www.ijrream.in Impact Factor: 2.9463

INTERNATIONAL JOURNAL OF RESEARCH REVIEW IN ENGINEERING AND MANAGEMENT (IJRREM)

An Efficient Architecture for Predicting Cardiovascular Disease with Medical Diagnosis System using Convolutional Neural Networks

[1]M. Navya Kala, [2]S.Venkataramana
[1]PG Student, [2]Associate Professor,
Department of Information Technology, S.R.K.R Engineering College, Bhimavaram, India

Abstract

Healthcare is one of the mandatory tasks to be done in human life. Cardiovascular disease is a broad category for a range of diseases that are affecting heart and blood vessels. Cardiology related diseases are incorporated with n number symptoms and different medication. For instance, In India most of the medical experts are preferring allopathy and China is following both allopathy and acupressure medicine. Machine Learning is the powerful tool for medical Industry and decision making for the medical experts and researchers; it helps to maximize the hypothesis of the possibilities and minimize the disease factors. Many algorithms are promoting medical field but we preferred Convolutional Neural Networks for these kind of imaging applications. Decision making in the medical processing is the complexity for any kind of disease because so many factors are playing against the human disease. In this work, we analyzed the earlier datasets with useful information and classified image, video, text data and store in database. Each datasets are produced with valid parameterized results with the help of Spark ML. We proposed convolutional neural network, or CNN for short, is a specialized type of neural network model designed for working with two-dimensional image data, although they can be used with one-dimensional and three-dimensional data. A convolution is a linear operation that involves the multiplication of a set of weights with the input, much like a traditional neural network. Besides we achieved more than 90% accuracy for prediction and prevention from this major disease. Because 2 – 2.5 million people are affected by this disease.

Keywords: CNN, Cardiovascular, Health care

1. Introduction

1.1 Cardiovascular Disease

Cardiovascular disease (CVD) – which includes heart disease and stroke – is the number one cause of death globally. An estimated 17.3 million people died from CVD in 2008, affecting men and women almost equally, and representing 30% of all global deaths. Low- and middle-income countries are disproportionally affected with over 80% of global CVD deaths. Of the total CVD deaths, an estimated 7.3 million were due to coronary heart disease and 6.2 million were due to stroke. Annual CVD deaths are projected to rise to 23.3 million by 2030 (mainly from heart attacks and strokes) if current trends are allowed to continue. The leading risk factor for CVD is high blood pressure, also known as raised blood pressure or hypertension, with one in three adults being affected. It is often referred to as the "silent killer" as many people are not aware they have it, yet it causes 9.4 million deaths each year including 51% of deaths due to strokes and 45% of deaths due to coronary heart disease.

People at high CVD risk can be identified early in primary care settings and inexpensive treatment is available to prevent many heart attacks and strokes. Survivors of a heart attack or stroke are at high risk of recurrences and at high risk of dying from them. The risk of a recurrence or death can be substantially lowered with a combination of drugs – statins to lower cholesterol, drugs to lower blood pressure, and aspirin.

1.2 Machine Learning - Cardiovascular Disease

Machine learning is a sub discipline of AI and can be categorized into three types of learning, i.e., supervised, unsupervised, and by reinforcement. The learning curves and the area under the curve (AUC) are important considerations when choosing a machine learning algorithm, while C-statistic is important in the choice of traditional methods of data processing. As in traditional statistics, machine learning requires a sufficient data set for training (the sample size in traditional statistics), and there should be no lack of adjustments, i.e., underfitting and overfitting (and alfa should not be greater than 0.05 in traditional statistics)

In today's digital world, several clinical decision support systems on heart disease prediction have been developed by different scholars to simplify and ensure efficient diagnosis. This

paper investigates the state of the art of various clinical decision support systems for heart disease prediction, proposed by various researchers using data mining and machine learning techniques. Classification algorithms such as the Naïve Bayes (NB), Decision Tree (DT), and Artificial Neural Network (ANN) have been widely employed to predict heart diseases, where various accuracies were obtained. Hence, only a marginal success is achieved in the creation of such predictive models for heart disease patients therefore, there is need for more complex models that incorporate multiple geographically diverse data sources to increase the accuracy of predicting the early onset of the disease.

Recently, authors have successfully experimented with the newest and most innovative neural network (NN) models and, more specifically, machine and deep learning techniques, such as the convolutional neural networks (CNN) and audio biometrics techniques. CNN has been utilized in arrhythmia detection, coronary artery disease detection, and beats classification. A deep belief network has been used to classify signal quality in ECG. Some researchers have implemented 11-layer CNN to detect MI. The authors have demonstrated the use of a shallow convolutional neural network, only focusing on inferior myocardial infarction. This network benefits from the use of varying filter sizes in the same convolution layer, which allows it to learn features from signal regions of varying lengths

1.3 Machine Learning

Machine Learning is a field of computer science that uses statistical techniques to give computer systems the ability to learn that is progressively improve performance on a specific task with data, without being explicitly programmed. The name machine learning evolved from the study of pattern recognition and computational learning theory in artificial intelligence, machine learning explores the study and construction of algorithms that can learn from and make predictions on data such algorithms overcome following strictly static program instructions by making data-driven predictions or decisions, through building a model from sample inputs. Machine learning is employed in a range of computing tasks where designing and programming explicit algorithms with good performance is difficult or infeasible.

1.4 Problem Statement

Cardiovascular diseases are the life-threatening diseases of the present-day world. Cardiovascular diseases deal with the problems related to heart and blood vessels. Cardiovascular diseases are one of the diseases that account for the loss of millions of lives every year. Lack of early prediction is the primary reason for the loss of lives. Deep Learning holds great potential for healthcare industry to enable systems to systematically use data and analytics to identify inefficiencies and best practices that improve care and reduce costs. Deep learning provides an efficient way of diagnosing diseases and provides numerous approaches to discover the hidden patterns or similarities present in the data. The predicted results can be used to prevent and thus reduce cost for surgical treatment and other expensive. Many input attributes can be taken but our goal is to predict with few attributes and faster efficiency in finding the risk of having heart disease. Our proposed model helps in predicting the results at an early stage more efficiently and Accurately.

2. Literature Review

Dinesh Kumar G [3] had discussed data pre-processing techniques like the removal of noisy data, removal of missing data, filling default values if applicable and classification of attributes for prediction and decision making at different levels. This is done by comparing the accuracies of applying rules to the individual results of Support Vector Machine, Gradient Boosting, Random forest, Naive Bayes classifier and logistic regression on the dataset taken in a region to present an accurate model of predicting cardiovascular disease

T Rajesh, Meenasha [4] have discussed topic is about prediction of heart disease by processing patient's dataset and a data of patients to whom we need to predict the chance of occurrence of a heart disease. They had analyzed Data – preprocessing, ID3 algorithm, Naïve bayes classification and K – Means clustering. Their experimental results demonstrated that Naïve Bayes results and decision tree results may change so for every prediction we need not have a comparison of both the algorithms so get accurate results and in the same way if we use only a single algorithm which cannot pre-process data we even can't get good accuracy so its better to have combination of algorithms like k-means, ID3 and k-means and Naïve

Bayes. Besides during small datasets in some other cases most of time decision trees direct us to a solution which is not accurate, but when we look at Naïve Bayes results Divya Krishnan et. al [4] They proposed preprocessing extensive approach to predict Coronary Heart Diseases (CHD). The approach involves replacing null values, resampling, standardization, normalization, classification, and prediction. This work aims to predict the risk of CHD using machine learning algorithms like Random Forest, Decision Trees, and K-Nearest Neighbours. Also, a comparative study among these algorithms on the basis of prediction accuracy is performed. Further, *K*-fold Cross Validation is used to generate randomness in the data. These algorithms are experimented over "Framingham Heart Study" dataset, which is having 4240 records. In their experimental analysis, Random Forest, Decision Tree, and K- Nearest Neighbour achieved an accuracy of 96.8%, 92.7%, and 92.89% respectively.

3. Implementation

In Proposed System, we are applying CNN techniques (Hybrid) in identifying suitable treatments for heart disease patients. Based on this work, we identified gaps in the research on heart disease diagnosis and treatment and proposes a model to systematically close those gaps to discover if applying Machine Language (ML) techniques to heart disease treatment data can provide as reliable performance as that achieved in diagnosing heart disease.

Advantages:

- It improves the speed.
- It has higher performance.
- Data redundancy were rectified before tuning the input files like audio, video or text

The figure 3.1 depicts the way of organizing the medical data sets with CNN classification for getting the prediction of cardiovascular decease. In this dataset is consisting with text, image and video files and they will be preprocessed and fine-tuned for maximizing the accuracy. In this paper, we used four modules for obtaining the result of cardiovascular decease prediction using CNN.

1. Dataset
2. Data preprocessing
3. Algorithm implementation (CNN)

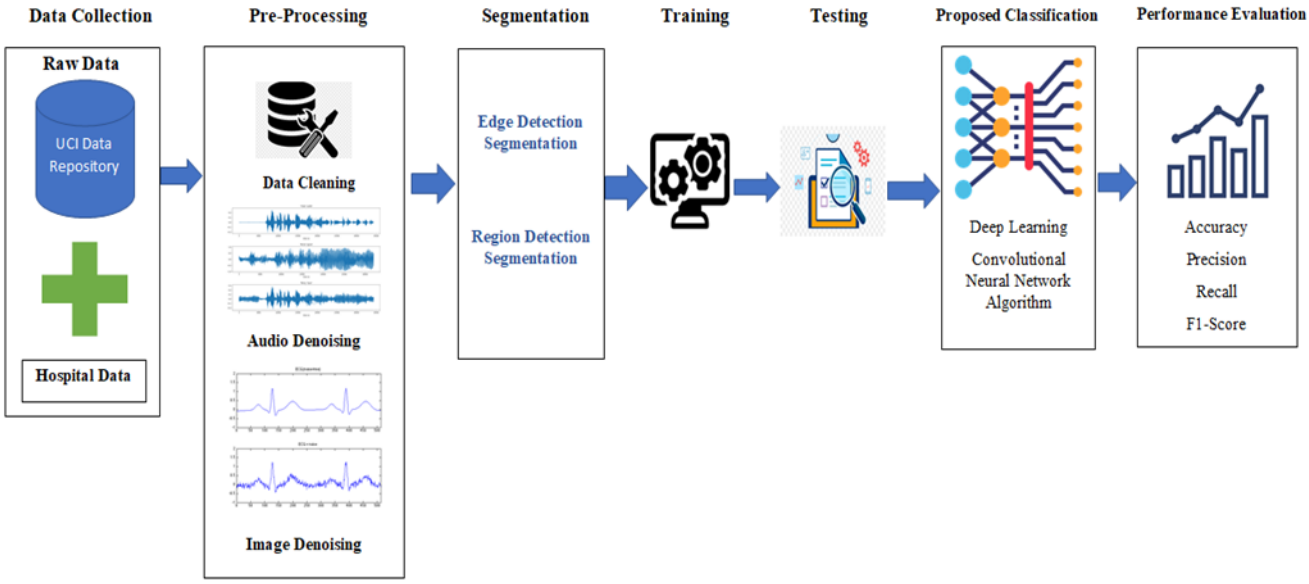

Figure 3.1 Proposed System Architecture

3.1.1 Dataset

Many techniques are adapted for predicting cardio vascular diseases. In this proposed work CNN is used to predict the risk of cardiovascular decease. In computer-aided heart disease diagnosis methods, where the data is obtained from some other sources and is evaluated by computer based applications. Computers have usually been used to build knowledge based clinical decision support systems which used the knowledge from medical experts, and transferring this knowledge into computer algorithms was done manually. This process is time consuming and really depends on the medical expert's opinion, which may be subjective. To handle this problem, machine learning techniques have been developed to gain knowledge automatically from examples or raw data. Medical diagnosis is an important but complicated task that should be performed accurately and efficiently and its automation would be very useful.

Data used for proposed system is obtained from UCI data repository. The data have been collected from 175 patients are used for proposed work. This database contains 76 attributes,

but experiment refers to use 16 of them. Cleaning and filtering of the data set is done to remove duplicate records, normalize the values, accounting for missing data and removing irrelevant data items.

Table – 3.1 Dataset Attributes

Name	Type	Description
Age	Continuous	Age in years
Sex	Discrete	1 = male 0 = female
Cp	Discrete	Chest pain type: 1 = typical angina 2 = atypical angina 3 = non-anginal pain 4 =asymptomatic
Trestbps	Continuous	Resting blood pressure (in mm Hg)
Chol	Continuous	Serum cholesterol in mg/dl
Fbs	Discrete	Fasting blood sugar > 120 mg/dl: 1 = true 0 = false
Restecg	Discrete	Resting electrocardiographic results: 0 = normal 1 = having ST-T wave abnormality 2 =showing probable or define left ventricular hypertrophy by Estes'criteria
Thalach	Continuous	Maximum heart rate achieved
Exang	Discrete	Exercise induced angina: 1 = yes 0 = no
Old peak ST	Continuous	Depression induced by exercise relative to rest
Slope	Discrete	The slope of the peak exercise segment : 1 = up sloping 2 = flat 3= down sloping
Ca	Discrete	Number of major vessels colored by fluoroscopy that ranged between 0 and 3.
Thal	Discrete	3 = normal 6= fixed defect 7= reversible defect
Diagnosis	Discrete	Diagnosis classes: 0 = healthy

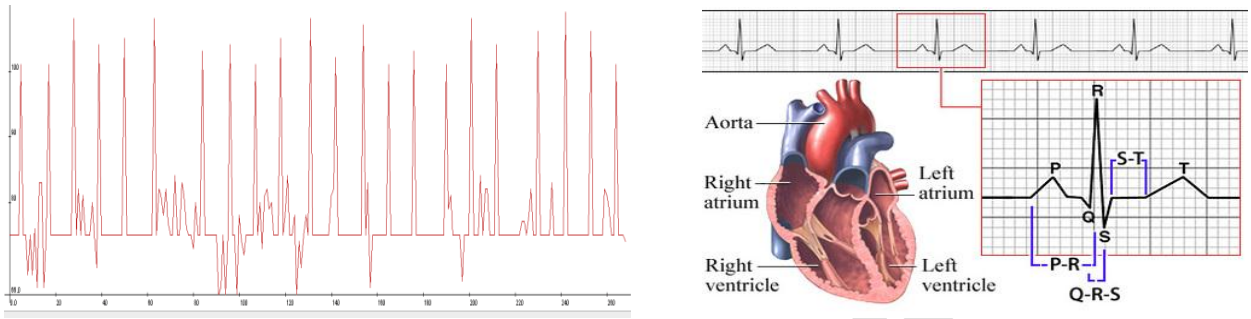

Figure 3.2 ECG Signals

3.2 Angiogram

An angiogram is an X-ray of the blood vessels. They can provide images of the blood vessels in many different organs. As a result, they often help doctors diagnose conditions affecting the heart, brain, arms, or legs. Angiograms can help doctors detect blood vessel abnormalities, including weakened blood vessels, plaque deposits, and blood clots [7].

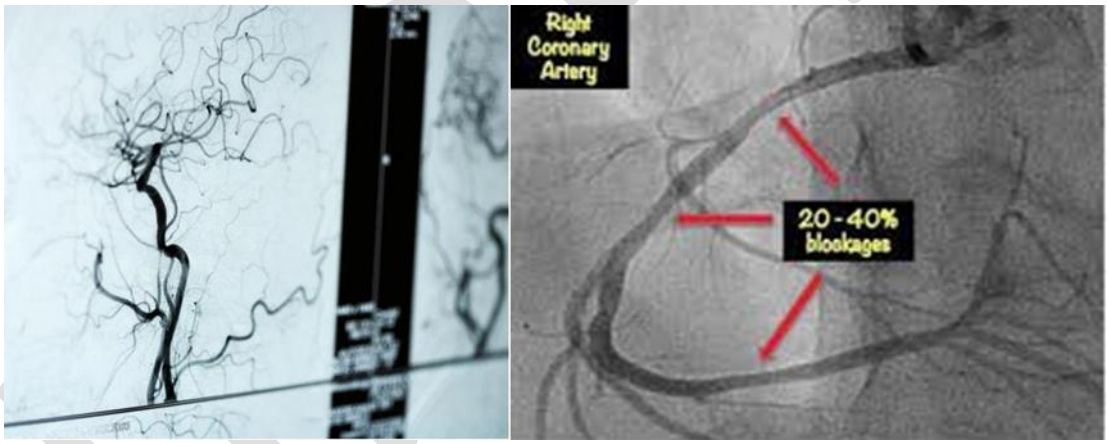

Figure 3.3 Angiogram Test result

After the dataset processing, medical experts suggest the following treatment

- Coronary artery bypass
- Balloon angioplasty (where a small balloon-like device is threaded through an artery to open the blockage)
- Valve repair and replacement
- Heart transplantation

- Artificial heart operations [8]

In our proposed system, we used some additional inputs which are not available in the existing model. We supposed to collect the physiological information about the patient with the following data sheet.

1	Name of the Patient
2	Symptoms
3	Food habit
4	Mentally Disturbed
5	Family Member's information about the patient

Table 3.2 Patient Data Sheet

3.3 Data Preprocessing

Any database is a collection of data objects. It can be also called them data samples, events, observations, or records. However, each of them is described with the help of different characteristics. In data science lingo, they are called attributes or features. Data preprocessing is a necessary step before building a model with these features.

- duplicates or semi-duplicates of the data records;
- data segments, which have no value for a particular research;
- Unnecessary information fields for each of the variables.
- Aggregation and normalization were implemented

Feature selection is the selection of variables in data that are the best predictors for the variable we want to predict [9]

3.4 Algorithm Implementation (CNN)

The figure depicts the various steps are involved in CNN for extracting the output of cardiovascular decease. It has majorly three layers; Convolution layer, pooling layer and fully connected layer. We used dataset with 16 parameters and personal data sheet from the patients are to be processed in the convolution layer. In this section, data will be trained and matched with the medical experts guidelines. This layer is confined with minimum data elements and processed for next layer. The convolution layer computes the output of neurons that are connected to local regions or receptive fields in the input, each computing a dot product between their weights and a small receptive field to which they are connected to in the input volume. Each computation leads to extraction of a feature map from the input image [10]

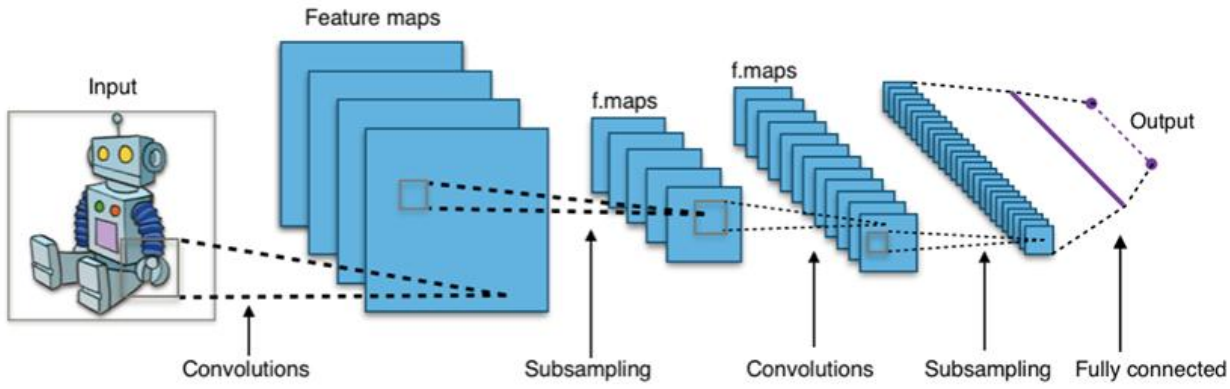

Figure 3.4 CNN implementation

4. Result and Discussion

4.1 For instance

```
plt.figure(figsize=[5,5])
# Display the first image in training data
plt.subplot(121)
plt.imshow(train_X[0,:,:], cmap='angiogram')
plt.title("Ground Truth : {}".format(train_Y[0]))
# Display the first image in testing data
plt.subplot(122)
```

```
plt.imshow(test_X[0,:,:], cmap='angiogram')
plt.title("Ground Truth : {}".format(test_Y[0]))
// For model creation we use the following implemetation
import keras
from keras.models import Sequential,Input,Model
from keras.layers import Dense, Dropout, Flatten
from keras.layers import Conv2D, MaxPooling2D
from keras.layers.normalization import BatchNormalization
from keras.layers.advanced_activations import LeakyReLU
batch_size = 32
epochs = 9
num_classes = 8
```

Next step, pooling layers provide an approach to down sampling feature maps by summarizing the presence of features in patches of the feature map. Two common pooling methods are average pooling and max pooling that summarizes the average presence of a feature and the most activated presence of a feature respectively. Eventually Fully connected layers are an essential component of Convolutional Neural Networks (CNNs), which have been proven very successful in recognizing and classifying images for computer vision. The CNN process begins with convolution and pooling, breaking down the image into features, and analyzing them independently. The result of this process feeds into a fully connected neural network structure that drives the final classification decision [11].

4.2 Cross Validation

It is a technique of validating a model with rotation. It is a method of evaluating a model for finding how the output of predictive analytics can be used to provide a generic solution for an unknown data set. So during training process the training data are divided in to k parts which are otherwise known as k fold. So out of k folds 1-fold used for validating the model and remaining k-1 fold are used for training the model. Here the value of k is taken as 10. During each time the fold which is used for evaluation changes. The accuracy of all the k-folds is noted and the average is calculated. The accuracy of all the 10- Folds are shown in the

following table-2 & Fig-3. After 10 fold cross validation are done the average training accuracy reported by the system as 96.3%.

4.3 Testing Accuracy

In this section the performance of the system is evaluated by using the test data set. Out of the total 59 records of test data 56 records are correctly predicted by the system which is given in the following confusion matrix table-3. The efficiency of the system is measured using its classification accuracy, precision and recall.

Table. 4.1 Cross Validation Score of 10 Folds

Fold No	**Accuracy**
1	92
2	93
3	94.8
4	95.4
5	96.8
6	96.3
7	97.4
8	98
9	98.6
10	98.4
Average Accuracy	**96.03**

Table 4.2 Confusion Matrix

Total Test Records	59
Correctly Predicted	56
TP (True Positive)	30
TN (True Negative)	26
FP (False Positive)	1
FN (False Negative)	2

4.4 Performance Evaluation

The accuracy of any classification model state how it is able to find any generalized solution to classify any unseen data in to any correct class. This knowledge of classification is obtained by the model during its training. Here the recital of the system is assessed by considering accuracy as a metric. In the following table no-4 the accuracy value of different literatures are compared with the accuracy of the proposed model. It is witnessed that the proposed model is having the best accuracy as compared to all the discussed literature by taking the same Cleveland data set which is shown in Fig-4. In addition to accuracy other performance metric like precision and recall percentage of the proposed method is also calculated for finding efficiency. The good percentage of recall indicates how the proposed model has identified the positive cases with less error. In the proposed model 13 features used. In comparison to this in some literatures less number of features are used. In fuzzy logic method 10 features are used, in artificial bee colony method 7 features are used, in principal component analysis method

7 features are used and in SVM Based integer-Coded Genetic Algorithm method 6 features are used. The main aim of using feature reduction is to remove the redundant and dependent feature in order to increase the accuracy. But in the proposed model no features are reduced. The proposed model gives the better accuracy without reducing any feature which is 94.91%. Hence it shows the efficiency of the model as compared to other literatures using feature reduction. In the literature, for diagnosing the coronary artery disease the researchers have used various traditional methods. These methods basically rely on Hypothesis to increase the accuracy level with huge data training to identify the pattern. The high accuracy is achieved using machine learning classification algorithms. But it has the biggest drawback like overfitting. The biggest advantage of the proposed system is that it is not only enhancing the classification accuracy of the system but also properly deal with overfitting. This makes the system more useful for the real-time environment.

Table 4.4 Comparison Table

Model Name	**Accuracy Obtained**
Random Forest(RF)	85.81
Artificial Bee Colony(ABC)	86.76
Ensemble-based Decision Support Framework(EDSF)	82
SVM Based Decision Support System with	72.55
Proposed System	94.91

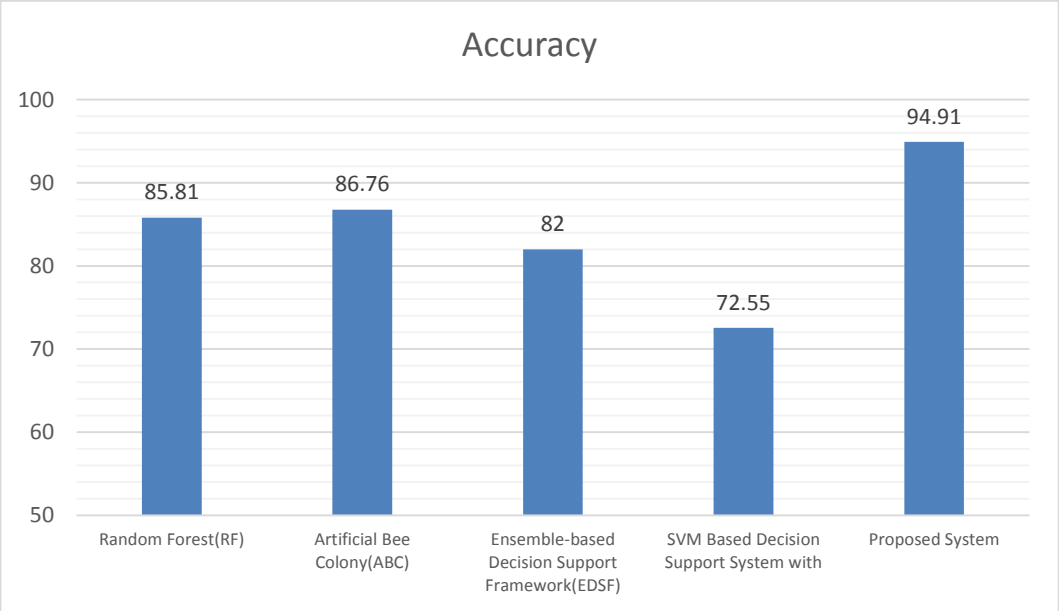

Figure 5.1 Chart

5. Conclusion

In this proposed, we proposed a new CNN-based heart disease prediction model. This model is evaluated on Cleveland dataset. The model is given 16 clinical features obtained from Cleveland dataset and personal data sheet from the patients as input. The training of the proposed model is done Convolutional Neural Network algorithm. The outcome of the model is whether heart disease is present or not with different degree of presence. To the best of our knowledge, none of the existing models is based on CNN. This project proposes an efficient

neural network with convolutional layers to classify significantly class-imbalanced clinical data.

5.1 Future Enhancements

For future endeavors, we propose to extend our algorithm to incorporate unstructured data as well. As of now, all attributes and laboratory tests considered have been approved by medical doctors.

References:

1. H. Chen, S.-Y. Huang, P.-S. Hong, C.-H. Cheng, and E.-J. Lin, "HDPS: Heart disease prediction system," in Computing in Cardiology, 2011, 2011, pp. 557-560.
2. S. U. Amin, K. Agarwal, and R. Beg, "Genetic neural network based data mining in prediction of heart disease using risk factors," in Information & Communication Technologies (ICT), 2013 IEEE Conference on, 2013, pp. 1227-1231.
3. J. S. Sonawane and D. Patil, "Prediction of heart disease using learning vector quantization algorithm," in IT in Business, Industry and Government (CSIBIG), 2014 Conference on, 2014, pp. 1-5.
4. S. Bashir, U. Qamar, and M. Y. Javed, "An ensemble based decision support framework for intelligent heart disease diagnosis," in Information Society (i-Society), 2014 International Conference on, 2014, pp. 259-264.
5. Ms. Rupali R. Patil, "Heart Disease Prediction System using Naive Bayes and Jelinek-mercer smoothing", International Journal of Advanced Research in Computer and Communication Engineering Vol. 3, Issue 5, May 2014, 6787
6. M. G. Feshki and O. S. Shijani, "Improving the heart disease diagnosis by evolutionary algorithm of PSO and Feed Forward Neural Network," in Artificial Intelligence and Robotics (IRANOPEN), 2016, 2016, pp. 48-53.
7. Gomathi K, ShanmugaPriyaa D. Multi disease prediction using data mining techniques. Int J Syst Softw Eng. 2016;4(2):12–4.

8. Miranda E, Irwansyah E, Amelga AY, Kom S, Maribondang MM, Kom S, Salim M, Kom S. Detection of cardiovascular disease risk's level for adults using Naive Bayes classifer. Healthc Inf Res. 2016;22(3):196–205.
9. Vincy Cherian, Bindu M.S., "Heart Disease Prediction Using Naïve Bayes Algorithm and Laplace Smoothing Technique", International Journal of Computer Science Trends and Technology (IJCST) – Volume 5 Issue 2, Mar – Apr 2017 Page 68
10. S.Nandhini, Monojit Debnath, Anurag Sharma, Pushkar, "Heart Disease Prediction using Machine Learning", .International Journal of Recent Engineering Research and Development (IJRERD) ISSN: 2455-8761 Volume 03 – Issue 10, October 2018, PP. 39-46
11. Dinesh Kumar G, Santhosh Kumar D, " Prediction of Cardiovascular Disease Using Machine Learning Algorithms", IEEE – 2018
12. T Rajesh, Meenasha," Prediction of Heart Disease Using Machine LearningAlgorithms", IJET 2018
13. Divya Krishnani, Anjali Kumari, Akash Dewangan, Aditya Singh," Prediction of Coronary Heart Disease using Supervised Machine Learning Algorithms", IEEE 2019
14. Liaqat Ali, Atiqur Rahman, Aurangzeb Khan, Mingyi Zhou, Ashir Javeed, Javed Ali Khan, "An Automated Diagnostic System for Heart Disease Prediction Based on x2 Statistical Model and Optimally Configured Deep Neural Network",DigitalObjectIdentifier10.1109/ACCESS.2019.2 904800.
15. Yogita Solanki1, Sanjiv Sharma "Analysis and Prediction of Heart Health using Deep Learning Approach",2019, International Journal of Computer Sciences and Engineering. Vol.7(8), Aug 2019, E-ISSN: 2347-2693.
16. Repaka, A. N., Ravikanti, S. D., & Franklin, R. G., "Design And Implementing Heart Disease Prediction Using Naives Bayesian", 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI).