overview of heart disease prediction . Int. J. Comput. Appl. 17(8), 43–48 (2011)

[15] Dai, H., Much, A.A., Maor, E., Asher, E., Younis, A., Xu, Y., Lu, Y., Liu, X., Shu, J., Bragazzi, N.L.: Global, regional, and national burden of ischaemic heart disease and its attributable risk factors, 1990–2017: results from the global burden of disease study 2017. Eur. Heart J.Qual. Care Clin. Outcomes 8(1), 50–60 (2022)

[16] Mozaffarian, D.; Benjamin, E.J.; Go, A.S.; Arnett, D.K.; Blaha, M.J.; Cushman, M.; de Ferranti, S.; Després, J.-P.; Fullerton, H.J.; Howard, V.J.; et al. heart disease and stroke statistics—2015 update: A report from the American Heart Association. Circulation 2015, 131, e29–e322.

[17] Maiga, J.; Hungilo, G.G.; Pranowo. Comparison of Machine Learning Models in Prediction of Cardiovascular Disease Using Health Record Data. In Proceedings of the 2019 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS), Jakarta, Indonesia, 24–25 October 2019; pp. 45–48.

[18] Khan MA. An IoT framework for heart disease prediction based on MDCNN classifier. IEEE Access 2020; 8:34717–27

[19] Pal M, Parija S. Prediction of heart diseases using random forest. J Phys Conf Ser 2021;1817(1).

[20] Durairaj M, Revathi V. Prediction of heart disease using back propagation MLP algorithm. Int J Sci Technol Res 2015;4(8):235–9.

[21] Rani P, Kumar R, Ahmed NMOS, Jain A. A decision support system for heart disease prediction based upon machine learning. J Reliab Intell Environ 2021;7(3): 263–75. https://doi.org/10.1007/s40860-021-00133-6.

[22] Nancy P, Swaminathan B, Navina K, Nandhine B, Lokesh P. Tuned Random Forest Algorithm for Improved Prediction of Cardiovascular Disease, 1; 2020. p. 1355–60.

[23] Almustafa KM. Prediction of heart disease and classifiers' sensitivity analysis. BMC Bioinf 2020;21(1):1–18.

[24] Akyol K, Çalik E, Bayir S¸ , S¸ en B, Çavus¸oılu A. Analysis of demographic characteristics creating coronary artery disease susceptibility using random forests classifier. Procedia Comput Sci 2015;62(Scse):39–46. https://doi.org/10.1016/j. procs.2015.08.407.

[25] Pal M, Parija S. Prediction of heart diseases using random forest. J. Phys. Conf. Ser. 2021;1817(1).

[26] Su X, et al. Prediction for cardiovascular diseases based on laboratory data: an analysis of random forest model. J Clin Lab Anal 2020;34(9):1–10. https://doi.org/ 10.1002/jcla.23421.

[27] S. Palaniappan and R. Awang, "Intelligent Heart Disease Prediction System Using Data Mining Techniques," Int. J. Comput. Sci. Netw. Secur., vol. 8, no. 8, pp. 343–350, 2008. https://doi.org/10.1109/AICCSA.2008.4493524.

[28] R. Asha and R. G. Sophia, "Diagnosis Of Heart Disease Using Datamining Algorithm," Glob. J. Comput. Sci. Technol., vol. 10, no. 10, pp. 1–6, 2010.

[29] M. Shouman, T. Turner, and R. Stocker, "Using decision tree for diagnosing heart disease patients," in Proceedings of the Ninth Australasian Data Mining Conference-Volume 121, 2011, pp. 23–30.

[30] N. Bhatla and K. Jyoti, "An analysis of heart disease prediction using different data mining techniques," Int. J. Eng., vol. 1, no. 8, pp. 1–4, 2012.

[31] C. Dangare and S. Apte, "A data mining approach for prediction of heart disease using neural networks," Int. J. Comput. Eng. Technol., vol. 3, no. 3, 2012.

[32] S. U. Amin, K. Agarwal, and R. Beg, "Genetic neural network- based data mining in prediction of heart disease using risk factors," in 2013 IEEE Conference on Information & Communication Technologies, 2013, pp. 1227–1231. https://doi.org/10.1109/CICT.2013.6558288.

[33] V. Chaurasia and S. Pal, "Early prediction of heart diseases using data mining techniques," Caribb. J. Sci. Technol., vol. 1, pp. 208–217, 2013.

[34] D. Chandna, "Diagnosis of heart disease using data mining algorithm," Int. J. Comput. Sci. Inf. Technol., vol. 5, no. 2, pp. 1678–1680, 2014.

[35] H. D. Masethe and M. A. Masethe, "Prediction of heart disease using classification algorithms," in Proceedings of the world Congress on Engineering and computer Science, 2014, vol. 2, no. 1, pp. 25–29.

[36] B. Venkatalakshmi and M. V. Shivsankar, "Heart Disease Diagnosis Using Predictive Data mining," Int. J. Innov. Res. Sci. Eng. Technol., vol. 3, no. 3, pp. 1–5, 2014.

[37] B. Bahrami and M. H. Shirvani, "Prediction and diagnosis of heart disease by data mining techniques," J. Multidiscip. Eng. Sci. Technol., vol. 2, no. 2, pp. 164–168, 2015.

[38] U. Shafique, F. Majeed, H. Qaiser, and I. U. Mustafa, “Data mining in healthcare for heart diseases,” Int. J. Innov. Appl. Stud., vol. 10, no. 4, p. 1312, 2015.

[39] M. Abdar, S. R. N. Kalhori, T. Sutikno, I. M. I. Subroto, and G. Arji, “Comparing Performance of Data Mining Algorithms in Prediction Heart Diseases.,” Int. J. Electr. Comput. Eng., vol. 5, no. 6, 2015. http://doi.org/10.11591/ijece.v5i6.pp1569-1576.

[40] L. Verma, S. Srivastava, and P. C. Negi, “A hybrid data mining model to predict coronary artery disease cases using non-invasive clinical data,” J. Med. Syst., vol. 40, no. 7, pp. 1–7, 2016. https://doi.org/10.1007/s10916-016-0536-z.

[41] J. Thomas and R. T. Princy, “Human heart disease prediction system using data mining techniques,” in 2016 international conference on circuit, power and computing technologies (ICCPCT), 2016, pp. 1–5. https://doi.org/10.1109/ICCPCT.2016.7530265.

[42] G. Manogaran, R. Varatharajan, and M. K. Priyan, “Hybrid recommendation system for heart disease diagnosis based on multiple kernel learning with adaptive neuro-fuzzy inference system,” Multimed. Tools Appl., vol. 77, no. 4, pp. 4379–4399, 2018.

[43] S. Poornima, S. Sanjay, and S. P.-J. Gayatric, “Effective heart disease prediction system using data mining techniques,” Int. J. Nanomedicine, pp. 121–124, 2018.

[44] C. B. Gokulnath and S. P. Shantharajah, “An optimized feature selection based on genetic approach and support vector machine for heart disease,” Cluster Comput., vol. 22, no. 6, pp. 14777–14787, 2019.

[45] S. Maji and S. Arora, “Decision tree algorithms for prediction of heart disease,” in Information and communication technology for competitive strategies, Springer, 2019, pp. 447–454.

[46] D. Chicco and G. Jurman, “Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone,” vol. 5, pp. 1–16, 2020.

[47] M. G. El-Shafiey, A. Hagag, E.-S. A. El-Dahshan, and M. A. Ismail, “A hybrid GA and PSO optimized approach for heart-disease prediction based on random forest,” Multimed. Tools Appl., vol. 81, no. 13, pp. 18155–18179, 2022.

[48] N. Absar et al., “The efficacy of machine-learning-supported smart system for heart disease prediction,” in healthcare, 2022, vol. 10, no. 6, p. 1137.

[49] Narin, A.; Isler, Y.; Ozer, M. Early prediction of Paroxysmal Atrial Fibrillation using frequency domain measures of heart rate variability. In Proceedings of the 2016 Medical Technologies National Congress (TIPTEKNO), Antalya, Turkey, 27–29 October 2016.

[50] Shah, D.; Patel, S.; Bharti, S.K. Heart Disease Prediction using Machine Learning Techniques. SN Comput. Sci. 2020, 1, 345

[51] Drozd˙ z, K.; Nabrdalik, K.; Kwiendacz, H.; Hendel, M.; Olejarz, A.; Tomasik, A.; Bartman, W.; Nalepa, J.; Gumprecht, J.; Lip, G.Y.H. ˙ Risk factors for cardiovascular disease in patients with metabolic-associated fatty liver disease: A machine learning approach. Cardiovasc. Diabetol. 2022, 21, 240.

[52] Alotaibi, F.S. Implementation of Machine Learning Model to Predict Heart Failure Disease. Int. J. Adv. Comput. Sci. Appl. 2019, 10, 261–268.

[53] Hasan, N.; Bao, Y. Comparing different feature selection algorithms for cardiovascular disease prediction. Health Technol. 2020, 11, 49–62.

[54] S. Mohan, C. Thirumalai, and G. Srivastava, "Effective heart disease prediction using hybrid machine learning techniques," IEEE access, vol. 7, pp. 81 542–81 554, 2019.

[55] V. Ramalingam, A. Dandapath, and M. K. Raja, "Heart disease prediction using machine learning techniques: a survey," International Journal of Engineering Technology, vol. 7, no. 2.8, pp. 684–687, 2018.

[56] A. Rajdhan, A. Agarwal, M. Sai, D. Ravi, and P. Ghuli, "Heart disease prediction using machine learning," International Journal of Research and Technology, vol. 9, no. 04, pp. 659–662, 2020.

[57] Y. Khourdifi and M. Bahaj, "Heart disease prediction and classification using machine learning algorithms optimized by particle swarm optimization and ant colony optimization," International Journal of Intelligent Engineering and Systems, vol. 12, no. 1, pp. 242–252, 2019.

[58] A. Jagtap, P. Malewadkar, O. Baswat, and H. Rambade, "Heart disease prediction using machine learning," International Journal of Research in Engineering, Science and Management, vol. 2, no. 2, pp. 352–355, 2019.

[59] A. U. Haq, J. P. Li, M. H. Memon, S. Nazir, and R. Sun, "A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms," Mobile Information Systems, vol. 2018, 2018.

[60] R. Katarya and S. K. Meena, "Machine learning techniques for heart disease prediction: a comparative study and analysis," Health and Technology, vol. 11, no. 1, pp. 87–97, 2021.

[61] Pavithra, V.; Jayalakshmi, V. Hybrid Feature Selection Technique for Prediction of Cardiovascular Diseases. Mater. Today Proc. 2021; in press.

[62] Reddy, K.V.V.; Elamvazuthi, I.; Aziz, A.A.; Paramasivam, S.; Chua, H.N.; Pranavanand, S. Heart Disease Risk Prediction Using Machine Learning Classifiers with Attribute Evaluators. Appl. Sci. 2021, 11, 8352.

[63] Ananey-Obiri, D.; Sarku, E. Predicting the Presence of Heart Diseases Using Comparative Data Mining and Machine Learning Algorithms. Int. J. Comput. Appl. 2020, 176, 17–21.
[64] Rustam, F.; Ishaq, A.; Munir, K.; Almutairi, M.; Aslam, N.; Ashraf, I. Incorporating CNN Features for Optimizing Performance of Ensemble Classifier for Cardiovascular Disease Prediction. Diagnostics 2022, 12, 1474.
[65] Ahamad, G.N.; Fatima, H.; Zakariya, S.M.; Abbas, M. Influence of Optimal Hyperparameters on the Performance of Machine Learning Algorithms for Predicting Heart Disease. Processes 2023, 11, 734
[66] N. Chandrasekhar and S. Peddakrishna; "Enhancing Heart Disease Prediction Accuracy through Machine Learning Techniques and Optimization", Processes 2023, 11, 1210.
[67] M. A. Hambali, M. D. Gbolagade and Y. A. Olasupo, "Heart Disease Prediction Using Principal Component Analysis and Decision Tree Algorithm", Journal of Computer Science an Engineering (JCSE), Vol. 4, No. 1, February 2023, pp. 1-14.
[68] P. Gupta and D. Seth, "Comparative analysis and feature importance of machine learning and deep learning for heart disease prediction", Indonesian Journal of Electrical Engineering and Computer Science, Vol. 29, No. 1, January 2023, pp. 451~459.