Prediction on Heart Disease Using Artificial Intelligence and Machine Learning Techniques

Preetham John
School of Informatics, Computing & Cyber Systems
Northern Arizona University, Flagstaff, AZ, U.S.A.
pj323@nau.edu

I. Abstract

Heart related diseases or Cardiovascular Diseases (CVDs) are the main reason for a huge number of deaths in the world over the last few decades and has emerged as the most life-threatening disease, in the whole world. Machine Learning algorithms and techniques have been applied to various medical datasets to automate the analysis of large and complex data [1]. Many researchers, in recent times, have been using several machine learning techniques to help the health care industry and the professionals in the diagnosis of heart related diseases.

Medicine and hospital resources these days have become more expensive and difficult to afford. There are lot of regions where medical resources are still out of reach. Working on this is one of my major interests. Analyzing the previous data and having given some useful data we can do the prediction analysis. Cardiomyopathy and Cardiovascular disease are some categories of heart diseases. The reduction of blood and oxygen supply to the heart leads to heart disease [2]. In this paper the data classification is based on supervised machine learning algorithms which result in accuracy, time taken to build the algorithm.

Keywords *: Cardiovascular Diseases; K-Nearest Neighbor, Support Vector Machine, Logistic Regression, Decision Tree Classifier, Random Forest Classifier, XG-Boost Classifier; K-Fold Cross-Validation; Featureless; Ensemble Models.*

II. Introduction

The Heart is an organ about the size of your fist, and it is one the most important organ in our body and it carries out one of the main functions necessary for life. It pumps blood to every part of our anatomy, And If it fails to function correctly, then the brain and various other organs will stop working, and within few minutes, the person will die. Heart diseases are considered as one of the most risky and prominent cause of death all around the world. There are many causes for a person get into this medical situation work related stress and bad food habits contribute to the increase in rate of several heart related diseases. We know that Artificial Intelligence (A.I.) is influencing mankind on a whole new different level. And one such field that intrigues me the most is the Medicine. AI in simple terms mean making a machine think and respond rationally. These days we have huge amounts of data and the size of these datasets are in billions so working manually on these and getting the required output is a challenging task. However, Automation has made our life easier.

III. Problem Statement

According to World Health Organization (WHO) Cardiovascular diseases (CVDs) are the number 1 cause of death globally, taking an estimated 17.9 million lives each year. CVDs are a group of disorders of the heart and blood vessels and include coronary heart disease, cerebrovascular disease, rheumatic heart disease and other conditions. Four out of Five CVD deaths are due to heart attacks and strokes, and one third of these deaths occur prematurely in people under 70 years of age.Medical organizations, all around the world, collect data on various health related issues. These data can be exploited using various machine learning techniques to gain useful insights. But the data collected is very massive and, many a times, this data can be very noisy. These datasets, which are too overwhelming for human minds to comprehend, can be easily explored using various machine learning techniques. Thus, these algorithms have become very useful, in recent times, to predict the presence or absence of heart related diseases accurately. There are a lot of difficulties in the field of medical as we see through the world from a different perspective. If this research is made available to all we can save a lot of lives [2]. Some of the important problems are listed below.These days the expenses of medical facilities is extremely high all over the world. Families with low income must suffer a lot with this kind of situations. We

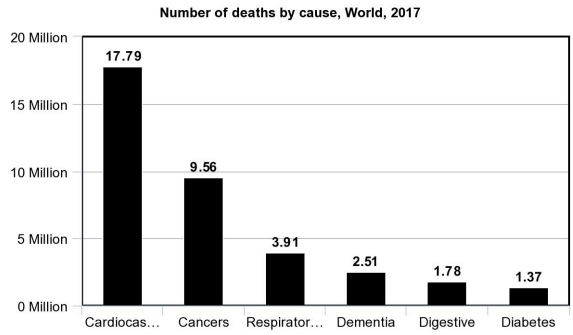

Fig. 1: Problem Statement : Real World Data

also need to understand the there will be some human errors at hospitals due to some issues but with proper algorithms we can make sure that they are fixed and can help us all.

Number of deaths by Cause: I have collected the data from different sources and plotted the graph for Number of deaths by Cause Where Y-axis has the labels *Cardiocased Diseases; Cancers; Respiratory Diseases; Dementia; Diabetes.* [Fig 1]

IV. Existing Model

We use some of the existing techniques that we can work on and give a sample picture of the algorithms and how you can use them to excel in our predictions for the betterment of the people to use it. Increasing the accuracy of the prediction is one of the main goals. We will be working with large number of datasets in order to get the proper diagnosis of the data. In Fig 2 we can see that there is a model which shows how we can progress of our work. We first collect the Raw Data and the data is then pre-processed which also means that we are training the data. The trained data is then passed to to the actual models the three main models that are K-Nearest Neighbor, Support Vector Machine and Logistic Regression and when have the results from this data training models we then pass it on to calculating the Accuracy of our test results. When we have the test results accuracy we then predict the possibilities or the potential of that person if at all he has a heart disease or is there a chance in the near future of him having the same. Algorithms: *K-Nearest Neighbor, Support Vector Machine, Logistic Regression.*

V. Proposed / Working Model.

The proposed system has been developed with the aim to classify people with heart disease and healthy people. The performances of different machine learning predictive models for heart disease diagnosis on full and selected models were tested. So in this model there is a implementation of many more algorithms that we are

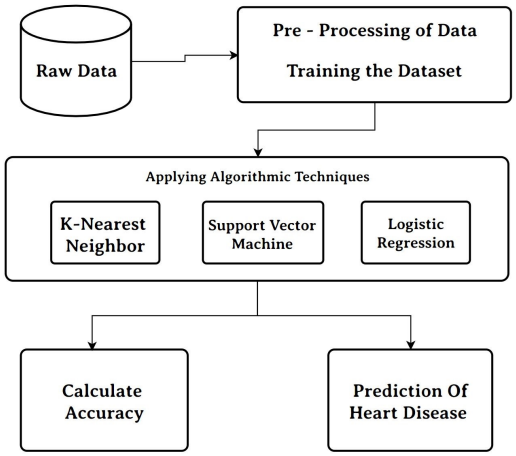

Fig. 2: Existing Model

using and more advanced methods for this project. We are pre-processing and training the data in a better way and in a proper way in order to get better results from the complete model. In order to make your model really robust, simply evaluating with a train/test split may not be enough for that reason we are introducing K-Fold cross validation into the system. By generating train/test splits across multiple folds, you can perform multiple training and testing sessions, with different splits. The main thing is that when we are working with the models we are also introducing classifiers for are model which are very helpful in the final model prediction. As we are using many models for the project good results are expected. We are introducing the classifier models in simple words classifier is nothing but it maps the input data to a specific category. What else can a classification model do ?. A classification model tries to draw some conclusion from the input values given for training. It will predict the class labels/categories for the new data is one of the advantage for model prediction of our experiment. Algorithms: *K-Nearest Neighbor, Support Vector Machine, Logistic Regression, Decision Tree Classifier, Random Forest Classifier, XGBoost Classifier* .

A. K-Nearest Neighbor

K-Nearest Neighbor technique is one of the most elementary but very effective classification techniques. It makes no assumptions about the data and is generally be used for classification tasks when there is very less or no prior knowledge about the data distribution. This algorithm involves finding the k nearest data points in the training set to the data point for which a target value is unavailable and assigning the average value of the found data points to it [1]. The k-nearest neighbors (KNN) algorithm is a simple, supervised machine learning algorithm that can be used to solve both classifi-

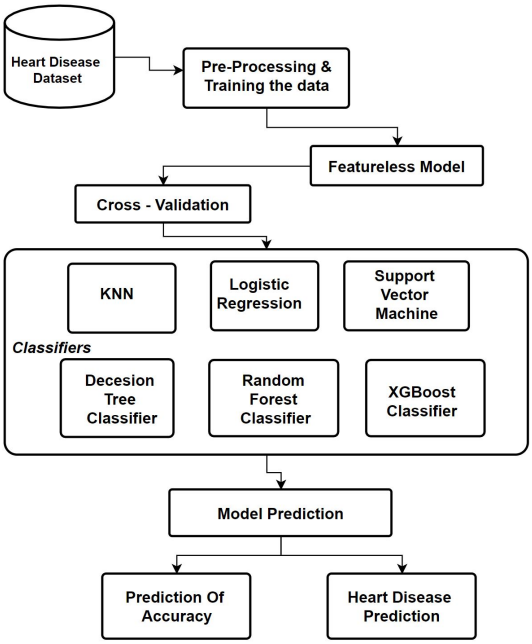

Fig. 3: Proposed / Working Model

cation and regression problems. It's easy to implement and understand but has a major drawback of becoming significantly slows as the size of that data in use grows.

B. Support Vector Machine

Support Vector Machine is an extremely popular supervised machine learning technique(having a predefined target variable) which can be used as a classifier as well as a predictor. For classification, it finds a hyperplane in the feature space that differentiates between the classes. An SVM model represents the training data points as points in the feature space, mapped in such a way that points belonging to separate classes are segregated by a margin as wide as possible. The test data points are then mapped into that same space and are classified based on which side of the margin they fall [1].

C. Logistic Regression

Logistic regression is named for the function used at the core of the method, the logistic function. The logistic function, also called the sigmoid function was developed by statisticians to describe properties of population growth in ecology, rising quickly and maxing out at the carrying capacity of the environment. It is an S-shaped curve that can take any real-valued number and map it into a value between 0 and 1, but never exactly at those limits. 1 / (1 + eˆ-value). Where e is the base of the natural logarithms (Euler's number or the EXP () function in your spreadsheet) and value is the actual numerical value that you want to transform. Below is a plot of the numbers between -5 and 5 transformed into the range 0 and 1 using the logistic function.

D. Decision Tree Classifiers

One of the Supervised learning algorithms is the decision tree classifier algorithm. Let us see how the classifier works - It performs effortlessly with continuous and categorical attributes. This algorithm divides the population into two or more similar sets based on the most significant predictors.Decision Tree algorithm, first calculates the entropy of each and every attribute. Then the dataset is split with the help of the variables or predictors with maximum information gain or minimum entropy. These two steps are performed recursively with the remaining attributes [1]. Decision Tree however did not perform the best in this test we got really poor results from this model and which was very helpful for our studies.

E. Random Forest Classifiers

Random forest classifier creates a set of decision trees from randomly selected subset of training set. It then aggregates the votes from different decision trees to decide the final class of the test object. Random forest has nearly the same hyper-parameters as a decision tree or a bagging classifier. Random forest adds additional randomness to the model, while growing the trees. Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean/average prediction [4]. In our work Random forest did perform well up-to one extent it gave us some pretty good results for the model we built.

F. XGBoost Classifiers

XGBoost is a decision-tree-based ensemble Machine Learning algorithm that uses a gradient boosting framework. In prediction problems involving unstructured data. It implements machine learning algorithms under the Gradient Boosting framework. XGBoost provides a parallel tree boosting (also known as Gradient Boosted Decision Tree GBDT, Gradient Boosting Machine GBM) that solve many data science problems in a fast and accurate way. t is a highly flexible and versatile tool that can work through most regression, classification and ranking problems as well as user-built objective functions. However it showed us pretty good results in our model. This was helpful in the prediction of our model.

No	Attributes	Description
1	Age	Age in year
2	Sex	0 for female and 1 for male
3	Cp	Chest pain type Value 1: typical angina Value 2: atypical angina Value 3: non-anginal pain Value 4: asymptomatic
4	Trestbps	Resting blood sugar in mm Hg on admission to the hospital
5	Chol	Serum cholesterol in mg/dl
6	Fbd	(Fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
7	Restecg	Resting ECG result
8	Thalach	Maximum heart rate achieved
9	Exang	Exercise induced angina
10	Oldpeak	ST depression induced by exercise relative to rest
11	Slope	Slope or peak exercise ST segment
12	Ca	Number of major vessels colored by fluoroscopy
13	Thal	Defect type
14	num	The predicted attribute

Fig. 4: Dataset Used

1) Validation Method of Classifiers - K-Fold Cross-Validation. : We used K-fold cross-validation (CV) method in this research paper. So what do we mean my K-fold cross-validation and how are we sing this in our system. In k-fold cross-validation, the data set is divided into k equal size of parts, in which k 1 groups are used to train the classifiers and remaining part is used for checking out performance in each step. The process of validation is repeated k times [4]. The classifier performance is computed based on k results. For CV, different values of k are selected. In our experiment, we used k = 10 because its performance is good. In the 10-fold CV process, 90 % data were used for training and 10% data were used for testing purpose. The process was repeated 10 times for each fold of process, and all instances in the training and test groups were randomly divided over the whole dataset prior to selection training and testing new sets for the new cycle. Lastly, at the end of the 10- fold process, averages of all performance metrics are computed [4].

VI. Understanding The Data Used

Datasets in a perfect world is a flawlessly curated group of observations with no missing values or irregularities. However, this is not true. It can be disordered, which means it needs to be clean and wrangles. Data cleaning is a essential part in data science problems. Machine learning models learn from data. It is crucial; however, that the data you feed them is precisely pre-processed and refined for the problem you want to solve. This includes data cleaning, pre-processing, feature engineering, and so on. The data that is being used is large so training the data is very necessary.

Table Sample Dataset Used : The data which is used is plotted in a tabular format with 14 rows and 3 columns. [Fig 4]

	Models	Training Accuracy %	Testing Accuracy %
1	Logistic Regression	86.79	86.81
2	K-Nearest Neighbor	86.79	86.91
3	Support Vector Machine	93.40	87.91
4	Decision Tree Classifier	100.0	78.02
5	Random Forest Classifier	100.0	82.42
6	XGBoost	98.58	83.52

Fig. 5: Results : Algorithm and their accuracy

VII. Project Work

The model i have designed has six main models that also includes three Classifiers which i mentioned earlier. so what i am working on is getting the best accuracy for these datasets. The new thing that i am trying to implement is introduction of new models to train the given data for better accuracy. The new model called the **Proposed Intelligent System Model** refer Fig. 3 where all the models are combined together to get more accurate results. Working on this model i have implemented K-Fold cross validation. In k-fold cross-validation, the data set is divided into k equal size of parts, in which k 1 groups are used to train the classifiers and remaining part is used for checking out performance in each step. The process of validation is repeated k times. The classifier performance is computed based on k results. For CV, different values of k are selected. As of now we have worked with these six data training models and have tried to derive the accuracy of these data. So what exactly this research is going to help me with ?. Well as we all now get an understanding that this is a prediction model with this we can develop in the future as a helpful software by usage of all the data. When a person can enter some of his personal data it can run and do the analysis giving us an output of the complete prediction whether the person's Heart is at risk or if is it safe.

Test Accuracy Results : The Results obtained are shown here with two columns test accuracy and training accuracy. [Fig 5].

Results Plot : All the results are plotted according to the performance of the model [Fig 6].

VIII. Results and Discussion

So many people in the world are working on this research in order to make this world a better place and help as much as possible. So all we are doing is trying get the best results in order to predict the correct the human errors and predict the correct analysis. We need to understand the necessity of the demand for this research. The most important thing is working with the data, So in this experiment K-Fold Cross-Validation (k = 10) Classifiers with 10-fold cross-validation method.

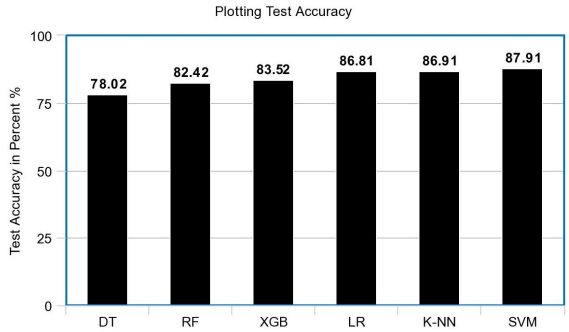

Fig. 6: Plotting the Test Accuracy Results

In 10-fold CV, 90 % was used for training the classifiers and 10% was used for testing. Finally, the average metrics of 10-fold methods were computed. As of now the best results that are obtained are with KNN model testing accuracy of 86.81% and also of the logistic regression with 86.91%Ȯn more and clearer observations we can see is that performance of Decision Tree performed poorly the testing results were around 78.02 % which was the lowest even if the training data gave us the results of 100%Ṫhe model which performed the best was surprisingly Support Vector Machine even though the training data test was 93.40% the testing performance was 87.91%Ṫhe Testing accuracy of Random Forest Classifier was 100% the testing accuracy 82.42%ẊGBoost performed with an accuracy of 83.52 %Ẇe are trying to work on more models to get the research done where will be trying to push the accuracy of my model to around 90% with the most available tools and the dataset. I have also Plotted and tabulated the results for easy understanding to verify refer Fig 5. for Algorithm results and their accuracy and also refer Fig 6. for Graph showing all the results of the model tabulated in a Acsending order. As we have progressed from my literature review there were many challenges and also many more models i have been working on will discuss in the next section:

1) There is a severe lack of abundant data. This seems like the main challenge as data in the area of Heart disease and the data is always kept safe no given access to the free world because a lot these data can be sold and also misused .
2) The accuracy of algorithms cannot be generalized there are many possible errors that are still possible by the mankind. To avoid these we work on maximum training of the given dataset for the accuracy.

IX. Future Work

We can see some really good results by the working of the proposed model Each of the above-mentioned

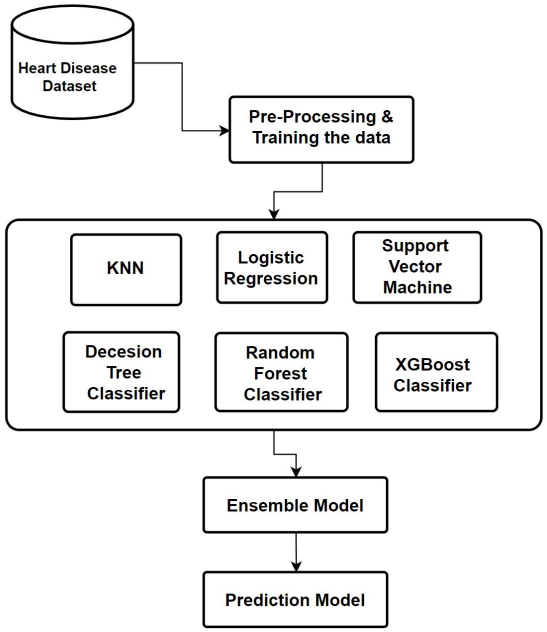

Fig. 7: Ensemble Model

algorithms have performed extremely well in some cases but poorly in some other cases. This will help us in analyzing more models in the future. Referring to other papers the most important thing which i will be implementing in this section is the Ensemble model. In ensemble modeling two or more related but different analytical models are used and produce their results are combined into a single score. TahiraMahboob et al. [3] have used an **Ensemble** of SVM, KNN, ANN and many other models to achieve an accuracy of 94.12% . So this is one of the most important model in this research which can help us get better accuracy. Will be implementing many more models in the near future for better accuracy and will be loaded into the ensemble model for better results.

A. Ensemble Model

What is an Ensemble Model ?. Ensemble modeling is a process where multiple diverse models are created to predict an outcome, either by using many different modeling algorithms or using different training data sets [11]. The ensemble model then aggregates the prediction of each base model and results in once final prediction for the unseen data. In the future will be working on more on these models with ensemble method. As this is a vast topic there is a lot more research that can be done especially on this one.

X. Conclusion

Classification algorithms are used to predict small set of relations between attributes in the databases to

build a correct classifier. The main contribution of the present study to attain high calculation accuracy for early diagnoses of heart diseases. The proposed hybrid associative classification is implemented on spicy environment. Finally, a skilled system is developed for the end user to check the risk of heart diseases on the basis of assumed parameters and the best associative classification method. The experimental results show that large number of the rules support to the better determines of heart diseases that even support the heart professional in their diagnosis in decisions [2].

XI. Acknowledgement

I sincerely thank Pro. Dr. Toby D Hocking Northern Arizona University, School of Informatics, Computing, and Cyber Systems for helping me in all the research. He has provided his immense support and guidance throughout the project.

XII. Reference

[1] Ramalingam, V. (2018). Heart disease prediction using machine learning techniques: A survey. Emitter: International Journal of Engineering Technology., 7(2.8), 684.

[2] Sridhar, A. (2018). Predicting heart disease using machine learning algorithm. International Research Journal of Engineering and Technology (Online), 6(4), 36.

[3] Tahira Mahboob, Rida Irfan and Bazelah Ghaffar et al. "Evaluating Ensemble Prediction of Coronary Heart Disease using Receiver Operating Characteristics", 978-1-5090-4815-1/17/31.002017$IEEE$.

[4] A. U. Haq, J. P. Li, M. H. Memon, S. Nazir, and R. Sun, "A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms," Mobile Inf. Syst., vol. 2018, Dec. 2018, Art. no. 3860146.

[5] S.Rajathi and Dr.G.Radhamani et al. "Prediction and Analysis of Rheumatic Heart Disease using kNN Classification with ACO ", 2016.

[6] P. A. Heidenreich, J. G. Trogdon, O. A. Khavjou et al., "Forecasting the future of cardiovascular disease in the United States: a policy statement from the American Heart Association," Circulation, vol. 123, no. 8, pp. 933–944, 2011.

[7] S. Ghwanmeh, A. Mohammad, and A. Al-Ibrahim, "Innovative artificial neural networks-based decision support system for heart diseases diagnosis," Journal of Intelligent Learning Systems and Applications, vol. 5, no. 3, pp. 176–183, 2013.

[8] Kanika Pahwa and Ravinder Kumar et al. "Prediction of Heart Disease Using Hybrid Technique For Selecting Features", 2017 4th IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics (UPCON).

[9] Thomas, G. D., Ensemble methods in machine learning. In Proc. of the First International Workshop on Multiple Classifier System (MCS 2000), 1–15, 2000.

[10] Shan Xu ,Tiangang Zhu, Zhen Zang, Daoxian Wang, Junfeng Hu and Xiaohui Duan et al. "Cardiovascular Risk Prediction Method Based on CFS Subset Evaluation and Random Forest Classification Framework", 2017 IEEE 2nd International Conference on Big Data Analysis.

[11] Saba Bashir, Usman Qamar, M.Younus Javed et al. "An Ensemble based Decision Support Framework for Intelligent Heart Disease Diagnosis" International Conference on Information Society (i-Society 2014).

[12] H. Kahramanli and N. Allahverdi, "Design of a hybrid system for the diabetes and heart diseases," Expert Systems with Applications, vol. 35, no. 1-2, pp. 82–89, 2008.

[13] S. Palaniappan and R. Awang, "Intelligent heart disease prediction system using data mining techniques," in Proceedings of IEEE/ACS International Conference on Computer Systems and Applications (AICCSA 2008), pp. 108–115, Doha, Qatar, March-April 2008.