Hybrid Feature Selection Algorithm and Ensemble Stacking for Heart Disease Prediction

Nureen Afiqah Mohd Zaini[1], Mohd Khalid Awang[2]
Faculty of Informatics and Computing, Universiti Sultan Zainal Abidin,
22000 Tembila, Terengganu, Malaysia

***Abstract*—In cardiology, as in other medical specialties, early and accurate diagnosis of heart disease is crucial as it has been the leading cause of death over the past few decades. Early prediction of heart disease is now more crucial than ever. However, the state-of-the-art heart disease prediction strategy put more emphasis on classifier selection in enhancing the accuracy and performance of heart disease prediction, and seldom considers feature reduction techniques. Furthermore, there are several factors that lead to heart disease, and it is critical to identify the most significant characteristics in order to achieve the best prediction accuracy and increase prediction performance. Feature reduction reduces the dimensionality of the information, which may allow learning algorithms to work quicker and more efficiently, producing predictive models with the best rate of accuracy. In this study, we explored and suggested a hybrid of two distinct feature reduction techniques, chi-squared and analysis of variance (ANOVA). In addition, using the ensemble stacking method, classification is performed on selected features to classify the data. Using the optimal features based on hybrid features combination, the performance of a stacking ensemble based on logistic regression yields the best result with 93.44%. This can be summarized as the feature selection method can take into account as an effective method for the prediction of heart disease.**

Keywords—Heart disease prediction; feature selection; stacking; accuracy

I. Introduction

The process of learning a function that maps an input to an output based on examples of input-output pairs is referred to as supervised learning in the field of machine learning. This task involves learning a function that maps an input to an output. It accomplishes this by drawing conclusions about a function based on a collection of samples from training that have been labelled [1]. In a variety of fields, including marketing, commercial applications, pattern recognition, image processing, classification, and prediction, feature selection has been utilized. It is common to encounter a sizable data collection and a high number of features while working with actual applications. Most of the time, just a few of the features are important and pertinent to the objective. Since the remaining features are viewed as unimportant and unnecessary, doing without them would not only affect performance but also classification accuracy. As a result, choosing a suitable and compact feature subset from the original features is crucial to improving classification performance and accuracy as well as overcoming the curse of dimensionality. To determine the importance of attributes, feature selection techniques are employed, and the aim is to minimize the number of input variables to those demands most relevant to the model. Aside from minimizing the number of attributes, feature selection also reduces processing time as well.

According to [2], medical records from the National Heart Institute Malaysia (IJN) discovered between January 1, 2009, and December 31, 2018, were used in a non-interventional study that looked back 10 years. From the IJN database, there were 3923 out of 4739 eligible and used in the analysis. Another study by [3] in 2019, conducted by the Department of Statistics of Malaysia, found that heart disease was the leading cause of death in Malaysia. Representing 15% of all fatalities requiring rapid medical attention. However, heart disease can be prevented by avoiding dangerous factors. In machine learning, varieties of algorithms such as supervised, unsupervised, semi-supervised, reinforcement, and transduction, are frequently employed. Supervised learning is the ability of an algorithm to synthesize knowledge from previously labelled data in order to predict future unlabelled cases [4].

In this study, 13 attributes from the UCI dataset are used for the experiment to determine the cause of heart disease. Nevertheless, not all attributes are useful, and a feature selection method is needed to prove the only important cause of heart disease. The choice of attributes based on the feature selection method might vary depending on the feature selection method used. The prediction of heart disease can be detected based on symptom from patients which make the specialist's task easier. When we talk about predicting heart disease, we should note that prediction is one of the applications of machine learning that is utilized frequently. With the assistance of machine learning, data mining is quickly becoming an essential part of the healthcare industry by employing classification and prediction techniques which are used to generate models that describe necessary classes [5].

It is commonly held risk factors such as age, sex, chest pain type, trestbps, chol, fasting blood sugar, restecg, thalach, exang, oldpeak, slope, number of major vessels, and thalassemia are the major risk factors for heart disease according to the dataset used. In light of these considerations, this research employed a feature selection method to build a heart disease risk assessment model that could aid specialists in making accurate early predictions [6].

Even though several feature selection strategies have been used in decision support systems for medical datasets, there is always the opportunity for improvement. The combination of feature selection algorithms and classifiers has to be tuned for heart disease datasets with a lot of feature space in order to deliver high performance. The proposed framework is based on a well-balanced mix of two different types of feature selection algorithms that work well together.

This study aims to propose a hybrid feature selection that combines both chi-squared and ANOVA techniques. Chi-squared is utilized for the selection of categorical features, whereas ANOVA is applied to numerical features. The research proposes to combine the highest rank from both techniques, and the five most influential features are derived from a total of 13 features. The five most influential features are then evaluated using an ensemble stacking approach to improve the accuracy of heart disease prediction.

This paper is organized as follows: Section II discovered related works consisting of accuracy achieved by the author using feature selection technique for the prediction of heart disease. Section III discusses the dataset to use for the experiment along with the feature selection technique and framework that visualize the whole process for the experiment. Lastly, section IV discussed the result obtained based on the experiment made, and in Section V, the conclusion is presented.

II. Literature Review

There are several elements that lead to heart disease, however the present approaches for heart disease prediction are inadequate and need to be improved. By using reduction approaches to remove some of the redundant features, the prediction accuracy might be improved. Feature selection is a process of selecting important attributes of the dataset. Pre-processing is the main step for selecting important attributes for a certain dataset. In this research, ten base classification algorithms and three subsets of meta-models are tested for the prediction of accuracy.

A. Filter Method

The filter method is one of the feature selection methods that independently evaluates the importance of each feature. The selected features are subsequently used as input for a model-building process.

Before induction can take place, the filter method is used to remove unwanted attributes using one paradigm which independently act[7]. Karl Pearson pioneered the use of chi-squared statistics for categorical data, but it will take some time before the asymptotic distribution of these statistics was thoroughly understood [8].

However, the valid conclusion from chi-squared depends on several assumptions such as [9]:

1) A cross-tabulation can be used to figure out actual frequencies. The chi-squared test should not be used for percentages or other derived statistics.

2) The two variables are nominal which is the categories have no natural ordering.

3) Independent observations.

4) More than 75%-80% of contingency table columns have an expected count of ≥ 5, and none have an expected count of 0.

Aside from chi-squared, [10] ANOVA test is another filter-based feature selection technique used in this research. By utilizing the SelectKBest class, the f_classif() function is called upon to determine the most important features. SelectKBest class may be found in the scikit-learn library which employs a scoring function to assign the features with the highest score.

According to [11], Classification and Regression Trees (CART), Gradient Boosting Machine (GBM), Adaboost, K-Nearest Neighbor (KNN), Multilayer Perceptron (MLP), Stochastic Gradient Descent (SGD), Support Vector Classifier (SVC) and Naïve Bayes are tested through feature selection to find the best accuracy algorithm. CART was found to have the best accuracy with 87.65%. Four important attributes from eleven features are selected based on the feature selection. The author uses the majority voting technique to find out the best attributes and the result proved that st_slope_flat and st_depression are the best and second highest results go to max_heart_rate_achieved, exercise_induced_angina, and cholesterol. The authors claimed these attributes are the leading cause of heart disease.

In 2019 [12], the author uses a rapid miner as a tool to test the accuracy of each algorithm. Six algorithms such as Decision Tree, Logistic Regression, Logistic Regression SVM, Naïve Bayes, and Random Forest are tested and the result found out Logistic Regression SVM is the highest with 84.85%. However, the author did not reveal the attributes of the leading cause.

On the other hand, [13] investigated the use of principal component analysis (PCA) in clinical aspect. To measure the effectiveness of reducing infection risk among university student, a pilot study of 200 volunteers was carried out. Essential clinical parameters were identified and confirmed by medical experts. From the clinical history variables with 49 parameters, the disease was identified through the use of PCA. PCA method was utilized to confirm the weightage of risk level towards the disease in order to ensure the system possesses the highest possible level of accuracy, reliability, and efficacy. Cumulative achieved with the use of PCA is 58.288% and the author proof optimal accuracy, reliability, and efficiency to conduct mass-screening of students.

The author in [14], found the most accurate algorithm achieved 85.00% using chi-squared feature selection with the BayesNet classifier. The dataset of heart disease is tested using principal component analysis (PCA), chi-squared testing, ReliefF, and symmetrical uncertainty. The author agreed to use PCA feature extraction with IBK and the result is highest for recall at 87.22% but the accuracy is low compared to the chi-squared result. Based on the results, cp is categorized as the most influential feature for heart disease prediction followed by exang, chol, and thal. Different features are ranked differently based on which feature selection is used.

Based on the dataset, this research [15] compares several machine learning techniques and determines the most efficient classification technique. KNN, NB, decision tree (J48), and RF are four different classification algorithms and other techniques, such as SVM were used to compare with affinity degree (AD) classification. All these algorithms are then tested on three different UCI dataset. As a result, J48 demonstrates the highest level of performance when compared to the other four classifiers as the purpose of this research is to investigate the compatibility of affinity to use for classification method.

The study by [16] affirms the use of the backward feature selection technique resulted in the highest accuracy of 88.52% using the decision tree algorithm. Algorithms such as random forest, support vector machine, decision tree, k-nearest neighbor, logistic regression, and gaussian naïve bayes are tested and the decision tree outperformed the other five algorithms. They also experimented with the accuracy using ten different feature selection techniques which are ANOVA, chi-squared, mutual information, ReliefF, forward feature selection, backward feature selection, exhaustive feature selection, recursive feature elimination, lasso regression, and ridge regression. As a result, backward feature selection is the most influential feature selection technique which leads to a better result.

Research done by [17], suggested dataset of 70000 patients and 11 features are tested with the chi-squared feature selection method. Features involved in this research consist of age, gender, height, weight, systolic blood pressure, diastolic blood pressure, cholesterol, glucose, smoking, alcohol intake, and physical activity. Seven algorithms and the chi-squared method were used to filter the most influential features. The author adjusted some features of the dataset to discover the factors that have the greatest impact on cardiovascular disease which resulted in weight and height as the most influential cardiovascular disease. As a result, Multi-Layer Perceptron achieved the highest accuracy with 87.23%.

The authors of the research [18], proposed two different datasets and use a feature selection technique to find the best features. The author also tested the ensemble classifier with a sampling technique to find the best accuracy. ANOVA is one of the feature selection techniques used by the author to find the best features for improved accuracy. The study [19] suggested a model predict numerous diseases as there are very few suggestions made about the detection of numerous diseases. The author takes into consideration conditions such as heart disease, diabetes, and kidney disease. There are only a few features in the dataset that will not affect how well the prediction system works and only important features will be taken into consideration for the decision-making. Chi-squared and ANOVA are applied to trace out the best features from the dataset. Exang, cp, ca, oldpeak and thalach are chosen as the most influential features.

A study conducted by [20], shows the size of the dataset increase as the complexity of the model increases. Classification and regression fields are tested respectively in this research for comparison purposes as they might be a potential resource for the researcher to decide on appropriate algorithms. Chi-squared as one of the feature selection methods is used for categorical, ordered with missing values, and ordered without missing values. The major benefit of chi-squared is, it decreases computing complexity through the merging procedure by decreasing the number of categories for each predictor.

Recently, [21] developed a heart failure survival prediction model with the help of an ensemble tree machine learning approach. Extreme Gradient Boosting (XGBoost) was demonstrated as the most accurate classifier with 83.00%. During the pre-processing stage, the unimportant feature will be removed to obtain better accuracy. The author uses ANOVA and chi-squared to analyze numerical and binary features, respectively. The most influential features consist of anemia, time, ejection_fraction, and serum_creatninine but ‘time’ features are counted as the highest contribution for the improvement of accuracy.

A comparison of the result obtained shows that different authors came out with different results. The highest accuracy achieved based on past work is 88.52% from the decision tree. Thal features can be categorized as the most influential features seems all the experiments conducted with feature selection show that thal ranked the most among other features. As will be shown in succeeding sections, we analyze and present a comparison with our feature selection technique together with the result of accuracy for heart disease prediction.

III. Methodology

This study is based on the UCI dataset of heart disease which consists of 303 datasets and 13 attributes. The original attributes consist of age, sex, cp, trestbps, chol, fbs, restecg, thalach, exang, oldpeak, slope, ca, thal, and target. Data dictionary in Table I will explain further the attributes involved.

TABLE I. Data Dictionary for Involved Features

Feature Name	Data Description
X_1= Age	age
X_2= Sex	1=male, 0=Female
X_3= Cp	chest pain type: Value 0=typical angina, Value 1=atypical angina, Value 2=non-anginal pain, Value 3=asymptomatic
X_4= Trestbps	resting blood pressure
X_5= Chol	serum cholestrol in mg/dl
X_6= Fbs	(fasting blood sugar > 120) 1=True, 0=False
X_7= Restecg	resting electrocardiographic results: Value 0= normal, Value 1=having ST-T wave abnormality, Value 2=showing probable or definite left ventricular
X_8= thalach	max heart rate achieved
X_9= Exang	exercise induced angine: 1=yes, 0=no
X_{10}=Oldpeak	ST depression
X_{11}= Slope	slope of peak exercise:Value 0=upsloping, Value 1=flat, Value 2= downsloping

X_{12}= Ca	number of major vessel(flourosopy)
X_{13}=Thal(Thalassemia)	0 = error (in the original dataset 0 maps to NaN's),1 = fixed defect,2 = normal,3 = reversable defect
Y= Target	0 = no disease,1 = disease

Cases of heart disease and non-heart disease are extracted from the dataset and displayed visually.

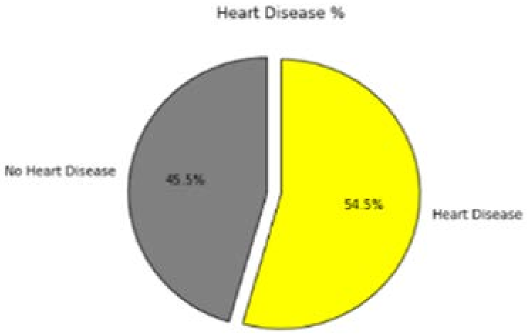

Fig. 1. Data visualization of heart disease and non-heart disease patients

According to Fig. 1, 54.50% of patients suffer from heart disease and the remaining 45.50% are free from heart disease.

The execution is accomplished using the following procedures:

1) UCI Cleveland dataset is obtained
2) Data visualization is performed
3) Dataset is divided into testing and training data
4) Applying algorithms method for training
5) Train the model
6) Heart disease prediction based on accuracy obtain

From the UCI dataset, 80% of the dataset has been assumed as training input for machine learning methods, and the model has been fitted accordingly. The remaining 20% is test data for predicting heart disease [22].

B. Pre-Processing

Dimensionality reduction is a pre-processing procedure that can eliminate irrelevant data, noise, and redundant features to improve the accuracy of learning features and save training time [23]. Data pre-processing often encompasses the following task [24]:

Data cleansing: The first stage in data cleansing is identifying mistakes and inconsistencies in the database by evaluating the data. In other words, this phase is known as data audits and will identify all forms of database irregularities [25].

Normalization: Initially, pre-processing is not only a method for transforming raw data into a clean dataset but it also improves the performance of machine learning. By way of explanation, if data is acquired from various sources, it is collected in a raw format which is incompatible with analysis and machine learning [26].

Feature discovery: Feature discovery is one of the pre-processing methods which is the data filtered from the pre-processing section. The advantage of feature discovery is extracting meaningful data from identified correlations of patterns [27].

Management of imbalance data: An example of an issue known as imbalance data classification is when the proportional class size of a dataset differs significantly by a significant margin from one another. From this, a group of a small number is represented as a minority class and the remaining belong to the other group represented majority class [28].

C. Feature Selection

Attribute or feature selection is a data reduction method that is applied to the dataset. This method decreases the size of the data by eliminating unnecessary or duplicate attributes. Methods for selecting features subset can be broken into four distinct categories which are the embedded method, wrapper method, filter method, and hybrid method [29]. In our research, the features are divided into numerical and categorical which is the filter method applied. As it operates independently from the induction algorithm, this method is faster than the wrapper approach and produces a better generalization. However, the chi-squared method favours selecting a subset with a large number of features, necessitating a threshold to select a subset [30].

According to [31], it is found that the filter approaches are effective, scalable, computationally straightforward, and independent of the classifier. In this research, categorical features consist of sex, fb, restecg, exang, slope, ca, thal and target while age, trestbps, chol, thalach, and oldpeak are numerical features. Chi-squared is used to generate categorical features and ANOVA is tested for numerical features. Both methods generate the features according to rank based on the importance of each feature. The Table II below shows the selected features for categorical and numerical features.

TABLE II. SELECTED FEATURES FOR CATEGORICAL AND NUMERICAL

Feature Selection	Selected Features
Categorical	X_{12}, X_3, X_9
Numerical	X_8, X_{10}

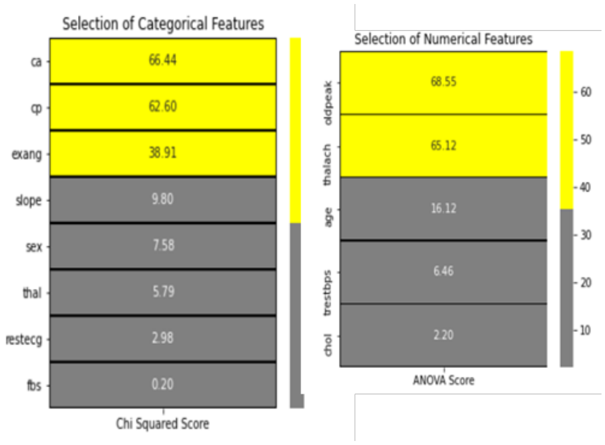

Fig. 2. Ranked each feature selection method for categorical and numerical features

Fig. 2 depicts the ranking of the involved features based on their importance. Ca, cp, and exang counted as the highest rank tested with chi-squared for categorical features while oldpeak and thalach counted as the most influential features for the numerical group using the ANOVA score.

D. Chi-Squared

Chi-squared is one of the techniques for categorical types of data. The chi-squared test determines if two categorical variables are significantly associated. Two-sided chi feature selection is tested between each categorical and binary outcome with a p-value. The features are retained with two-sided p<0.05 [32].

Several steps involved in the chi-square process are explained as follows [33]:

Step 1: All features from the original dataset are selected.

Step 2: Utilize the chi-squared () function from the scikit-learn to figure out whether the two features are independent or not. Use (1) to find the chi-squared score for each of the following features.

$$X^2 = \sum \frac{\left(f_O - f_E\right)^2}{f_E} \tag{1}$$

Step 3: The value with the highest chi-squared value probably relies on the target feature and is therefore selected for model creation. SelectKBest() was utilized to choose the five features with the highest chi-squared value.

Step 4: The next step is to determine a threshold to construct a subset for the number of features represented by n. The optimal number of features with the highest Chi test score is utilized based on the top five ranking features. In this research, five features with the highest Chi test score are tested to create the original feature subset.

According to [34], the strategy for the chi-squared method is incrementally adding important characteristics to the feature subset. At each level, this method will determine the significance threshold and discards features that fall below it. As a result, the chi-squared strategy is more efficient than similar step-wise selection methods. Most of the studies prove the use of the chi-squared method among other feature selection methods improves most of the classifiers' performance and accomplishes outstanding results [35].

Based on [36], up to 1900, the evolution of the chi-squared test process can be divided into six stages. Six related stages included:

1) From the multivariate error law to the multivariate normal distribution.

2) Exponent distribution in multivariate normal density.

3) Multinomial distribution approximation by multivariate normal density.

4) Evaluation of the exponent when the moment is multinomial.

5) The definition to which probability refers.

6) Provision for the effect of estimating an undetermined parameter.

E. ANOVA

Analysis of Variance (ANOVA) is another technique used for the classification method. ANOVA is tested for numerical feature from the dataset and the ratio between variances from two different samples are formulated [33]. For completion of the ANOVA technique, the below step is applied [33]:

Step 1: All features are selected from the original dataset

Step 2: The target feature function from scikit-learn is calculated using ANOVA F-score for each feature. Below (2), (3), (4) are the following formula to calculate ANOVA.

$$F = \frac{\text{variance between groups}}{\text{variance within groups}} \tag{2}$$

$$\text{Variance between groups} = \frac{\sum_{i}^{n} n_i \left(\bar{Y}_i - \bar{Y}\right)^2}{(k-1)} \tag{3}$$

$$\text{Variance within groups} = \frac{\sum_{i=1}^{k} \sum_{j=1}^{ni} \left(Y_{ij-} \bar{Y}_i\right)^2}{(n-k)} \tag{4}$$

Step 3: The result from the test is used to perform feature selection which enables the removal of features that are unrelated to the target variable. The most influential features with the lowest variance are chosen in the experiment and tested with SelectKBest(); *K* represents the number of features for the final dataset.

Step 4: The number of features(*n*) with the highest ranking is used to create various feature subsets.

Research conducted by [37], shows the use of ANOVA can enhance the accuracy which is a 9.1% increase from 72.70%.

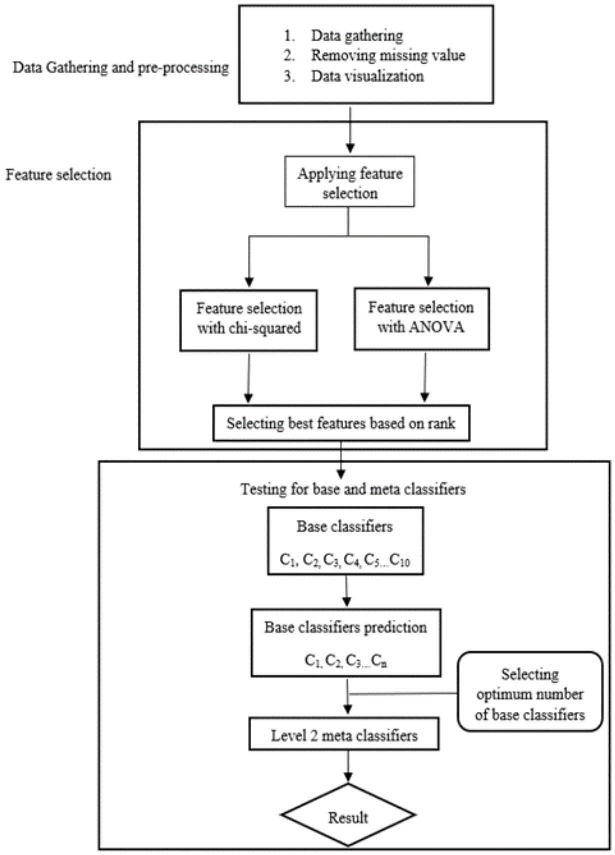
Data Gathering and pre-processing
1. Data gathering
2. Removing missing value
3. Data visualization
Feature selection
Applying feature selection
Feature selection with chi-squared
Feature selection with ANOVA
Selecting best features based on rank
Testing for base and meta classifiers
Base classifiers
$C_1, C_2, C_3, C_4, C_5 ... C_{10}$
Base classifiers prediction
$C_1, C_2, C_3 ... C_n$
Selecting optimum number of base classifiers
Level 2 meta classifiers
Result

Fig. 3. Proposed framework for feature selection method

About Fig. 3, several steps are applied including data gathering and pre-processing before feature selection is applied. 13 attributes from the dataset are extracted to remove the missing value and visualize the data accordingly. Before we go deeper for base and meta classifiers, the feature selection method is applied to the data. Feature selection with the chi-squared technique is applied for categorical features while the ANOVA technique is applied for numerical features. Accuracy is tested for each feature selection method and the best accuracy is selected before we filter the important features. From the experiment made, five important features have been sorted out.

The data are then tested for base and meta-classifier methods. Ten base algorithms consisting of logistic regression (LR), support vector classifier (SVC), random forest (RF), extra tree classifier (ETC), naïve bayes (NB), extra gradient boosting (XGB), decision tree (DT), k-nearest neighbor (KNN), multilayer perceptron (MLP), and stochastic gradient descent (SGD) is tested and result obtained is used to find the optimum number of base classifiers. Then, meta-classifiers are applied for MLP, LR, NB, and SVC algorithms.

IV. RESULTS AND DISCUSSION

The proposed work is using the chi-squared method for categorical features and ANOVA for the numerical feature. 13 features from the UCI dataset are reduced to five features and tested accordingly. Highly rank of features are tested using the required method and there is an improvement in terms of accuracy for each algorithm.

Table III will further explain the involvement of five attributes for chi-squared and ANOVA and the achieved accuracy for each feature selection method.

TABLE III. FIVE CHOSEN ATTRIBUTES FROM CHI-SQUARED AND ANOVA

Selected attributes	**Data Dictionary**
X_{12} =ca	Number of major vessels
X_3 = cp	Chest pain type:Value 0= typical angina,Value 1=atypical angina, Value 2=non-anginal
X_9 = exang	exercise induced angine:1=yes,0=no
X_{10}= oldpeak	ST depression
X_8 =thalach	max heart rate achieved

Chi-squared and ANOVA technique feature selection was the focus of the subsequent testing phase. Ca, cp, exang, oldpeak, and thalach was chosen as the first five attributes selection which is superior to those of another feature selection algorithm.

Accuracy tests for both the base and meta classifiers using these five features and the result show an improvement from the accuracy of base classifiers. Results for both techniques of feature selection which are chi-squared and ANOVA are contracted in the following table.

TABLE IV. ACCURACY OF CLASSIFIERS BEFORE AND AFTER FEATURE SELECTION IS APPLIED

Feature selection				
Algorithms	***Base Classifier***	***Base classifier (after FS applied)***	***Meta classifier***	***Meta-classifiers (after FS applied)***
LR	85.24	91.80	90.16	93.44
RF	83.60	86.89		
KNN	81.96	86.89		
DT	72.13	81.97		
NB	85.24	88.52		
SVC	86.88	91.80	83.60	91.80
XGB	85.24	81.97		
MLP	88.52	90.16	88.52	91.80
SGD	83.60	86.89		
ETC	86.88	83.61		

From Table IV, logistic regression obtains the highest accuracy compared to the other nine algorithms. For the level 1 base classifier, 85.24% is achieved before feature selection is applied and increases to 6.56% after feature selection is applied. Level 2 meta-classifier, increase from 90.16% to 93.44%.

For SVC, the accuracy increases by 4.92% from 86.88% for base classifiers and meta-classifiers, the accuracy achieved is 91.80% from 83.60%. MLP achieved 90.16% accuracy from 88.52% for base classifiers while the accuracy spike from 88.52% to 91.80% for meta-classifiers.

Classification and regression trees (CART) have an acquired accuracy of 87.65%, according to the literature [11]. The author makes an effort to boost precision by employing feature selection and an ensemble technique. There has been some improvement, but the accuracy is still low. In the current research, we suggested the same process but with a new set of features and an alternative method of feature selection. Logistic regression was able to provide a success rate of 93.44 percent, which is an increase over earlier efforts.

V. CONCLUSION

The main goal of this work is to develop hybrid feature selection method for heart disease prediction that combines chi-squared and ANOVA approaches. ANOVA is used to choose numerical data, whereas Chi-squared is used to pick categorical features. The involved algorithms are logistic regression, k-nearest neighbor, decision tree, random forest, gaussian naive bayes, extra gradient boosting, support vector classifier, multilayer perceptron, stochastic gradient descendent, and additional tree classifier. Various algorithms are tested for base classifiers. The meta-classifier is evaluated using the logistic regression, support vector classifier, and multilayer perceptron methods. Then, feature selection techniques are used to evaluate the base and meta-classifiers.

We decided to assess the efficacy of two distinct feature-selection algorithms in this study. Chi-squared tests and analysis of variance are utilized as feature selection methods. The experimental results show that the accuracy of heart disease prediction may be improved by employing the hybrid feature selection technique.

In addition to the utilization of feature selection techniques, the selected features from the dataset are also something that have to be emphasized. The chi-squared test and the analysis of variance (ANOVA) are used to evaluate the results of the experiment regarding five characteristics, namely ca, cp, exang, oldpeak, and thalach. The logistic regression method had a performance that was 93.44% better than the other ensemble stacking techniques. Because the accuracy of the approach might change depending on the dataset that is being used, it will be possible in the future to evaluate the technique of feature selection using a variety of different datasets.

REFERENCES

[1] M. Batta, “Machine Learning Algorithms - A Review,” Int. J. Sci. Res., vol. 18, no. 8, pp. 381–386, 2018, doi: 10.21275/ART20203995.

[2] A. Mohd Ghazi, C. K. Teoh, and A. A. Abdul Rahim, “Patient profiles on outcomes in patients hospitalized for heart failure: a 10-year history of the Malaysian population,” ESC Hear. Fail., vol. 9, no. 4, pp. 2664–2675, 2022, doi: 10.1002/ehf2.13992.

[3] N. A. F. Abu Bakar et al., “Association between a dietary pattern high in saturated fatty acids, dietary energy density, and sodium with coronary heart disease,” Sci. Rep., vol. 12, no. 1, pp. 1–11, 2022, doi: 10.1038/s41598-022-17388-5.

[4] N. Endut, W. M. A. F. W. Hamzah, I. Ismail, M. K. Yusof, Y. A. Baker, and H. Yusoff, “A Systematic Literature Review on Multi-Label Classification based on Machine Learning Algorithms,” TEM J., vol. 11, no. 2, pp. 658–666, 2022, doi: 10.18421/TEM112-20.

[5] I. Tougui, A. Jilbab, and J. El Mhamdi, “Heart disease classification using data mining tools and machine learning techniques,” Health Technol. (Berl)., vol. 10, no. 5, pp. 1137–1144, 2020, doi: 10.1007/s12553-020-00438-1.

[6] S. I. Ansarullah, S. M. Saif, P. Kumar, and M. M. Kirmani, “Significance of Visible Non-Invasive Risk Attributes for the Initial Prediction of Heart Disease Using Different Machine Learning Techniques,” Comput. Intell. Neurosci., vol. 2022, 2022, doi: 10.1155/2022/9580896.

[7] M. Hall and L. Smith, “Feature subset selection: a correlation based filter approach,” Proc. Int. Conf. Neural Inf. Process. Intell. Inf. Syst., pp. 855–858, 1998.

[8] S. E. Fienberg, “The Use of Chi-Squared Statistics for Categorical Data Problems,” J. R. Stat. Soc. Ser. B, vol. 41, no. 1, pp. 54–64, 1979, doi: 10.1111/j.2517-6161.1979.tb01057.x.

[9] P. Schober and T. R. Vetter, “Chi-square Tests in Medical Research,” Int. Anesth. Res. Soc., vol. 129, no. 2, p. 2019, 2019.

[10] M. S. Pathan, A. Nag, M. M. Pathan, and S. Dev, “Analyzing the impact of feature selection on the accuracy of heart disease prediction,” Healthc. Anal., vol. 2, p. 100060, 2022, doi: 10.1016/j.health.2022.100060.

[11] S. Diwan, G. S. Thakur, S. K. Sahu, M. Sahu, and N. K. Swamy, “Predicting Heart Diseases through Feature Selection and Ensemble Classifiers,” J. Phys. Conf. Ser., vol. 2273, no. 1, 2022, doi: 10.1088/1742-6596/2273/1/012027.

[12] S. Bashir, Z. S. Khan, F. Hassan Khan, A. Anjum, and K. Bashir, “Improving Heart Disease Prediction Using Feature Selection Approaches,” Proc. 2019 16th Int. Bhurban Conf. Appl. Sci. Technol. IBCAST 2019, pp. 619–623, 2019, doi: 10.1109/IBCAST.2019.8667106.

[13] A. Ismail et al., “Development of COVID-19 Health-Risk Assessment and Self-Evaluation (CHaSe): a health screening system for university students and staff during the movement control order (MCO),” Netw. Model. Anal. Heal. Informatics Bioinforma., vol. 11, no. 1, 2022, doi: 10.1007/s13721-022-00357-3.

[14] R. Spencer, F. Thabtah, N. Abdelhamid, and M. Thompson, “Exploring feature selection and classification methods for predicting heart disease,” Digit. Heal., vol. 6, pp. 1–10, 2020, doi: 10.1177/2055207620914777.

[15] R. Mohd Rosdan, W. S. Wan Awang, and W. A. Wan Abu Bakar, “Comparison of affinity degree classification with four different classifiers in several data sets,” Int. J. Adv. Technol. Eng. Explor., vol. 8, no. 75, pp. 247–257, 2021, doi: 10.19101/IJATEE.2020.762106.

[16] K. Dissanayake and M. G. M. Johar, “Comparative study on heart disease prediction using feature selection techniques on classification algorithms,” Appl. Comput. Intell. Soft Comput., vol. 2021, 2021, doi: 10.1155/2021/5581806.

[17] A. Alfaidi, R. Aljuhani, B. Alshehri, H. Alwadei, and S. Sabbeh, "Machine Learning: Assisted Cardiovascular Diseases Diagnosis," Int. J. Adv. Comput. Sci. Appl., vol. 13, no. 2, pp. 135–141, 2022, doi: 10.14569/IJACSA.2022.0130216.

[18] A. Lakshmanarao, A. Srisaila, and T. S. R. Kiran, "Heart disease prediction using feature selection and ensemble learning techniques," Proc. 3rd Int. Conf. Intell. Commun. Technol. Virtual Mob. Networks, ICICV 2021, no. Icicv, pp. 994–998, 2021, doi: 10.1109/ICICV50876.2021.9388482.

[19] R. Shanthakumari, C. Nalini, S. Vinothkumar, E. M. Roopadevi, and B. Govindaraj, "Multi Disease Prediction System using Random Forest Algorithm in Healthcare System," 2022 Int. Mob. Embed. Technol. Conf. MECON 2022, pp. 242–247, 2022, doi: 10.1109/MECON53876.2022.9752432.

[20] M. Jena and S. Dehuri, "Decision tree for classification and regression: A state-of-the art review," Inform., vol. 44, no. 4, pp. 405–420, 2020, doi: 10.31449/INF.V44I4.3023.

[21] P. A. Moreno-Sanchez, "Development of an Explainable Prediction Model of Heart Failure Survival by Using Ensemble Trees," Proc. - 2020 IEEE Int. Conf. Big Data, Big Data 2020, pp. 4902–4910, 2020, doi: 10.1109/BigData50022.2020.9378460.

[22] M. Kavitha, G. Gnaneswar, R. Dinesh, Y. R. Sai, and R. S. Suraj, "Heart Disease Prediction using Hybrid machine Learning Model," Proc. 6th Int. Conf. Inven. Comput. Technol. ICICT 2021, pp. 1329–1333, 2021, doi: 10.1109/ICICT50816.2021.9358597.

[23] R. Zebari, A. Abdulazeez, D. Zeebaree, D. Zebari, and J. Saeed, "A Comprehensive Review of Dimensionality Reduction Techniques for Feature Selection and Feature Extraction," J. Appl. Sci. Technol. Trends, vol. 1, no. 2, pp. 56–70, 2020, doi: 10.38094/jastt1224.

[24] M. Kang and J. Tian, "Machine Learning: Data Pre-processing," Progn. Heal. Manag. Electron., pp. 111–130, 2018, doi: 10.1002/9781119515326.ch5.

[25] W. M. N. W. Z. Fakhitah Ridzuan, "A Review on Data Cleansing Methods for Big Data." Elsevier, p. 8, 2019. doi: 10.1016/j.procs.2019.11.177.

[26] J. Jo, "Effectiveness of Normalization Pre-Processing of Big Data to the Machine Learning Performance," J. KIECS., vol. 14, no. 3, pp. 547–552, 2019, doi: https://doi.org/10.13067/JKIECS.2019.14.3.547.

[27] M. J. Hamid Mughal, "Data mining: Web data mining techniques, tools and algorithms: An overview," Int. J. Adv. Comput. Sci. Appl., vol. 9, no. 6, pp. 208–215, 2018, doi: 10.14569/IJACSA.2018.090630.

[28] H. Kaur, H. S. Pannu, and A. K. Malhi, "A systematic review on imbalanced data challenges in machine learning: Applications and solutions," ACM Comput. Surv., vol. 52, no. 4, 2019, doi: 10.1145/3343440.

[29] S. Sreelakshmi and K. G. Preetha, "Innovations in Bio-Inspired Computing and Applications," Adv. Intell. Syst. Comput., vol. 424, no. Ibica, pp. 139–149, 2016, doi: 10.1007/978-3-319-28031-8.

[30] S. Noelia, "Filter Methods for Feature Selection – A," pp. 178–187, 2007.

[31] Y. B. Wah, N. Ibrahim, H. A. Hamid, S. Abdul-Rahman, and S. Fong, "Feature selection methods: Case of filter and wrapper approaches for maximising classification accuracy," Pertanika J. Sci. Technol., vol. 26, no. 1, pp. 329–340, 2018.

[32] S. E. Awan, M. Bennamoun, F. Sohel, F. M. Sanfilippo, B. J. Chow, and G. Dwivedi, "Feature selection and transformation by machine learning reduce variable numbers and improve prediction for heart failure readmission or death," PLoS One, vol. 14, no. 6, pp. 1–13, 2018, doi: 10.1371/journal.pone.0218760.

[33] K. Dissanayake and M. G. M. Johar, "Comparative study on heart disease prediction using feature selection techniques on classification algorithms," Appl. Comput. Intell. Soft Comput., vol. 2021, p. 17, 2021, doi: 10.1155/2021/5581806.

[34] F. Kamalov, H. H. Leung, and S. Moussa, "Monotonicity of the $\chi 2$ -statistic and Feature Selection," Ann. Data Sci., vol. 9, no. 6, pp. 1223–1241, 2022, doi: 10.1007/s40745-020-00251-7.

[35] N. Alotaibi and M. Alzahrani, "Comparative Analysis of Machine Learning Algorithms and Data Mining Techniques for Predicting the Existence of Heart Disease," Int. J. Adv. Comput. Sci. Appl., vol. 13, no. 7, pp. 810–818, 2022, doi: 10.14569/IJACSA.2022.0130794.

[36] R. L. Plackett, "Karl Pearson and the Chi-squared Test," Int. Stat. Rev., vol. 51, no. 1, pp. 59–72, 1983.

[37] M. F. Ihsan, S. Mandala, and M. Pramudyo, "Study of Feature Extraction Algorithms on Photoplethysmography (PPG) Signals to Detect Coronary Heart Disease," 2022 Int. Conf. Data Sci. Its Appl. ICoDSA 2022, vol. 4, no. 2, pp. 300–304, 2022, doi: 10.1109/ICoDSA55874.2022.9862855.