Journal of Computer Science

Original Research Paper

# Study and Analysis of Prediction Model for Heart Disease Data Using Machine Learning Techniques

**B. Santhi and K. Renuka**

*School of Computing, SASTRA University, Thirumalaisamudram, Thanjavur, Tamil Nadu, India*

*Article history*
Received: 14-07-2019
Revised: 09-09-2019
Accepted: 13-03-2020

Corresponding Author:
B. Santhi
School of Computing,
SASTRA University,
Thirumalaisamudram,
Thanjavur, Tamil Nadu, India
Email: shanthi@cse.sastra.edu

**Abstract:** Heart disease is the number one cause of death for all communities of individuals in advanced countries and a major problem for emerging nations too. Doctors' availability to care for the general population could not catch up with the present demand for healthcare. So, there is a severe need for a support system to assist save individuals. With novel ML frameworks and big data repositories, our motive is to design a machine learning model to predict heart disease at the earliest, help prioritize hospital consultations and improve accuracy. For this study, several analyzes were carried out on the Cleveland heart disease data set with 303 patients records, using five different classifiers namely Support Vector Machine (SVM), Random forests, Ordinal Regression, Logistic Regression and Naïve Bayes. Feature selection using chi- squared statistical test and correct tuning of hyperparameters maximized classification accuracy of the Support vector machine (Radial basis function) from 40% to 85%. By incorporating rules based on the statistical patterns observed, the efficiency was further enhanced to 95%. On the other side, seeing it as a 5-class classification, multi-class imbalance issue was addressed using suitable sampling techniques that resulted in 96% accuracy for 5-class data. We evaluated model efficiency using k-fold cross validation and confusion matrix. This study shows that the classification accuracy could be significantly improved by balancing the dataset using sampling and by properly tuning hyperparameters after feature selection.

**Keywords:** Machine Learning, Cardio Vascular Disease, Cardiology, Support Vector Machine, Accuracy, Hyperparameter Optimization, Sampling Techniques

## Introduction

Heart disease is the number one cause of death for all groups of people in developed countries and is a great burden for developing nations too. Not only, 82% of the total Disability-Adjusted Life Year (DALYs) are because of coronary heart disease but also attribute to 43% of all cardio vascular disease deaths (https://www.heart.org/idc/groups/ahamah-ublic/@wcm/@sop/@smd/documents/downloadable/ucm_470704.pdf; https://www.who.int/cardiovascular_diseases/en/cvd_atlas_13_coronaryHD.pdf). But the death rates vary among countries. High income countries have rates of approximately 38%.

In Eastern Europe, the rates are more as 58% to as low as 10% in Sub-Saharan Africa (Gaziano *et al*., 2010). This variable disease prevalence is due to multiple factors. After the industrial revolution, countries started moving to industrial, post-industrial states and the social, environmental and structural changes led to many lethal and chronic diseases. Thus growth in industrial development gives a bane of more risks to health as a bonus. A recent study by HUNT says that pregnancy complications are also a cause for heart disease among women (Markovitz *et al*., 2019; CDC, 2019).

Age is one of the factors for heart diseases because our blood vessels become less flexible when we get older. So, blood would find it hard to flow freely thus increasing the pressure. Sometimes, fats get deposited on the walls of artery obstructing the pumping of blood leading to attacks. Poor exercise and bad nutrition remain as the major cause (https://www.healthline.com/health-news/how-poor-diet-raises-your-risk-of-dying-from-heart-disease).

Body ailments such as high blood pressure, diabetes are added risk factors. About four out of five people

Science Publications

who die of coronary heart disease are 65 or older. The importance of cardiovascular symptoms, unavailability of a doctor at all places and times, led to the search of other ways of detecting diseases.

Machine learning, a wide discipline with mathematics and computer science as its roots, offers a number of new algorithms and methodologies to build inferential and predictive data-driven models. With novel ML frameworks and large repositories of data, now ML started concentrating on the healthcare sectors. Currently it has been shown in clinical cardiology; ML is more skilled in predicting cardiac and all-cause death than manual approaches used individually for clinical or imaging purposes. Most frequently used machine learning algorithms are Logistic Regression, Artificial Neural Networks, Support Vector Machines, Tree-based methods and ensemble methods (Brownlee, 2019). Typically, the data sets used in ML projects are divided into training, validation and test subsets; training sets covering the bulk of all available data are used to primarily develop the model, validation sets are used to estimate the overall performance of the model or to fine-tune its hyper parameters.

Having discussed the significance of the problem and machine learning algorithms, this study gives an insight on related works in Section 2, workflow of the research in Section 3, analyses of the outcomes in Section 4 and conclusion in Section 5.

## Related Work

Several techniques in this area have been proposed to enhance the level of accuracy in heart disease prediction. Different scholars have used and contributed to different to techniques in this field. Suvarna *et al*. (2017) used the Particle Swarm Optimization technique, which is an inherently distributed algorithm where the problem solution arises from the interactions between many simple individual agents called particles. They also used a moderately altered PSO version with the constricted PSO factor that produced competitive outcomes. Dhanashree *et al*. (2013) used Naïve Bayes. Durairaj and Revathi (2015) attempted to detect the presence of heart illness using Artificial Neural Network's Back Propagation (MLP) Multilayer Perceptron. For the categorization of ECG beat, Emanet (2009) used a random forest algorithm and discrete wavelet transformation. Five types of ECG signals were ranked with a score of 99.8%. Because the Random Forest algorithm operates very fast, offers great effectiveness and there is no cross verification, it can be helpful for the ECG's long-term beat ranking. Parthiban *et al*. (2011) applied Naive Bayes, which uses a minimal level training set to produce an efficient prediction model. Parthiban and Subramanian (2008) provided a fresh strategy centered on a co-active neuro-fuzzy inference scheme for which the genetic model was used to automatically adjust the parameters of the visual network as well as to select the function set. Purushottam and Sharma (2015) have developed a model that can effectively find the rules for predicting patients' risk levels based on their health parameter. The rules were prioritized according to the necessities of the user. Srivastava and Bhambhu (2010) have compared the efficiencies of different kernels in SVM for all data samples. Radhimeenakshi's (2016) work incorporates the classes of Heart Disease, utilizing Support Vector Machine (SVM) as well as Artificial Neural Network (ANN). Examination is completed among two strategies on the premise of accuracy and training time. Amin *et al*. (2013) focused on selecting key features using genetic neural network based analysis and other techniques such as regression. Observations show that the prediction accuracy has improved greatly for the proposed method of applying feed forward neural network classifier using risk factors of heart disease. They also used a method involving two most effective techniques of data mining, neural networks and genetic algorithms. The implemented hybrid system utilizes the genetic algorithm's global optimization advantage to initialize neural network weights. Tomov (2018) have found a 5-layer neural network architecture (HEARO-5) using regularization, optimization and k-way cross-validation to tune the design. Using the above created model, they obtained 99% accuracy and 0.98 Mathews Correlation coefficient. Yan *et al*. (2006) used a computing model based on a three-layer neural network of Multilayer Perceptron (MLP) to create a decision support framework for the treatment of five significant heart illnesses. Their experimental findings have shown that the chosen decision framework based on MLP can reach a large amount of precision (63.6-82.9%) in heart disease diagnosis.

## Proposed Work Flow

This work (with flow Fig. 1) utilizes heart disease data from the Cleveland Database. Data have been preprocessed where missing values are handled and sampling methods are used to manage the problem of class imbalance. Features are then chosen using the feature Importance plot acquired from classifiers and chi-square test (as a way of Dimensionality reduction). To train and test, these processed data are supplied to the existing classifiers (SVM Linear and Gaussian Kernel, Naive Bayes, Logistic and Ordinal Regression). This method is repeated with hyperparameter tuning to obtain the best performance which is verified during the post-processing stage. Using a five-point summary and plots built between the features; rules have been constructed and with precisely tuned hyperparameters, all the classifiers (enhanced) are used to train and test the data. The effectiveness of the enhanced classifiers has been proven using performance metrics in the post-processing stage by comparing the precision of the results with the performance of the existing classifiers.

The inference from this analysis is that more precise outcomes are achieved with correct tuning of hyperparameters after incorporating rules based on statistical concepts. Hence the substitution of rules and hyperparameters provided a new set of classifiers which are termed as Enhanced SVM, Enhanced Naive Bayes and Enhanced Regression models. Three different studies were conducted on the dataset with the above-mentioned workflow, namely considering the data as binary class, 3-class and multi-class data which are discussed in the next sections.

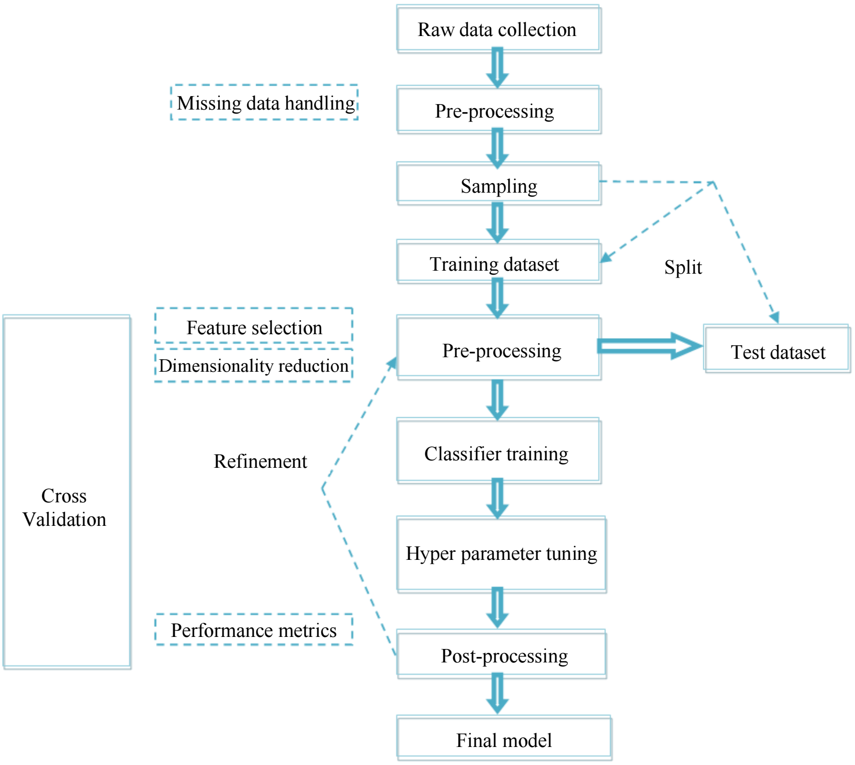


**Fig. 1:** Work flow diagram

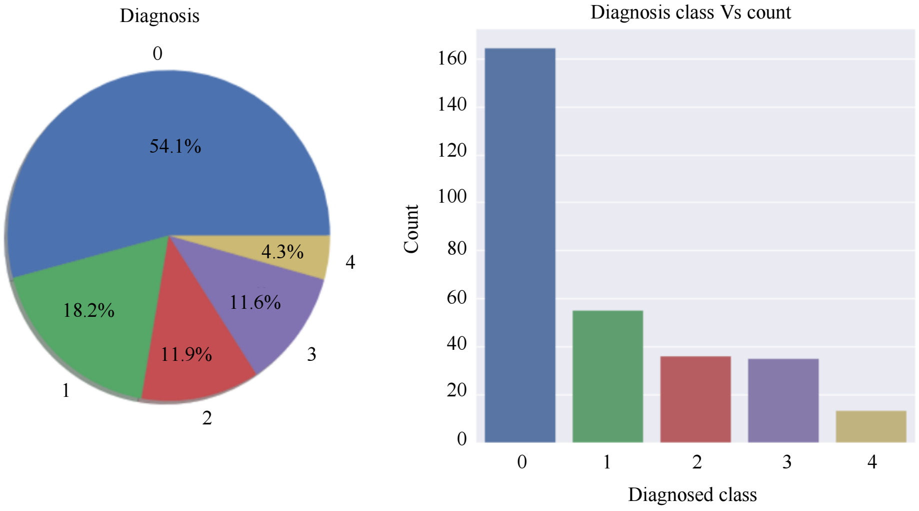


**Fig. 2:** 5-class data visualization before sampling (original data is 5-class data)

### Data

The Cleveland Database (https://archive.ics.uci.edu/ml/datasets/Heart+Disease) used in this study has records of 75 attributes from 303 patients. But the dataset selects and lists only 14 of them. They are age, sex, chest pain as cp, resting blood pressure as trestbps, cholesterol as chol, blood sugar as fbs, electro-cardiographic results as restecg, heart rate achieved (maximum) as thalach, exercise induced angina as exang, ST depression in ecg as oldpeak, slope of the peak exercise ST as slope, number of major vessels colored by fluoroscopy as ca, thalassemic results as thal and the variable to be predicted "num" for abnormality. The abnormality level is within range 0-4. Cleveland database experiments focused on attempting to distinguish existence from absent (values 1,2,3,4) (value 0).

### Data Sample

A data set sample is as follows (Table 1):

### Data Visualization

As per the database (Fig. 2), the data is a multi-class data with five different levels 0-4. The plot above shows the count of patients in each level of disease for e.g., 54.2% being level 0 (a high class-imbalance could be seen). The standard machine learning algorithms in this scenario could be biased and inaccurate. Ways to improve the performance in such ***class-imbalance*** problem are:

- Try to gather more information
- Change the performance metrics for proper assessment
- ***Resampling the dataset (used in our study)***
- Generate synthetic samples
- Trying various algorithms
- Trying penalized models

### Supervised Learning Types

Supervised algorithms in this study are:

### Linear Regression

Regression analysis is a statistical method for the assessment and simulation of the relationship between variables. If we depict the dependent variable as y and the independent variable as x, the equation between these two variables is a simple line:

$$Y_i = \beta_i + \beta_i X_i + \varepsilon_i \tag{1}$$

where, $Y_i$ is the dependant variable, $\beta_0$ is the $Y$ intercept, $\beta_1$ is the slope intercept, $X_i$ is the independent variable and $\varepsilon$ is the Random error term is the distinction between the measured value of $y$ and the line ($\beta_0 + \beta_1 x$). It is useful to think of it as a statistical mistake which is the random variable which accounts for the model's inability to match precisely the information. As far as the notion of statistical autonomy is concerned $x$ is the regressor variable and $y$ is the reaction variable. The outcome (dependent variable) in linear regression is continuous.

### Logistic Regression

Logistic regression is a statistical model which, although there are many more complicated extensions, utilizes a logistic function in its fundamental form to model a binary dependent variable. Logistic regression (or logit regression) estimates the parameters of a logistic model (a form of binary regression) in regression analysis. This is done here and could also be used in instances involving more than two categories using multinominal regression.

### Ordinal Regression

Ordinal regression is a statistical method used with a collection of independent variables to predict behavior of ordinal class dependent variables which is the outcome of the ordinal type and the independent variable can be either categorical or continuous. Ordinal regression in stats is a sort of regression assessment used to predict an ordinal variable, i.e., a variable whose value happens on an indefinite scale where only the relative ordering between the different values is relevant.

### Naive Bayes

It is a supervised learning algorithm often connected with Natural Language Processing (NLP). This is a straightforward algorithm depending on likelihood. It provides a quick model construction for small data collection with all class conditions. They predict the likelihood of class affiliation such as the likelihood that a specified tuple will belong to a specific category using BAYES THEOREM:

$$P(H \mid C) = \frac{P(H) * P(C \mid H)}{P(C)} \tag{2}$$

where, $P(H|C)$, $P(H)$, $P(C)$ and $P(C|H)$ denotes posterior probability of '$H$' given the evidence, prior probability, priori probability that the evidence itself is true and the likelihood of the evidence '$C$' being true respectively.

### Support Vector Machines

It is a supervised learning that is used for classification as well as regression. It involves making a hyper plane that best separates the classes. In other words, considering the labeled training data (controlled learning), an ideal hyperplane is produced by the algorithm that categorizes new instances. If this cannot be accomplished in the input space the objective will be accomplished in a higher dimensional space by mapping the input vector through a non- linear function. But for time and storage limitations it is extremely dependent on the size of the dataset.

For better classification, kernel features are used to map information points into a greater dimensional space. There are various kernel kinds. By using a linear kernel, the data is separated linearly by using a straight line. It is good at classification of two classes at a time. For linear kernel the equation for prediction for a new input using the dot product between the input ($x$) and each support vector ($x_i$) is calculated as follows:

$$f(x) = B(0) + sum\left(a_i * \left(x, x_i\right)\right) \qquad (3)$$

This is an equation in which the internal products of a new input vector ($x$) are calculated with all support variables in training data. The learning method has to estimate the coefficients $B(0)$ and $x_i$ (for each sample from the training data). Polynomial kernel is similar to vectors in a feature space over polynomials of the original variables. The polynomial kernel and exponential kernel are expressed as follows:

$$K\left(x, x_i\right) = 1 + sum\left(x * x_i\right)^d \qquad (4)$$

$$K\left(x, x_i\right) = \exp\left(-gamma * sum\left(\left(x, x_i^2\right)\right)\right) \qquad (5)$$

Radial Gaussian kernel is best for non linear data. Sigmoid kernels are also used (Fig. 3).

### *Feature Selection by Chi –Square Test*

A chi-squared test is a statistical hypothesis test in which the sampling distribution of the test statistics is a chi-squared distribution if the null hypothesis is accurate. Chi-squared test' is often used as brief for Pearson's chi-squared test without any other qualification. The chi-squared test is done to ascertain if a significant difference exists in one or more categories between the expected values and the measured values used to select the feature.

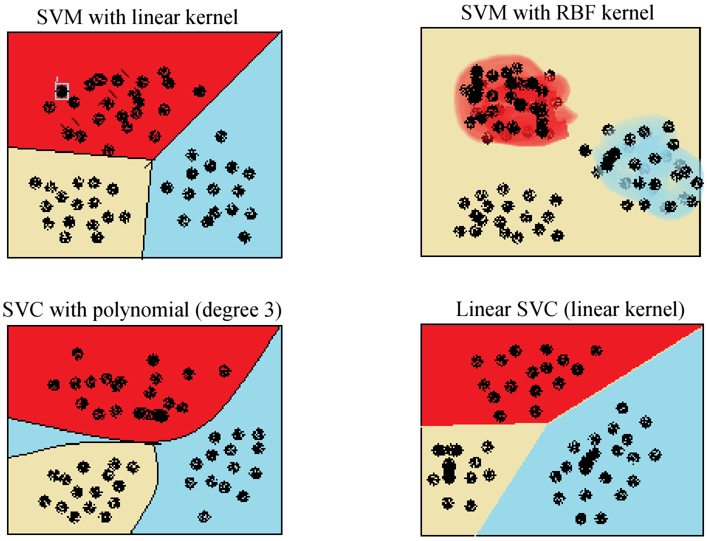


**Fig. 3:** Impact of gamma

| | | Predicted class | | |
|---|---|---|---|---|
| | | Yes | No | Total |
| Actual class | Yes | TP | FN | P |
| | No | FP | TN | N |
| | Total | P’ | N’ | P+N |

**Fig. 4:** Confusion Matrix

### *Cross Validation*

The initial data $A$ is randomly split in k-fold cross-validation into $k$ mutually exclusive subsets $A_1$, $A_2$... $A_k$, each of roughly the same size. Training and testing are performed $k$ times. Subset $A_1$ is chosen as the test set in phase 1 and the left over subsets are used for training the model. Likewise, for other $k$ values, the method is performed for $k$ number of times. It prevents overfitting.

### *Confusion Matrix*

Given m classes, a confusion matrix consists of a table of at least m by m size. Using this performance of the model is measured (Fig. 4).

## Results Analysis

Firstly, the dataset was considered as binary-class data (presence or absence of disease) instead of levels 0-4. Five different classifiers SVM with Linear kernel, SVM with Gaussian Kernel, Naive Bayes, Logistic Regression and Ordinal Regression are used in this study. The first line of the Table 2 shows the resulting accuracy of all the five classifiers with the initial dataset (after pre-processing the missing values) taking into account all the features in the modeling. The model generated using the training set with the five classifiers presented approximately 85% percent precision with test data except for the SVM Gaussian kernel, which only provided 40% accuracy.

Since not all features present in the dataset will be significant in properly anticipating the heart abnormality, feature selection techniques have been introduced to reduce the input column space of the data set. Thus, from the 13 characteristics, the top 5 features were selected by leaving out the insignificant features using the classifiers' Importance plot Fig. 5 based on their weightage during prediction. This also produced the same accuracy as before with an additional advantage of decreasing size. It was anticipated that the classifiers would give 85% accuracy with only the top five characteristics. SVM Gaussian kernel showed an enhancement in effectiveness by delivering 75% accuracy after feature selection. The hyperparameters such as C, gamma were altered due to the tunable nature of its hyperparameters and such a finest mix of SVM Gaussian coefficients yielded 85 percent precision with selection of features. The results of this process are listed in the second row of the Table 2.

The Third and fourth rows of the Table 2 list the results when dimensionality reduction is carried out on the original dataset using Chi-square test with k (output dimension) being 4 and 5.

The dataset is imbalanced here. Thus, sampling methods such as up sampling and down sampling were used to balance the dataset, resulting in an enhanced output of 92 percent precision for SVM with Gaussian kernel (with tuned hyperparameters C and gamma) which is entered in the fifth and sixth rows of the Table 2.

While evaluating the outcomes of the dataset's classifiers, five-point summary and bar charts (Fig. 6a-6c and 7) showing relationships between the reduced features of the dataset and the column to be predicted "Diagnosis", domain knowledge-based pattern analysis was made and those patterns were designed as rules to create some sort of multivariate constrained rules. After these rules were included, the efficiency of all the classifiers enhanced (Fig. 8). 95 percent precision was given by SVM with Gaussian kernel as shown in the last row of the Table 2 which is more than the standard ensemble algorithm's accuracy on this dataset.

Similarly, second study is made by considering the five- level data as a three-level data and repeating all the above analysis (original dataset with no feature selection, after feature selection using feature importance plot and chi-square test) on the preprocessed data yielded the following results (1-4 rows of Table 3).

There is a serious downfall in the accuracies due to class-imbalance issue. But, after sampling there is a hike in accuracy (5th row of Table 3) in SVM with Gaussian kernel with proper mix of hyperparameters.

Table 4 shows the same analyses on the five class data where the initial accuracy was too low (first row of Table 4). But, with feature selection, tuning of hyperparameters and upsampling yielded a fair accuracy for SVM with Gaussian Kernel due to its high tunable nature.

**Table 1:** Sample data set cp-chest pain, trestbps – resting BP, fbs-fasting sugar, exang -exercise induced angina (1 yes), ca-number of colored major vessel, thalach – peak heart rate achieved, oldpeak-ST depression induced, slope- slope of ST peak exercise, num-diseased or not

| | age | sex | cp | trestbps | chol | fbs | restecg | thalach | exang | oldpeak | slope | ca | thal | num |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 63.0 | 1.0 | 1.0 | 145.0 | 233.0 | 1.0 | 2.0 | 150.0 | 0.0 | 2.3 | 3.0 | 0.0 | 6.0 | 0 |
| 2 | 67.0 | 1.0 | 4.0 | 160.0 | 286.0 | 0.0 | 2.0 | 108.0 | 1.0 | 1.5 | 2.0 | 3.0 | 3.0 | 1 |
| 3 | 67.0 | 1.0 | 4.0 | 120.0 | 229.0 | 0.0 | 2.0 | 129.0 | 1.0 | 2.6 | 2.0 | 2.0 | 7.0 | 1 |
| 4 | 37.0 | 1.0 | 3.0 | 130.0 | 250.0 | 0.0 | 0.0 | 187.0 | 0.0 | 3.5 | 3.0 | 0.0 | 3.0 | 0 |
| 5 | 41.0 | 0.0 | 2.0 | 130.0 | 204.0 | 0.0 | 2.0 | 172.0 | 0.0 | 1.4 | 1.0 | 0.0 | 3.0 | 0 |

**Table 2:** All classifiers on Binary class data

| | SVM Linear | SVM Gaussian | Naïve Bayes | Log.Reg | Ordinal Reg |
|---|---|---|---|---|---|
| Original | 0.8688 | 0.4080 | 0.8524 | 0.8524 | 0.8524 |
| Feature Importance Plot | 0.8360 | 0.8524 | 0.8360 | 0.8360 | 0.8360 |
| Chi square test k = 4 | 0.8688 | 0.8688 | 0.8524 | 0.8688 | 0.8688 |
| Chi square test k = 5 | 0.8360 | 0.8524 | 0.8196 | 0.8360 | 0.8360 |
| Upsampling | | | | | |
| | 0.9125 | 0.9250 | 0.9050 | 0.9250 | 0.9250 |
| DownSampling | 0.8210 | 0.875 | 0.8392 | 0.8210 | 0.8210 |
| Proposed | **0.9504** | **0.9508** | | | |
| | | **C = 10000 Gamma = 0.0001** | **0.9429** | **0.9414** | **0.9418** |

**Table 3:** All classifiers on 3-class data

| | SVM Linear | SVM Gaussian | Naïve Bayes | Log.Reg | Ordinal Reg |
|---|---|---|---|---|---|
| Original | 0.5573 | 0.4592 | 0.5901 | 0.6393 | 0.6721 |
| Feature Importance Plot | 0.6229 | 0.5737 | 0.5700 | 0.6229 | 0.6390 |
| Chi square test k = 4 | 06550 | 0.4090 | 0.6050 | 0.6557 | 0.6557 |
| Chi square test k = 5 | 0.5409 | 0.4262 | 0.5901 | 0.5737 | 0.6393 |
| Upsampling | 0.6767 | 0.9200 | | | |
| | | C = 10000 | | | |
| | | Gamma = 0.0001 | 0.6868 | 0.6363 | 0.6666 |
| DownSampling | 0.5667 | 0.7330 | 0.7000 | 0.6000 | 0.7667 |

**Table 4:** All classifiers on 5-class data

| | SVM Linear | SVM Gaussian | Naïve Bayes | Log.Reg | Ordinal Reg |
|---|---|---|---|---|---|
| Original | 0.4590 | 0.4750 | 0.4098 | 0.4590 | 0.5409 |
| Feature Importance Plot | 0.4750 | 0.4918 | 0.4590 | 0.4918 | 0.5240 |
| Chi square test k = 4 | 0.4918 | 0.5081 | 0.4900 | 0.5081 | 0.4918 |
| Chi square test k = 5 | 0.4562 | 0.4910 | 0.4910 | 0.5081 | 0.5245 |
| Upsampling | **0.6951** | **0.9695** | | | |
| | | **C = 10** | | | |
| | | **Gamma = 0.01** | **0.5182** | **0.5914** | **0.4634** |
| DownSampling | 0.5000 | 0.7330 | 0.5000 | 0.5330 | 0.4000 |

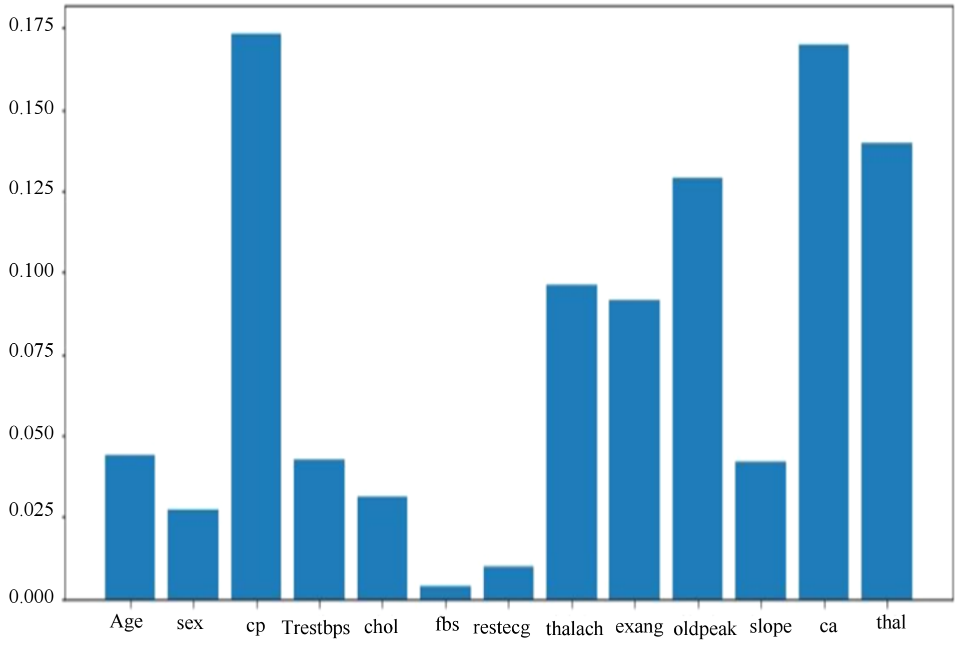
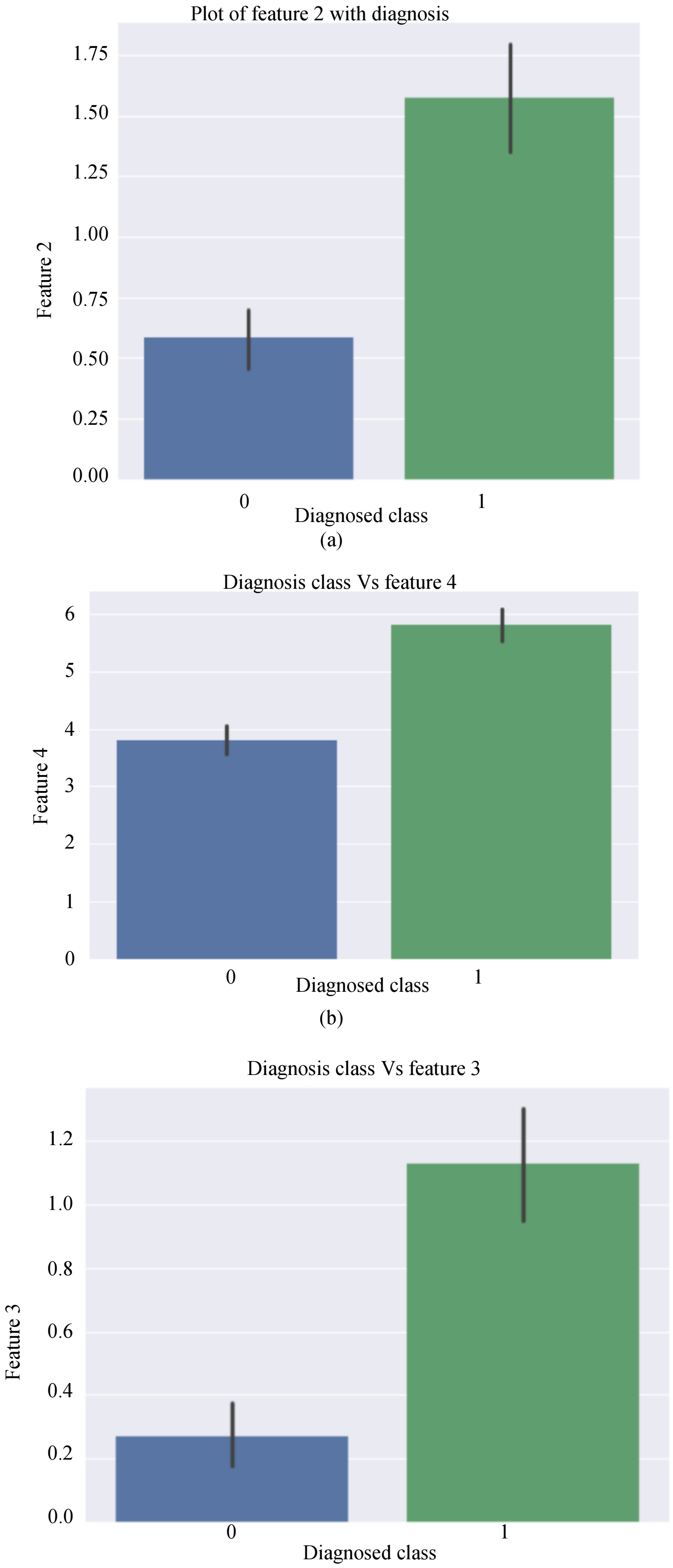


**Fig. 5:** Feature importance plot



(a)

(b)

(c)

**Fig. 6:** (a) Feature 2 Vs Diagnosis; (b) Feature 4 Vs Diagnosis; (c) Feature 3 Vs Diagnosis

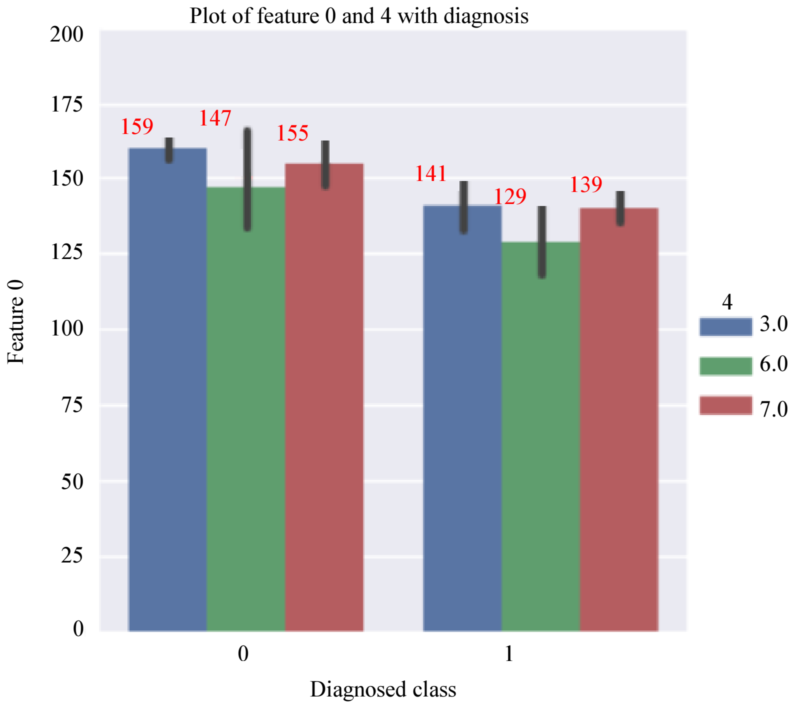


**Fig. 7:** Bar plot showing the statistical patterns, showing the relationships between the label (Diagnosed class), Feature 0 (with two levels), Feature 4 (with 3 levels) such that these values are used in the construction of rules

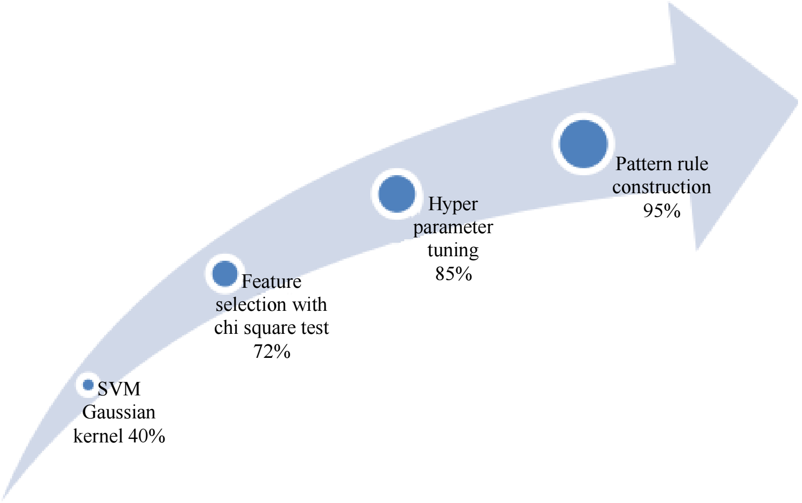

SVM Gaussian kernel 40%
Feature selection with chi square test 72%
Hyper parameter tuning 85%
Pattern rule construction 95%

**Fig. 8:** SVM Gaussian kernel on binary-class data

The observations are:

### *Hyperparameter Tuning*

At first the prediction accuracy for Support vector Machines with Radial Basis Function was 40%. Then, feature selection increased its performance from 40% to 72%. With tuning of hyperparameters, the accuracy reached 85%.

### *Rule Construction*

By visualizing three features at a time using different plots (count plot, bar plot) and obtaining five point summary of the dataset (mean, median, mode and quartiles), a pattern was observed. Rules were constructed to exploit the observed pattern which increased the accuracy to 95%.

### *Balancing the Data Through Sampling*

And multi class data imbalance problem was handled using sampling techniques leading to an accuracy of 96%.

## Conclusion

This study focuses primarily on developing a machine learning model for predicting heart disease at the early stage, helping to prioritize patient consultations and improving accuracy. We used the Cleveland heart disease database from the UCI machine learning repository, which has records of 75 characteristics from 303 patients. But the data set only chooses and lists 14 of them. The same dataset is implemented with different classification methods namely Support vector machine (linear and Gaussian kernel), Regression (logistic and ordinal) and naive bayes under distinct types of conditions, such as original, data after feature selection using feature importance plot, chi squared statistical test and sampling (upsampling and downsampling) techniques. Among all SVM classifiers with Gaussian kernel there was an improvement in efficiency at each stage, i.e. from 40% to 72% after feature selection and from 72% to 85% after correct hyperparameter tuning. The output improved to 95 percent with the inclusion of rules based on observed statistical patterns. Similar to the study of the binary class data, the dataset is as such taken as a data of 5 classes. Multi-classification related issues such as the issue of class imbalance have been solved using suitable upsampling methods. Thus, the study demonstrates that the precision of the classification could be considerably enhanced by balancing the dataset using sampling and correctly tuning hyperparameters after feature selection is done.

Future work involves expanding the analysis to build a more comprehensive system that involves ECG and other records of graphic data. More characteristics can give the algorithm more information to learn from, create a more complicated design and ensure a more precise and comprehensive prediction. In the future, the model will also be developed into a quality software with a user friendly interface, for easy use by doctors and patients.

## Acknowledgment

This research was carried out at SASTRA University, School of Computing, and we would like to extend our gratitude to the institution for providing us with the necessary computing resources and guidance.

## Author's Contributions

**Santhi, B.:** Structural design of the paper, Dataset selection and content revision.

**Renuka, K.:** Coding , Analysis, Content Writing and formatting.

## Ethics

There are no ethical issues in publishing this manuscript.

## References

Amin, S.U., K. Agarwal and R. Beg, 2013. Genetic neural network based data mining in prediction of heart disease using risk factors. Proceedings of the IEEE Conference on Information Communication Technologies, Apr. 11-12, IEEE Xplore Press, Thuckalay, Tamil Nadu, India. DOI: 10.1109/CICT.2013.6558288

Brownlee, J., 2019. A tour of the most popular machine learning algorithms. Machine Learning Algorithms.

CDC, 2019. Women and heart disease. Center for Disease Control and Prevention.

Dhanashree, S.M., M.P. Bote and S.D. Deshmukh, 2013. Heart disease prediction system using naive bayes. Int. J. Enhanced Res. Sci. Technol. Eng., 2: 1-5.

Durairaj, M. and V. Revathi, 2015. Prediction of heart disease using back propagation MLP algorithm. Int. J. Sci. Technol. Res., 4: 235-239.

Emanet, N., 2009. ECG beat classification by using discrete wavelet transform and random forest algorithm. Proceedings of the 5th International Conference on Soft Computing, Computing with Words and Perceptions in System Analysis, Decision and Control, Sept. 2-4 IEEE Xplore Press, Famagusta, Cyprus, pp: 1-4. DOI: 10.1109/ICSCCW.2009.5379457

https://archive.ics.uci.edu/ml/datasets/Heart+Disease

https://www.healthline.com/health-news/how-poor-diet-raises-your-risk-of-dying-from-heart-disease

https://www.heart.org/idc/groups/ahamah-ublic/@wcm/@sop/@smd/documents/downloadable/ucm_470704.pdf

https://www.who.int/cardiovascular_diseases/en/cvd_atlas_13_coronaryHD.pdf

Gaziano, T.A., A. Bitton, S. Anand, S. Abrahams-Gessel and A. Murphy, 2010. Growing epidemic of coronary heart disease in low- and middle-income countries. Curr. Probl. Cardiol., 35: 72-115. 10.1016/j.cpcardiol.2009.10.002

Markovitz, A.R., J.J. Stuart, J. Horn, P.L. Williams and E.B. Rimm, 2019. Does pregnancy complication history improve cardiovascular disease risk prediction? Findings from the HUNT study in Norway. Eur. Heart J., 40: 1113-1120. DOI: 10.1093/eurheartj/ehy863.

Parthiban, G., R. Appusamy and S. Srivatsa, 2011. Diagnosis of heart disease for diabetic patients using naive bayes method. Int. J. Comput. Applic. DOI: 10.5120/2933-3887.

Parthiban, L. and R. Subramanian, 2008. Intelligent heart disease prediction system using CANFIS and genetic algorithm. Int. J. Biol. Med. Sci., 3: 157-160.

Purushottam, K.S. and R. Sharma, 2015. Efficient heart disease prediction system using decision tree. Proceedings of the International Conference on Computing, Communication Automation, May 15-16, IEEE Xplore Press, Noida, India, pp: 72-77. DOI: 10.1109/CCAA.2015.7148346

Radhimeenakshi, S., 2016. Classification and prediction of heart disease risk using data mining techniques of support vector machine and artificial neural network. Proceedings of the 3rd International Conference on Computing for Sustainable Global Development, Mar. 16-18, IEEE Xplore Press, New Delhi, India, pp: 3107-3111.

Srivastava, D. and L. Bhambhu, 2010. Data classification using support vector machine. J. Theoretic. Applied Inform. Technol.

Suvarna, C., A. Sali and S. Salmani, 2017. Efficient heart disease prediction system using optimization technique. Proceedings of the International Conference on Computing Methodologies and Communication, Jul. 18-19, IEEE Xplore Press, Erode, India, pp: 374-379. DOI: 10.1109/ICCMC.2017.8282712

Tomov, S. and S. Tomov, 2018. On deep neural networks for detecting heart disease.

Yan, H., Y. Jiang, J. Zheng, C. Peng and Q. Li, 2006. A multilayer perceptron-based medical decision support system for heart disease diagnosis. Exp. Syst. Applic., 30: 272-281. DOI: 10.1016/j.eswa.2005.07.022