

Heart Disease Classification Based on Hybrid Ensemble Stacking Technique

Ahmed Elsheikh

Department of Computer Science
Faculty of Computers and Information
Menofia University

{ ahmed-m.elsheikh@ci.menofia.edu.eg }

Nader Mahmoud

Department of Computer Science
Faculty of Computers and Information
Menofia University

{ nader.mahmoud@ci.menofia.edu.eg }

Arabi Keshk

Department of Computer Science
Faculty of Computers and Information
Menofia University

{ arabi.keshk@ci.menofia.edu.eg }

Abstract—Heart diseases are considered one of the leading death rates for humanity in the recent decades. The early diagnosis and prediction of heart disease becomes a critical subject in medical domain. Data mining techniques are usually used for finding anomalies, patterns and correlations within large data sets, thus it's crucial for clinical data analysis and various disease prediction. Ensemble approaches have proven to be quite effective in solving a variety of classification problems. In this research, we propose a hybrid ensemble stacking model with different feature engineering algorithms. The proposed ensemble model is based on five base models: Random Forest, Decision Tree, K-Nearest Neighbour (K-NN), Support Vector Machine (SVM), and Naïve Bayes for heart disease diagnosis. Logistic Regression meta model is used to merge base models predictions. We have examined various feature selection approaches such as: Brute Force, Principal Component Analysis (PCA), Classification and Regression Tree (CART) Feature Importance, and Logistic Regression based Recursive Feature Elimination. The proposed approach has been experimentally validated and evaluated on different dataset : UCI Cleveland and UCI Statlog. A quantitative evaluation shows that the combination of the ensemble model with brute force as feature selection technique yields a top accuracy of 97.8% for heart disease classification. the proposed stacking model has proven it's efficiency and overcomes existing approaches in heart diseases classification

Keywords—Heart Disease; Data Mining; Classification; Ensemble Learning; Stacking; Feature Selection.

1. INTRODUCTION

Heart disease, also known as Cardiovascular Disease (CD) is considered as the leading cause of death worldwide for decades. In 2016, and as reported by the world health organization, 17.9 million people passed away worldwide as a result of CD [1]. The death rate of heart disease can be reduced in case of early prediction and warning for the presence of the disease. Hence, it is highly required to develop a system to improve the prediction and classification of CD. In recent decades, data mining approaches have been researched to improve prediction process in medical field, where datasets become available [2].

Data mining is the process of finding hidden patterns, knowledge, and anomalies within large data sets. It is considered as the core part of Knowledge Discovery in Database (KDD)

process, which involves series of steps such as Data Preparation, Data Selection, Data Transformation and Data Mining which include several tasks such as prediction, classification, and clustering. The data mining techniques have been extensively studied in medical domain to improve various disease diagnosis such as: stroke, cancer, and diabetes detection. Recently, several authors study different data mining techniques for CD diagnosis and prediction. Diagnosis efficiency is the key factor in medical field, the development of most accurate diagnosis applications for CD has been revealed in literature [3-7]. Hybrid techniques have been researched to improve medical diagnosis efficiency, where data mining techniques have been integrated to produce hybrid models that can outperform performance of single models [9,11,25,29].

Moreover, researchers study the integration of various classification approaches such as: Support Vector Machine (SVM), K-Nearest Neighbour (K-NN), Logistic Regression, Decision Tree, Naïve Bayes and Vote (hybrid algorithm of Naïve Bayes with Logistic Regression), with different combination of dataset attributes for CD detection [7]. The Vote hybrid algorithm yielded the highest accuracy of 87.4% using 9 significant features of UCI Cleveland dataset [8]. A hybrid model based on Particle Swarm Optimization (PSO) has been proposed for rule discovering and CD diagnosis [9]. The model yielded an accuracy of 85% using 13 significant attributes on Z-Alizadeh Sani dataset [10]. The study in [11] introduced a hybrid model that combine Artificial Neural Network (ANN) and Genetic Algorithm (GA) for the diagnosis of CD across Z-Alizadeh Sani dataset. GA played an important role for tuning ANN's weights. Their proposed model (ANN + GA) has improved ANN's efficiency by around 10% and achieve an accuracy of 93.85%. In [12], a Classification and Regression Tree (CART) model for CD diagnosis has been presented. The model reduces the features subset to only five features and achieved an accuracy of 92.41% on Z-Alizadeh sani dataset. In [13] a classification model that combine KNN with GA has been introduced, where the model achieves an accuracy of 87.1% with (K = 5) and 5-fold cross validation.

Several research work have been presented in the literature that used a single classification model for CD detection like Neural Network (NN) [7,11,14,30], Naive Bayes [7,14], SVM [7], CART Decision Tree [7,12], K-NN [14]. However, since every classification model has its own point of strengths and

Study	Dataset	Technique	Accuracy	Number of Features
[7]	Cleveland	Vote technique	87.4%	9
[9]	Z-Alizadeh Sani	Hybrid PSO based technique	85%	13
[11]	Z-Alizadeh Sani	Hybrid ANN-GA technique	93.85%	22
[12]	Z-Alizadeh Sani	CART technique	92.41%	5
[13]	Statlog	Hybrid KNN-GA technique	87.1%	All features
[17]	Z-Alizadeh Sani	Weighted-average Voting technique	98.97%	5
[18]	Z-Alizadeh Sani	En-MFS technique	93.7%	34

Table1 Summarization of existing work for CD classification.

weaknesses, thus it cannot guarantee a high degree of accuracy in all cases. On the other hand, ensemble learning methods are considered as multiple classifier systems that combine various classifiers together in order to improve the performance of individual classifiers [15-16]. In comparison with single classification models, ensemble methods have potential to improve disease diagnosis accuracy. A prior literature studies [17,18] have used the ensemble classification techniques for CD diagnosis and detection. In [17], a weighted-average voting ensemble technique based on Random Forest, SVM, and KNN has been used for CD prediction, the technique yielded a classification accuracy of 98.97% using 5 features selected from Z-Alizadeh sani dataset. The study in [18] presents an ensemble based multiple feature selection (En-MFS) technique that yielded an accuracy of 93.7% in CD prediction using Z-Alizadeh sani dataset. Table 1 provides summarization of CD related work.

In this paper we propose an efficient ensemble stacking model to improve CD prediction. The proposed model is based on set of heterogenous classifiers including Random Forest, Decision Tree, K-NN, SVM, and Naïve Bayes as base models. Further, a meta model based on Logistic Regression is used to combine base models prediction and produce the final decision. In experimental evaluation, clinical datasets are used, however such datasets suffer from redundant/inconsistent features that may have a negative impact on the model efficiency [19]. Additionally, feature selection is an important phase In data mining approaches, which aims to reduce the dimensionality of the input features by selecting most significant feature subset which contributes most to output. In this paper, various feature selection methods such as: Brute Force, PCA, CART Feature Importance, and Logistic Regression based Recursive Feature Elimination are employed to identify the important attributes from CD dataset. The UCI Cleveland dataset has been selected for building and training the proposed ensemble model. The proposed model has been evaluated on a different dataset, UCI Statlog dataset where set of experiments carried out to assess the efficiency and performance of our model using 10-fold cross validation, and final model evaluation is reported.

The major objectives for this research are as follows :

- (1) Identifying most relevant features from CD dataset using various feature selection approaches such as: Brute Force, PCA, CART Feature Importance, and Logistic Regression based Recursive Feature Elimination.
- (2) An efficient ensemble stacking based on set of heterogeneous classifiers, i.e., Random Forest, Decision Tree, K-NN, SVM, and Naïve Bayes.
- (3) A meta model based Logistic Regression is used to combine all of these heterogeneous models' prediction and produce the final decision.

The rest of the paper is organized as follows. Section 2 discusses the materials (e.g., datasets) and methods utilized in our study. Experimental results and analysis of feature selection and the proposed method are presented in Section 3. Finally, Section 4 presents the conclusions for this research.

2. MATERIALS AND METHODOLOGY

This section firstly provides description about the materials (e.g., dataset) utilized in our experiments. Then, we shortly discuss the feature engineering techniques considered in our research. Moreover, we describe the proposed hybrid stacking model used for CD diagnosis.

2.1. Heart Disease Dataset

2.1.1. Dataset Description

The CD datasets utilized in this research were selected from UCI repository [8]. The Cleveland dataset has been used for building the proposed approach, which contains 303 nearly completed records. It is highly used dataset, thus helps to provide a comparative detection accuracy results with other research work. The Statlog dataset was selected to evaluate the performance of proposed model and extracted features. Both

Cleveland/Statlog datasets have the same structure and symptoms attributes, which are 13 different attributes related to heart disease and single predicated class attribute for disease classification. Table 2 provide detailed description for both dataset features. The main difference between two datasets is the values used to represent the predicated class attribute named "NUM". The Statlog dataset represents the predicated class with a binary class values 1 (absence of disease) and 2 (presence of disease). On the other hand, Cleveland dataset uses multiclass values that range from 0 to 4, where 0 means absence of disease and 1-4 represent disease presence with different intensity (where 4 being highest disease intensity).

2.1.2. Dataset Preprocessing

Firstly, we need to handle missing data, which represents records with incomplete observations that may cause biased results and affects the final model efficiency. Cleveland dataset contains 6 records with missing values. All these records were identified and deleted manually from dataset. Therefore, the records number has been reduced to 297 from 303. Statlog dataset consists of 270 records which are all complete with no missing values. Secondly, we perform **Output Attribute Transformation**. Thus, the values of predicated class attribute called "NUM" in Cleveland dataset was converted from multiclass with five levels from 0 to 4 [0 represents absence of CD and 1-4 represents the presence of CD with different stages] to binary class with just two levels 0 and 1 [0 represents CD absence and 1 represents CD presence]. This unifies both Cleveland/Statlog values for the predicated class attribute "NUM". Dataset categorical attributes were preprocessed using **One Hot Encoding**, which is important to avoid model confusion and invalid results. Consider an attribute "THAL", that may contain three different categories values (3, 6, and 7). Thus, it may lead to model confusion, whereas the category with higher value (7) will dominate other categories with lower values (3 and 6). Therefore, one hot encoding solves this issue by splitting categorical attribute into different columns depending on number of categories. Thus, "THAL" attribute will be splitted into three different columns. Each column represents one category with values 0 and 1. Finally, we perform **Normalization** which considered a crucial step when dealing with numeric attributes that have different scale. Without normalization higher scale attributes will dominate lower scale ones and this will result in biased conclusions. In our experiments we first standardize each numeric attribute based on mean and standard deviation. Next, normalization task applied for all numeric attributes to rescale each one based on [0 – 1] scale.

2.2. Feature Engineering Techniques

Feature engineering techniques are desirable to reduce the dimensionality of the input features, improve model performance and save model computational time [19]. Traditionally, selecting the most relevant features affects on the power and predictability of final model. Several feature selection methods such as: SVM features weight [20], Information Gain [21], Correlation [22], PCA [22], Recursive Feature Elimination [23], and Boruta wrapper-based feature selection [24] are used for selecting the important features of

CD. In this paper we have examined different feature reduction techniques on UCI Cleveland dataset such as: Brute Force, PCA, CART Feature Importance, and Logistic Regression based Recursive Feature Elimination. All of these methods prove to be efficient in medical field.

Table 1 Description of UCI Cleveland/Statlog dataset attributes

Feature	Description	Values
Age	Patient Age (years)	Numeric
Sex	Gender	Nominal [Male=1, Female=0]
Cp	Chest Pain	Nominal [typical angina=1, atypical angina=2, non-anginal pain=3, asymptomatic=4]
Trestbps	Rest Mode Blood Pressure	Numeric
Chol	Levels of Cholesterol	Numeric
Fbs	Fasting blood sugar > 120 mg/dl	Nominal [true=1, false=0]
Restecg	Electrocardiographic Results	Nominal [normal=0, abnormality of ST-T wave=1, probable or left ventricular hypertrophy=2]
Thalach	Highest Heart Rate	Numeric
Exang	Exercise Induced Angina	Nominal [yes=1, no=0]
Oldpeak	ST Depression Segments Induced at exercise Compared to rest	Numeric
Slope	ST Segments Peak Slop	Nominal [upsloping=1, flat=2, downsloping=3]
Ca	Number of Flourosopy Major vessels	Numeric
Thal	Defect of Heart	Nominal [normal=3, fixed defect=6, reversable defect=7]
Num	Diagnosis	Nominal

2.2.1. Brute Force

The UCI Cleveland dataset has 13 features that represent different symptoms used for CD classification. Brute Force feature selection technique starts by generating every possible combination with 4 features from the 13 features and examined each feature combination with our ensemble stacking model. In the next step, the process is repeated in order to generate longer feature combination with 5 features instead of 4 features. The Brute Force process is repeated till generating all possible combinations for 13 attributes of UCI Cleveland dataset. Accordingly, total combinations generated from 13 features, ignoring empty set, is 8191. In our experiments, the lower limit for a combination features is set to 4 features. All combinations with one, two, and three features are ignored, therefore a total of 7814 different combinations are generated and examined with the proposed ensemble stacking model. The feature subset ['age', 'sex', 'chol', 'fbs', 'slope', 'thal'] achieves the highest performance, which are considered for evaluation.

2.2.2. Principal Component Analysis (PCA)

PCA is another dimensionality reduction algorithm that is used when dealing with correlated features. It transforms these features to a set of linearly and uncorrelated variables named principal components. PCA algorithm starts by standardize the input features with reference to mean and standard deviation. Then the covariance matrix is computed to identify the correlation between every two features with three possible values as output. 1) Positive values stands for highly correlation, thus two features decrease and increase similarly. 2) Negative values stands for two features are contradictory. 3) Zero is for uncorrelated features. Finally, the Eigen values and Eigen vectors of covariance matrix are calculated, the higher Eigen value the more significance corresponding Eigen vector. In our experiment, we examined the highest 9 Eigen vectors for evaluation with proposed ensemble stacking model.

2.2.3. Classification and Regression Tree (CART) Feature Importance

CART Feature Importance are also examined as a feature selection technique that rank dataset features based on Gini Impurity (GI) metric. CART classification tree utilizes different measures like GI in splitting criterion to determine the optimal feature used for node splitting. GI shows the probability of an observation being misclassified to random class label based on the dataset's class distribution. The formula of GI computed as:

$$GI = 1 - \sum_{i=1}^n P_i^2$$

n : number of output classes.

P : probability of correctly classify an observation to class i .

Figure 1 shows features ranking using CART feature importance technique. In our experiment, we examined the highest 7 features ['age', 'cp', 'chol', 'thalach', 'oldpeak', 'ca', 'thal'] among all dataset features for evaluation.

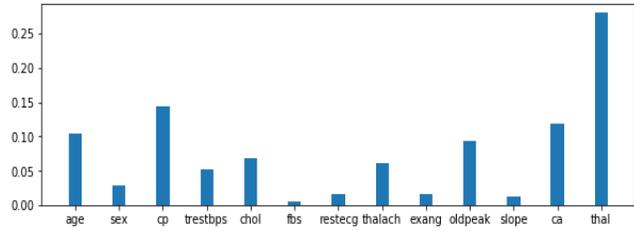


Figure 1 Feature selection using CART feature importance

2.2.4. Recursive Feature Elimination

Recursive Feature Elimination is considered as a wrapper-based feature selection algorithm that uses a machine learning algorithm (i.e. in our case Logistic Regression) for determining the importance of features. Recursive Feature Elimination starts by using all dataset features to train and fit Logistic Regression model and feature importance is determined for each one. Recursive Feature Elimination model is repeated recursively and each round the features with least significance are ignored and important features are kept for next round till the best features are selected. In our experiments, top 7 significant features were ['sex', 'cp', 'fbs', 'exang', 'oldpeak', 'slope', 'ca'], which will be used to examine and evaluate the proposed ensemble stacking model.

2.3. Proposed Ensemble Model

The proposed approach combines effective feature selection methods with a hybrid stacking model in order to improve CD classification and detection. The key idea of stacking is to combine different predictions of various models using a meta model. This architecture improves classification accuracy over individual classifiers, due to the usage of multiple classification models that have different abilities in solving the classification issues. Figure [2] Provide an overview of the proposed hybrid stacking model. The architecture of our stacking model contains three main phases **Data Partitioning Stage** (Sec 2.2.1), **Base Models Stage** (Sec 2.2.2) and **Meta Model Stage** (Sec 2.2.3).

2.3.1. Data Partitioning Stage

It is an initial stage in which we partition and prepare the data for the following model stages. The proposed approach utilizes two different datasets at this stage. UCI Cleveland dataset is used to train the ensemble stacking model, and UCI Statlog dataset used as holdout partition to evaluate the proposed model.

2.3.2. Base Models Stage

Also known as (Level-0 Classifiers), in this stage we consider five different classification algorithms: Random Forest, Decision Tree, K-NN, SVM, and Naïve Bayes which have proven their effectiveness in dealing with CD disease prediction. The reason of using an odd number of models (5) at **Base Model Stage** is to avoid equal CD and Non-CD predictions for an observation. The features subsets identified at (Sec. 2.2) were selected from the UCI Cleveland dataset, which were used for training the base models using 10-fold cross validation and all

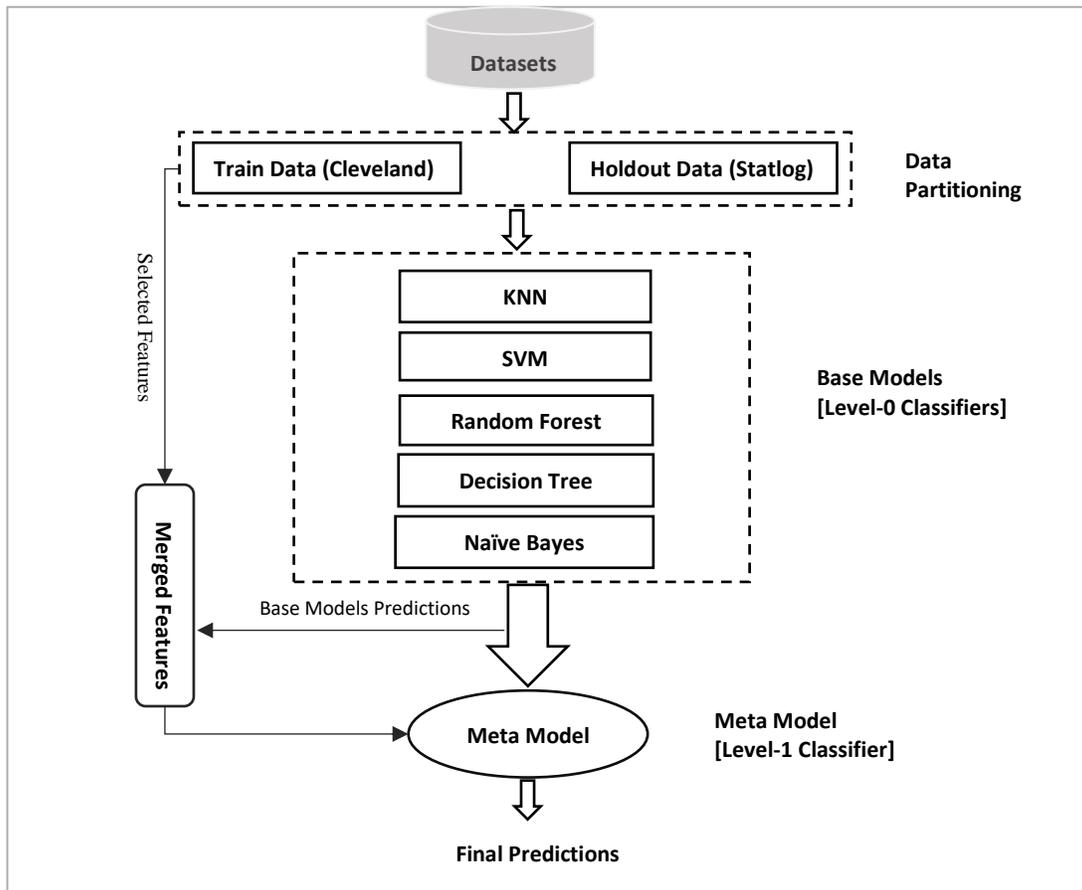


Figure 2 Overview of proposed stacking model

out-of-fold predictions are retained for each base model. All predictions of base models are joined with extracted features and used as the input to the *Meta Model Stage*, as illustrated in Figure 2.

2.3.3. Meta Model Stage

In this phase, we combine the predictions of the base models using a meta model classifier, for which we use Logistic Regression. The five prediction vectors of the base models are joined with the input features, which are used next to train the meta model using 10-fold cross validation. Figure 2 shows the training procedure for the proposed hybrid stacking model. The model is trained on UCI Cleveland dataset. The model is then evaluated on holdout data obtained in *Data Partitioning Stage*.

3. EXPERIMENTAL RESULTS

Several experiments have been conducted for model evaluation and performance validation. All Experiments in this research are carried out for 10 times for each model and we report the average of these 10 runs. Experiments are performed on a machine with an intel core i5 processors, 16GB RAM and

windows 10 OS. The Scikit Learn which is an efficient python library for data mining are used to implement all our experiments. Section 3.1 presents the performance evaluation metrics used for evaluating base models and the final stacking model. Section 3.2 provides a performance comparison between each individual base model and our stacking model for CD diagnosis and classification. In Section 3.3 we present the performance variation when applying different feature engineering methods with our ensemble stacking model. Section 3.4 provides a benchmark comparison between the proposed ensemble stacking model and other existing work in the field of CD diagnosis and prediction.

3.1. Performance Evaluation Metrics

This section provides an overview of the performance evaluation for the individual base models and the ensemble stacking model in CD diagnosing and prediction. The performance evaluation for any machine learning based model is evaluated with the help of confusion matrix parameters. The binary classification problem contains four possible outcomes in confusion matrix, namely True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN).

TP : CD patients who are correctly classified to be CD.

TN : Healthy persons who are correctly classified to be Non-CD.
 FP : Healthy persons who are incorrectly classified to be CD (Type I error).

FN : CD patients who are incorrectly classified to be Non-CD (Type II error).

The performance accuracy is the metric utilized for our experiments which defined as the ratio of correctly classified instances to the total number of instances. The mathematical representation for accuracy is as follows :

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

3.2. Individual Base Model Evaluation

All five base classifiers, i.e., Random Forest, Decision Tree, K-NN, SVM, and Naïve Bayes are evaluated individually on UCI Cleveland dataset and results are compared with the proposed hybrid stacking model. All dataset 13 features are applied for evaluation without any feature selection. Table[3] shows performance accuracy comparison between each individual base models and the proposed hybrid stacking model using 10-fold cross validation. It can be clearly seen that the hybrid stacking ensemble model outperforms other individual base models in CD classification with an accuracy of 86.4%. This clearly prove that ensemble techniques can guarantee a high degree of accuracy comparing to the individual classifiers. Moreover, results also indicate that Random Forest algorithm overcomes other individual classifiers with an accuracy of 85.6%, as it is considered a kind of ensemble techniques called Bagging, that combines a set of decision trees in parallel manner. Other individual classifiers namely, SVM, K-NN, Naïve Bayes and Decision Tree achieve an accuracy of 83.4%, 82.6%, 82.3% and 76.2% respectively in CD classification. The hyperparameters for individual models are tuned empirically as presented in Table [4].

3.3. Feature Engineering Evaluation

Firstly, we discuss the experiments of selecting CD dataset significant features using set of feature engineering methods, i.e., Brute Force, PCA, CART Feature Importance, and Logistic Regression based Recursive Feature Elimination. The training partition generated at *Data Partitioning Stage* (i.e., UCI Cleveland dataset) has been used firstly to identify the CD relevant features that will help at improving CD diagnosis. The Brute Force feature selection technique examined all possible combinations of UCI Cleveland dataset features with the proposed hybrid stacking model. The combinations lower limit is set to 4 features using trial-error procedure. The combination of ['age', 'sex', 'chol', 'fbs', 'slope', 'thal'] considered as the best performing subset, which will be used later for evaluation. The PCA feature extraction algorithm has been examined with the UCI Cleveland dataset. PCA transforms the dataset features into set of linearly and uncorrelated variables called principal

components. The highest 9 principal components have considered for evaluation. The CART Feature Importance technique also applied to UCI Cleveland dataset in order to select the dataset most relevant features. The top 7 important features, i.e., ['age','cp', 'chol', 'thalach','oldpeak', 'ca', 'thal'] only used for evaluation. The Recursive Feature Elimination which considered a wrapper-based feature selection technique that uses in our case Logistic Regression for selecting UCI Cleveland dataset significant features. The top 7 significant features identified by the Recursive Feature Elimination were ['sex', 'cp','fbs','exang', 'oldpeak', 'slope', 'ca'].

Table 3 Performance comparisons between base models and our stacking model

Algorithm	Accuracy
Random Forest	85.6%
KNN	82.6%
Decision Tree	76.2%
SVM	83.4%
Naïve Bayes	82.3%
Stacking Model	86.4%

Table 4 Hyperparameters used for individual classifiers

Algorithm	Hyperparameters
Random Forest	n_estimators = 100, criterion = gini, max_depth = 10, max_features = auto, min_samples_split = 6, bootstrap = true
KNN	n_neighbors = 10, algorithm = auto, weights = distance
Decision Tree	criterion = gini, splitter = best, max_depth = 10, max_features = auto, min_samples_split = 3, bootstrap = true
SVM	Kernel = RBF, degree = 3, gamma = 0

Table 5 Performance accuracy for different Feature Engineering methods with hybrid stacking model.

Algorithm	Accuracy	Features Selected
Stacking + Brute Force	97.8%	['age', 'sex', 'chol', 'fbs', 'slope', 'thal']
Stacking + PCA	91.8%	9 Principal Components
Stacking + CART Feature Importance	94%	['age','cp', 'chol', 'thalach','oldpeak', 'ca', 'thal']
Stacking + Recursive Feature Elimination	92.5%	['sex', 'cp','fbs','exang', 'oldpeak', 'slope', 'ca']

Table 6 Benchmark comparison of proposed model with previous studies.

Study	Technique	Number of Features	Accuracy
Proposed	Brute Force + Hybrid Stacking Ensemble Model	6	97.8%
[7]	BF + Vote (Naïve Bayes + Logistic Regression)	9	87.41%
[25]	PSO + CFS + MLP	7	90.28%
[26]	SVM Stacking	9	92.22%
[27]	IG + MLP	8	80.99%
[28]	Hybrid PSO + Emotional NN	7	84%

Later, the significant features identified by the previously mentioned experiments using UCI Cleveland dataset were selected from the holdout partition generated at *Data Partitioning Stage* (i.e., UCI Statlog dataset). The selected subsets were examined with the proposed hybrid stacking model using 10-fold cross validation. First, the feature subset was selected from the holdout Statlog dataset and passed to Base Models Stage, predictions are made for each base model classifier, these predictions along with the input features provide the data used to evaluate our meta model and the proposed stacking model. Table [5] summarizes the performance accuracy for each feature engineering method with the proposed hybrid stacking model. The hybrid stacking model with Brute Force method as feature selection achieves the highest performance accuracy with 97.8% using six features. The Brute Force feature selection outperforms other feature engineering methods as it iterates over all possible combinations of the dataset attributes in order to select the best performing subset, which in our case ['age', 'sex', 'chol', 'fbs', 'slope', 'thal'].

3.4. Benchmark with Previous Studies

The performance of the proposed hybrid stacking model with brute force feature selection has been benchmarked with the state-of-art approaches in the field of CD diagnosis and prediction using UCI datasets. As noticed in Table [6], the performance of the proposed model overcomes other studies with only six features, i.e., (age, sex, chol, fbs, slope, thal), it achieves an accuracy of 97.8% for CD diagnosis. The comparison proves the efficiency and strength of ensemble techniques over single classifiers. However, there are some disadvantages with the proposed model which should be considered in the future work. First, some base classifiers clearly outperform other ones, therefore the use of weights will be necessary in such cases. To achieve this, an evolutionary technique must be applied to select the proper weights for all base classifiers. Moreover, the brute force feature selection

technique will be too difficult to be considered with datasets with higher dimensions.

4. CONCLUSION

In this research, a hybrid ensemble stacking model is introduced to improve the detection of CD. The proposed model is based on set of heterogenous classifiers including Random Forest, Decision Tree, K-NN, SVM, and Naïve Bayes as base models. Further, a meta model based on Logistic Regression is used to combine base models prediction and produce the final decision. Set of feature selection techniques was involved to select most significant features that have high impact on the model efficiency. The UCI Cleveland dataset was selected for model building and significant features identification. The UCI Statlog dataset was selected to evaluate the performance of proposed model and extracted features. Experimental evaluation shows that the combination of the ensemble model with brute force as feature selection technique yields a top accuracy of 97.8% which outperforms existing approaches in CD classification (see Table 6).

REFERENCES

- [1] (WHO), W. H. (2017). Cardiovascular diseases (CDs) – Key Facts. [https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-\(cvds\)](https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)).
- [2] Srinivas, K., Rani, B.K., Govrdhan, A., 2010a. Applications of data mining techniques in healthcare and prediction of heart attacks. *Int. J. Comput. Sci. Eng. (IJCSE)* 2(02), 250–25562.
- [3] Alexander CA, Wang L. Big data analytics in heart attack prediction. *J Nurs Care.* 2017;6(2). <https://doi.org/10.4172/2167-1168.1000393>.
- [4] Maini E., Venkateswarlu B., Gupta A. (2019) Applying Machine Learning Algorithms to Develop a Universal Cardiovascular Disease Prediction System. In: Hemanth J., Fernando X., Lafata P., Baig Z. (eds) *International Conference on Intelligent Data Communication Technologies and Internet of Things (ICICI) 2018*. ICICI 2018. Lecture Notes on Data Engineering and Communications Technologies, vol 26.
- [5] Tspirouras MG, et al. Automated diagnosis of coronary artery disease based on data mining and fuzzy modeling. *IEEE Transactions on Information Technology in Biomedicine* 2008;12(4):447–58.
- [6] Ekta Maini, Bondu Venkateswarlu, Baljeet Maini, Dheeraj Marwaha. Machine learning-based heart disease prediction system for Indian population: An exploratory study done in South India, 2021, ISSN 0377-1237, <https://doi.org/10.1016/j.mjafi.2020.10.013>.
- [7] Shafenor Amin M, Kia Chiam Y, Dewi Varathan K. Identification of significant features and data mining techniques in predicting heart disease. *Telematics Inf.* 2018.
- [8] R. Detrano, V.A.M. Center, Long beach and cleveland clinic foundation and university of california, irvine machine learning repository, 2019, Cleveland Heart Disease Dataset Available at <http://archive.ics.uci.edu/ml/datasets/Heart+Disease> Date last accessed: 20-12.
- [9] Zomorodi-moghadam, M., et al., Hybrid Real-Binary Particle Swarm Optimization for Rule Discovery in the Diagnosis of Coronary Artery Disease. *Expert Syst.* (2019), doi: 10.1111/exsy.12485.
- [10] Z. A. Sani, R. Alizadehsani, M. Roshanzamir, Z-Alizadeh Sani data set, 2020, [Online]. Available: <https://archive.ics.uci.edu/ml/datasets/Z-Alizadeh+Sani>, Accessed January.
- [11] Arabasadi Z, et al. Computer aided decision making for heart disease detection using hybrid neural network-Genetic algorithm. *Computer methods and programs in biomedicine* 2017; 141:19–26.
- [12] M.M. Ghiasi, et al. Decision tree-based diagnosis of coronary artery disease: CART model, *Computer Methods and Programs in Biomedicine* 192 (2020) . Article No: 105400
- [13] M. Akhil jabbara, B.L. Deekshatub and Priti Chandrac., 2013. Classification of Heart Disease Using K-Nearest Neighbor and Genetic Algorithm”, *Procedia Technology*, Vol.10
- [14] R. Alizadehsani , J. Habibia , M.J. Hosseini, H. Mashayekhi, R. Boghrati , A. Ghan- deharioun , B. Bahadorian , Z. Alizadeh, Sani, a data mining approach for diagnosis of coronary artery disease, *Comput. Methods Programs Biomed.* 111 (2013) 52–61.
- [15] Mohapatra, S., D. Patra, and S. Satpathy, An ensemble classifier system for early diagnosis of acute lymphoblastic leukemia in blood microscopic images. *Neural Computing and Applications*, 2014. 24(7-8): p. 1887-1904.
- [16] A.R., E. Naderi, and A. Osareh. Parallel weak learners, a novel ensemble method. in *Computational Intelligence and Computing Research (ICCIC)*, 2010 IEEE International Conference on. 2010. IEEE.
- [17] Durgadevi Velusamy, Karthikeyan Ramasamy, Ensemble of heterogeneous classifiers for diagnosis and prediction of coronary artery disease with reduced feature subset, *Computer Methods and Programs in Biomedicine*, Volume 198, 2021, 105770, ISSN 0169-2607, <https://doi.org/10.1016/j.cmpb.2020.105770>.
- [18] C. J. Qin , Q. Guan , X. P. Wang , Application of ensemble algorithm integrating multiple criteria feature selection in coronary heart disease detection, *Biomed. Eng. Appl. BasisCommun.* 29 (6) (2017) 1–11.
- [19] Shilaskar S, Ghatol A. Feature selection for medical diagnosis: Evaluation for cardiovascular diseases. *Expert Systems with Applications* 2013;40(10):4146–53.
- [20] A. Ben-Hur , J. Weston , A user’s guide to support vector machines, *Data Min. Tech. Life Sci.* 609 (2010) 223–239.
- [21] J. Han, M. Kamber , *Data Mining: Concepts and Techniques Morgan Kaufmann*, second ed., 2006.
- [22] I. Guyon, A. Elisseeff, An introduction to variable and feature selection, *J. Mach. Learn. Res.* 3 (2003) 1157–1182.
- [23] A. Ogunleye, Q.G. Wang, XGBoost model for chronic kidney disease diagnosis, *IEEE/ACM Trans. Comp. Biol. Bioinf.* (2019), doi: 10.1109/TCBB.2019.2911071.
- [24] M.B. Kursu, W.R. Rudnicki, Feature selection with the Boruta package, *J. Stat. Softw.* 36 (11) (2010) 1–13.
- [25] Verma L, Srivastava S, Negi P. A hybrid data mining model to predict coronary artery disease cases using non-invasive clinical data. *Journal of medical systems* 2016;40(7):178.
- [26] L. Ali , A. Niamat , J.A. Khan , N.A. Golilarz , X. Xingzhong , A. Noor , R. Nour , S.A.C. Bukhari , An optimized stacked support vector machines based expert system for the effective prediction of heart failure, *IEEE Access* 7 (2019) 54007–54014 .
- [27] A. Khemphila and V. Boonjing, "Heart Disease Classification Using Neural Network and Feature Selection," 2011 21st International Conference on Systems Engineering, Las Vegas, NV, USA, 2011, pp. 406-409, doi: 10.1109/ICSEng.2011.80.
- [28] Afzal Hussain Shahid, M.P. Singh, A Novel Approach for Coronary Artery Disease Diagnosis using Hybrid Particle Swarm Optimization based Emotional Neural Network, *Biocybernetics and Biomedical Engineering*, Volume 40, Issue 4, 2020, Pages 1568-1585, ISSN 0208-5216, <https://doi.org/10.1016/j.bbe.2020.09.005>.
- [29] Shahid AH, Singh M. Computational intelligence techniques for medical diagnosis and prognosis: Problems and current developments. *Biocybernetics and Biomedical Engineering* 2019.
- [30] L. Verma, S. Srivastava, P. Negi ,A hybrid data mining model to predict coronary artery disease cases using non-invasive clinical data, *Journal of Medical Sys.* 40 (178) (2016) 1–7 . (7)

