Heart Disease Detection using Bio-Markers

1st Prashant Pradip Jadiya
Department of Electrical and Computer Eng.
Stevens Institute of Technology
Hoboken, USA
pjadiya@stevens.edu

2nd Deepika Vasu Kumar
Department of Mathematical Sciences
Stevens Institute of Technology
Hoboken, USA
dvasukum@stevens.edu

3rd Abhinayrao Janagama
Department of Mathematical Sciences
Stevens Institute of Technology
Hoboken, USA
ajanagam@stevens.edu

***Abstract*—This project focuses on identifying Heart disease with bio-makers. To solve this problem and get this topic into research and analysis we used multiple machine learning algorithms with techniques like Cross validation, SMOTE for upsampling, feature extraction, and incremental effect of features. This will be used as an edge to existing solutions as we combined mostly all methodologies and analysed in proper manner.**

***Keywords*—Particle Swarm Optimization, SMOTE, Machine learning, Heart disease, Bio-marker, Neural Network**

I. Introduction

Heart disease is the label for various types of heart conditions. Several symptoms and risk factors involve that lead to heart disease. People around the world are highly vulnerable to heart disease than any other disease. Several reports show that men are more vulnerable to heart disease than women in US and all around the world. Every 1 out of 4 men die due to heart disease regardless of their race and ethnicity in US. [1]

In this paper, dataset from University of California-Irvine(UCI) Machine Learning repository is used to predict the presence of heart disease given the history of patients. The first step is to analyze and understand the feature set that contributes towards the presence of Heart disease. Exploratory Data Analysis(EDA) is used to understand the data types, handling missing values, understanding mean,median and percentile of each features, data visualization, normalizing and one hot encodeing, Preprocessing the data.Dataset has 14 features, 13 independent features and 1 dependent feature. It has 4 binary features (sex, fbs, exang, Target), 5 categorical features (cp, restecg, slope, ca, thal) and 5 continuos features (age, trestbps, chol, thalach, oldpeak).

The purpose of this paper is to help people detect the presence of heart disease in earlier stage more accurately and reduce sudden death rate. As the outcome of this process plays a crucial role in patient's life, accuracy of outcome has to be more precise. To achieve more accuarte result, 6 form of same dataset is used to build models such as Logistic Regression(LR), Support Vector Machines(SVM), Latent Dirichlet Allocation(LDA), Guassian Naive Bayes(GNB), Gaussian Process Classifier(GPC), Decision Tree(DT), K-Nearest Neighbors(KNN), Random Forest(RF), AdaBoost(AB), Stochastic Gradient Descent(SGD) and neural networks using three methodologies without Cross Validation(CV), with CV and with SMOTE. Feature selection techniques such as Partical Swarm Optimization(PSO) and Chi square (SelectKBest) are used to understand the significance of features and their contribution towards increasing the accuracy.

II. Related Work

R. Perumal et al. [2] suggested technique which includes PCA for feature reduction and machine learning algorithms such as LR, SVM, and KNN with accuracy values 87% , 85% and 69% respectively. D. Ananey-Obiri et al. [3] induced some other models like DT and Gaussian Naive bayes with feature reduction performed using SVD (singular value decomposition) and they got 4 features from 13. They got accuracy score of 82.75% .

S. Mohan et al. [4] made hybrid random forest with linear model that can be called as HRFLM; made to enhancement in prediction of Heart disease with Random forest they got best error rates. S. K. J. et al. [5] analyzed and compared the results of two models such as Decision Tree classifier with 91% accuracy and Naive Bayes classifier with 87% accuracy.

S. Ekız et al. [6] compared the performance of two tools such as MATLAB and Weka. Implemented classification models for heart disease dataset. Several classification algorithms like SVM, DT, Radial Basis Kernal, Subspace Discriminant are used to build model. Compared the results on the accuracy and time taken to process for each algorithm. Accuracy metrics for Naive Bayes and Decision Tree algorithm were 87% and 91%, respectively.

III. Our Solution

Our approach to this problem is to first we learn the dataset, get insights about it and find correlated variables to make use of them in training a machine learning algorithm. We have implement four approaches to tackle the problem as it is more preferred in research and deep analysis.

A. Description of Dataset

Dataset is retrieved from this link. This dataset has 303 instances, with 75 attributes but only 13 of them can be used. This attributes can be called bio-markers as they are age, sex, chest pain type (0-asymptomatic, 1-atypical angina, 2-non-anginal pain, 3-typical angina, trestbps is person's resting blood pressure (mm Hg), chol is person's cholesterol measurement (mg/dl), fbs is fasting blood sugar (if it is $>$ 120 mg/dl then value=1 else 0), restecg is resting electrocardiographic results (0: probable, 1: normal, 2: abnormality), thalach is maximum heart rate, exang is exercise induced angina (1:yes, 0: no), oldpeak is ST (ECG plot) depression, slope indicates the slope of peak exercise ST segment (0:downslopping, 1:flat, 2:upslopping), ca is number of vessels (0-3), thal is thalassemia evidence (0:null, 1:fixed, 2:normal blood flow, 3: abnormal

blood flow), target shows whether that patient has heart disease or not (0:yes, 1:no). This table shows baseline characteristics of the dataset.

There are attributes who have p-value below 0.001; they are significant. We got this p-value using T-test and ANOVA test. Only two features who is not having p-value lesser than 0.001 are trestbps and chol.

SN	Attributes	Descriptions	Attribute types	Mean ± SD	*p*-value
1	age	Age	Continuous	54.4 ± 9.1	<0.001
2	sex	Sex	Categorical	-	<0.0001
3	cp	Chest pain type	Categorical	-	<0.0001
4	trestbps	Resting blood pressure	Continuous	70.9 ± 15.5	0.013
5	chol	Serum cholestrol	Continuous	-	0.136
6	fbs	Fasting blood sugar	Categorical	122.1 ± 11.3	<0.0001
7	restecg	Resting Electrocardiogram	Categorical	-	<0.0001
8	thalach	Maximum heart rate	Continuous	178.6 ± 96.6	<0.001
9	exang	Exercise induced angina	Categorical	-	<0.001
10	oldpeak	ST depression	Continuous	154.5 ± 33.5	<0.001
11	slope	Slope of peak exercise	Categorical	-	<0.001
12	ca	Number of major vessels	Categorical	-	<0.001
13	thal	Thalassemia	Categorical	-	<0.001

Fig. 1: Baseline Characteristics of Dataset

We adopted some visualisation techniques like box-plot, pair plot, scatter plot to analyse the data in terms of finding correlations, outliers and to understand the dataset that we have.

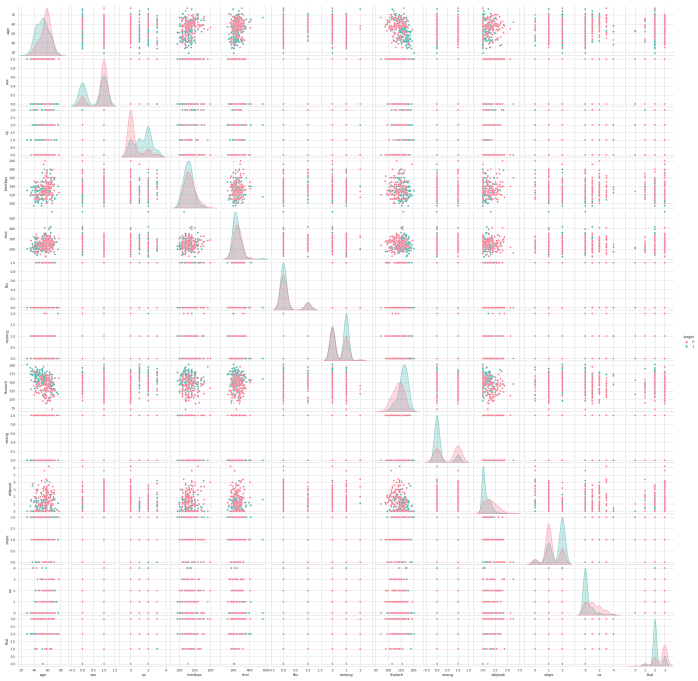

Fig. 2: Pair plot

The pair plot enables us to know the pairwise relationships specifically best set of features in the dataset. As noticed in the above pair plot, pairs like chol-trestbps, chol-age are having positive correlation and thalach-age are having negative correlation. Discrete plots simply indicate the categorical features.

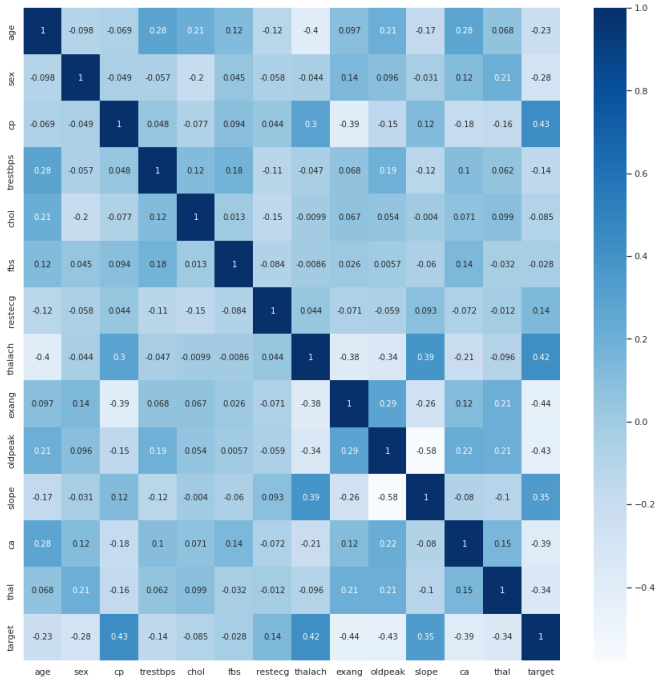

Fig. 3: Correlation Matrix

From the above correlation matrix it is evident that not a single feature has strong correlation that stands out. But features like cp, thalach, slope are the top features that has positive correlation compared to other features. Features like exang, oldpeak, ca are in the top list for negative correlation.

By both plots we can conclude that some of the best features to make use of in training a model are **thalach**, **cp**, and **slope**. There are some of the outliers in the dataset for specific attributes like trestbps, chol, thalach, and oldpeak; we have dropped them to improve the accuracy. And some values which are falsely entered in dataset. For example, in ca attribute the value should only between 0-3 but there are instances where value is 4. So we have considered them as null values and replaced them with median value. Some facts are concluded from the dataset such as: Lesser the number of vessels, more the chances of having heart disease; Diabetes increase the chance of having heart disease.

B. Methodology

We plan to use numerous machine learning algorithms such as Logistic Regression, K-Nearest Neighbours, Support Vector Machine, Linear Discriminant Algorithm, Naive Bayes, Gaussian Process Classifier, Decision Tree, Random Forest, Stochastic Gradient Descent, Boosting algorithm like AdaBoost and Neural Network with 9 Layers. Additionally, we have four methodologies i.e. 1) Traditional machine learning 2) Cross Validation 3) up-sampling with SMOTE 4) Including Feature selection. In parallel we have made six datasets from the same origin i.e. df1: Original dataset, df2: cleaned null values and removed duplicates from df1, df3: handled outliers from df2, df4: applied MinMaxScaling in df2, df5: applied standard scaling in df2, df6: applied RobustScaling in df2. We are going to implement all eleven algorithms including

Neural Network and comparing each algorithm with respective dataset; then we will conclude which dataset is good for respective algorithm.

For Feature selection, we have implemented chi-square and PSO algorithm. Moreover, to do more analysis, we have done Incremental effect of features in the end.

C. Implementation Details

As discussed above, we executed this problem in four ways; so all four methods will have different results and those will be discussed in following.

1) Traditional Method: Traditional method means that just applying the algorithms to the datasets without any add-ons after applying train test split (70% for train and 30% for test). Comparing all algorithms with this method gives us the best performing algorithm for scaled dataset and for non-scaled dataset.

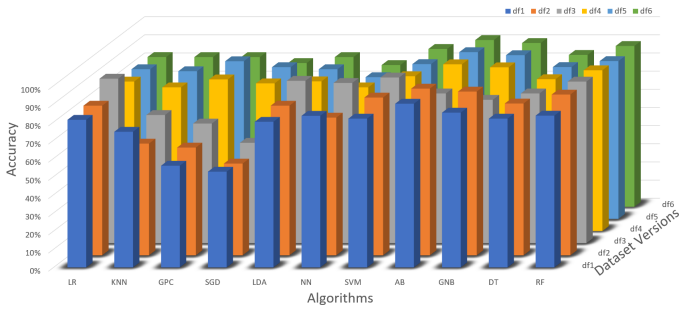

Fig. 4: Results for Traditional Method

From this method we can induce that AdaBoost is most accurate algorithm as it gives 91.80% accuracy whether dataset is scaled or not.

2) Cross-Validation: As we thought, traditional method might be overfitted or biased, so that we tried to do Cross validation as in 5-fold CV.

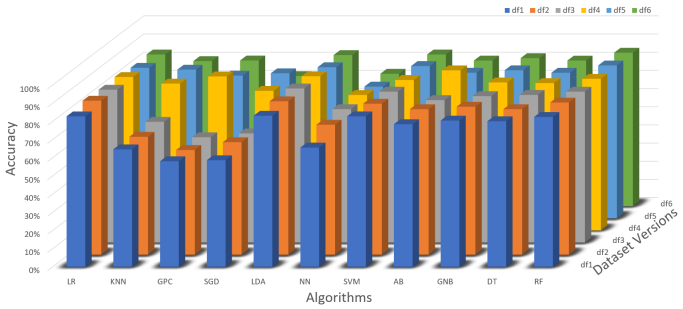

Fig. 5: Results for Cross-Validation

Here, we got decrement in our accuracy chart as we expected. But best performing algorithm remains same. AdaBoost is accurate at 88.09% for scaled dataset. Though, Logistic regression results are improved by 2 to 3 percentage compare to previous method.

3) Up-sampling with SMOTE: While executing the algorithms, we have experienced that the dataset is baised. Not balanced with target values i.e. Out of 303 patients, we have 165 patients who don't have Heart disease and 139 who has disease. So we tried to implement some balancing techniques like SMOTE. We implemented it to balance data and multiply the existing instances with 10. So this being 10x SMOTE.

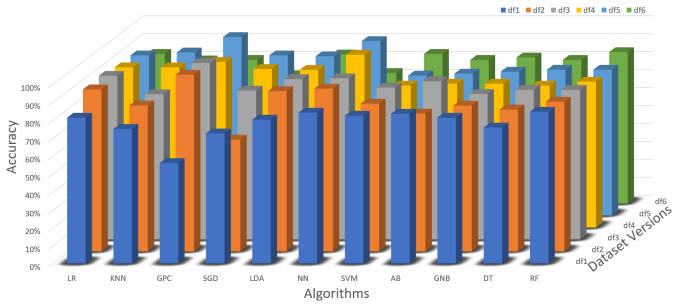

Fig. 6: Results for SMOTE

4) Feature Selection: Alternative way to make our models accurate, is doing dimensionality reduction and feature selection. We have used population intelligence algorithms Particle Swarm Optimization algorithm in order get some significant features of the dataset. As a result, we got top significant columns (five) and fed it to algorithm to compare it with traditional method. We have choosen GPC for the same.

PSO algorithm is an optimization algorithm which mimics the behaviour of finding food in birds and animals. It is similar with Genetic algorithm that finds best fit for future generation. Here this means by particles, process includes updating the particle position, velocity and fitness. Then it will return the best position and fitness of the particle in the result. Here we have chosen n_particles=30 and got five most prominent features. [7]

We have implemented another feature selection technique using chi-square (SelectKBest). As results are attached in chart, PSO was significantly improving the performance for all dataset versions (i.e. df1, df2 etc) whereas, Chi-square feature selection gave pretty good results but it was stable for all versions of datasets.

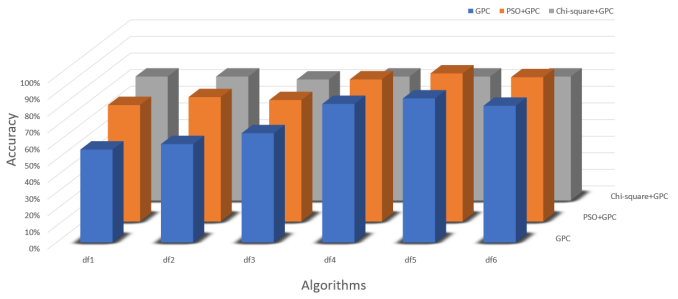

Fig. 7: Comparison of GPC, PSO+GPC, Chi-square+GPC

5) Additional findings: After analysing all methodologies we have performed we were impressed with results and

findings we got. Though, there were several methodology to analyse our process in terms of **Incremental Effect of Features**. Here we have considered Age and Sex as Base features of the dataset. So we will test each from rest of features combining with base features. For example, in first iteration we will train model on Base features and oldpeak; then will test the accuracy. After that in 2nd iteration we'll combine base features and trestbps for training and testing the model, and so on.

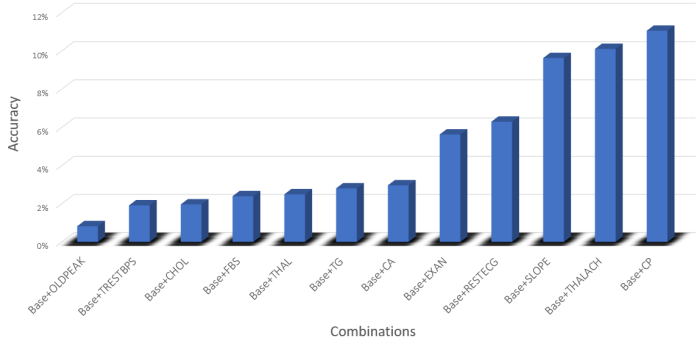

Fig. 8: Incremental Effect of Features for LR with df3

In the result, we can see that **cp** feature is most prominent for the accuracy of Logistic Regression algorithm. This is a kind of research methodology that focuses on individual feature performance.

IV. Future Work

As we have maximized all of the possibilities of methodology to improve performance, now we are looking for some real-world (from doctors) datasets which has two doctor's opinions. There we can test our model for **Inter-variability analysis** that how our model performs for same patient and compare it with other doctor's response.

References

[1] Virani SS, Alonso A, Aparicio HJ, Benjamin EJ, Bittencourt MS, Callaway CW, et al. Heart disease and stroke statistics—2021 update: a report from the American Heart Association *Circulation* 2021;143:e254–e743.

[2] Ramya Perumal, Kaladevi AC, "Early Prediction of Coronary Heart Disease from Cleveland Dataset using Machine Learning Techniques", *IJAST*, vol. 29, no. 06, pp. 4225 - 4234, May 2020.

[3] Ananey & Obiri, Daniel Sarku, Enoch. (2020). Predicting the Presence of Heart Diseases using Comparative Data Mining and Machine Learning Algorithms. *International Journal of Computer Applications*. 176. 975-8887. 10.5120/ijca2020920034.

[4] S. Mohan, C. Thirumalai and G. Srivastava, "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques," in *IEEE Access*, vol. 7, pp. 81542-81554, 2019, doi: 10.1109/ACCESS.2019.2923707.

[5] S. K. J. and G. S., "Prediction of Heart Disease Using Machine Learning Algorithms.," 2019 *I*st International Conference on Innovations in Information and Communication Technology (ICIICT), 2019, pp. 1-5, doi: 10.1109/ICIICT1.2019.8741465.

[6] S. Ekız and P. Erdoğmuş, "Comparative study of heart disease classification," 2017 *E*lectric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT), 2017, pp. 1-4, doi: 10.1109/EBBT.2017.7956761.

[7] Xue, Bing & Zhang, Mengjie & Browne, Will. (2013). Particle Swarm Optimization for Feature Selection in Classification: A Multi-Objective Approach. *I*EEE transactions on cybernetics. 43. 1656-1671. 10.1109/TSMCB.2012.2227469.