



Prediction of Heart Disease using CNN

Shubhanshi Singhal¹, Harish Kumar², Vishal Passricha³

¹Assistant Professor, Computer Science and Engineering Department, TERII, Kurukshetra, Haryana, INDIA

²Assistant Professor, Department of Computer Engineering, YMCAUST, Faridabad, Haryana, INDIA

³Assistant Professor, Computer Engineering Department, NIT, Kurukshetra, Haryana, INDIA

Abstract: Heart disease is one of the deadly diseases. A large population in the world is suffering from this problem. As we consider death rate and a large number of people who are suffering from heart disease, it is revealed how important is early diagnosis of heart disease. There are many traditional methods of prediction for such illness but they are not looking sufficient. There is an urgent need of medical diagnosis system that can predict the heart diagnosis at an early stage and offers more accurate diagnosis than traditional methods like Logistic Regularization, Lasso, Elastic Net, and Group Lasso regularization. Machine learning approaches are gaining much popularity nowadays. In this paper, Convolutional Neural Networks (CNNs) are used to design an early stage prediction and medical diagnosis system. 13 clinical features are supplied as input to CNN. Modified backpropagation training method is used to train the CNN. During testing, it is observed that CNN offers more than 95% accurate results by predicting absence and presence of heart disease.

Keywords: Cleveland Dataset, Convolutional Neural Networks, Heart Disease, Modified Backpropagation.

I. Introduction

According to statistics available by the World Health Organization (WHO) [1], heart disease is the most killing disease. Billions of people around the world are suffering from heart disease and 12 million people are dying every year out of them. The main reason for this large number of death is that the problem is not identified at an early stage [2]. If the heart disease is predicted at an early stage, then we can prevent the death of lots of patients. Early prediction of heart disease always plays an important role in the control and diagnosis. By early prediction, an efficient and more accurate treatment may be offered to the patient. Therefore, a need to develop such an early prediction and medical diagnosis system arises day by day. The important key points of such a system should be that it must offer high accuracy with less operating cost. Developing such a system using deep learning approaches for prediction of heart disease may provide more accurate diagnosis than traditional methods on lesser cost. In this paper, CNN-based heart disease prediction model is proposed for an automated medical diagnosis. This model is trained using a modified backpropagation training method and used for early prediction. The model's performance is evaluated on Cleveland dataset and it offers 95% accurate results. The rest of this paper is organized as follows: section 2 covers the insight studies on heart disease i.e. related work. In section 3, we explain the CNN and modified backpropagation training method. Section 4 presents the proposed work, data source, the experimental setup used, and results obtained.

II. Related Work

Many researchers are working in the field of heart disease prediction. Shen et al. [3] initially, proposed a self-applied questionnaire (SAQ) based study to predict heart disease. This study is based on the analysis of the common risk features of the disease and other data collected in SAQ. Dundee rank factor score is used to validate their study. This study is based on statistically 3 risk factors (blood pressure, smoking, and blood cholesterol) together with sex and age to determine the risk of having heart disease. This study is implemented using multi-layered feed-forward neural network trained with the backpropagation method. This model offers 98% of relative improvement over relative operating characteristics.

Chen et al. [4] proposed an idea of heart disease prediction. They used learning Vector Quantization which is one of the artificial intelligence technique for classification and prediction purpose. Initially, 13 clinical features i.e. age, cholesterol, chest pain type, exercise, induced angina, max heart rate, fasting blood sugar, number of vessels colored, old peak, resting ECG, sex, slope, thal, and trestbps are identified for prediction. Secondly, Artificial Neural Network (ANN) is applied for classification purpose. Finally, training of neural networks are performed using backpropagation to evaluate the heart disease prediction system. Nearly 80% accuracy is achieved on testing set for heart disease prediction by the given system.

Amin et al. [5] proposed a hybrid system based on genetic algorithm and ANN. This system predicts the heart disease based on the risk factor. They trained ANN in a similar way i.e. using backpropagation as trained previously. They mainly highlight the two major disadvantages of the backpropagation algorithm. First one is that finding out the initial weights which are globally optimized is almost impossible. Second, backpropagation is slow in convergence. To solve these issues, they applied the genetic algorithm to optimize the weights of ANN and got better performance than simple ANN. They achieved an accuracy of 96.2% on the training set and the accuracy of 89% on the testing set.

Sonawane and Patil [6] proposed another methodology based on ANN to predict heart disease. In this, the ANN is trained by Vector Quantization algorithm using random order incremental training. Total 13 neurons are used in input layer to take the clinical features of heart disease dataset. To obtain less error and high accuracy, only a single neuron is taken as an output layer which shows that heart disease is present or not. The system performance is improved by training the network through a high number of epochs. 85.55% accuracy is achieved by this prediction model.

Bashir et al. [7] proposed a hybrid machine learning model based on Decision Tree, Support Vector Machine, and Naïve Bayes. They used three different classifiers to obtain the majority voting scheme. This scheme works in two different steps. In the first step, three classifiers produce their decision separately. In the second step, these decisions are combined to acquire a new model based on the majority voting scheme. The accuracy obtained by this model is higher than previous models. The results are 74% sensitivity, 82% accuracy, and 93% specificity for predictively of the heart disease.

Feshki and Shijani [8] proposed one model on heart disease prediction by using feature selection and classification approach with a specific dataset. The proposed model works in three steps. In the first step, the dataset is partitioned into two subsets i.e. sick and healthy people dataset. In the second step, the subset having the highest accuracy is found using particle swarm optimization with feed-forward backpropagation algorithm as a classifier. In this approach, four different algorithms: C4.5, multilayer perceptron, sequential minimal optimization, and feed-forward backpropagation NN were used. Feature selection and backpropagation feed-forward NN with particle swarm optimization is found as the most efficient method. This method offers maximum 91.94% accurate results.

T. Mane [9] presented the big data based model. They utilized the Hadoop Map platform for the reduction of such big data. In the clustering, K-means [10] and a decision tree algorithm are used for data mining purpose. They performed the improvement with the help of ID3. This method is found effective and efficient in the prediction of heart disease as compared to other methods.

III. Research Methodology

This field is originally kindled by psychologists and neurobiologists who sought to develop and test computational analogy of neurons. In heart disease prediction system, disease risk factors are taken as input variables. ‘Disease present’ and ‘disease absent’ are the output variables. There are many methods to perform learning of model like supervised learning, unsupervised learning, and reinforcement learning etc. In this experiment, the dataset is already available. The supervised learning method is best when the dataset is available hence it is used for training purpose.

A. Convolutional Neural Networks

CNNs are an advanced version of deep neural networks having spectral layers to learn lower and higher level features [11]. CNNs are effective models for predicting the statistics, modeling etc. Simply, three extra concepts i.e. local filters, max-pooling, and weight sharing make it powerful over DNNs. The architecture of CNN used for predicting heart disease is illustrated in figure 1. The CNN consists of a few pairs of convolution and max-pooling layers. Pooling layer always follows the convolutional layer. Pooling is generally applied along frequency domain. The max-pooling offers good results for variability problem. A max-pooling layer takes the maximum filter activation from different positions within a specified window. By this step, it generates a lower resolution version of the convolution features. The max-pooling layer makes the architecture more tolerable to minor differences in positions of the object’s parts and leads to faster convergence. Finally, fully connected layers combine inputs from all the positions into a 1-D feature vector. Then the softmax activation function layer is used for the classification of the overall inputs.

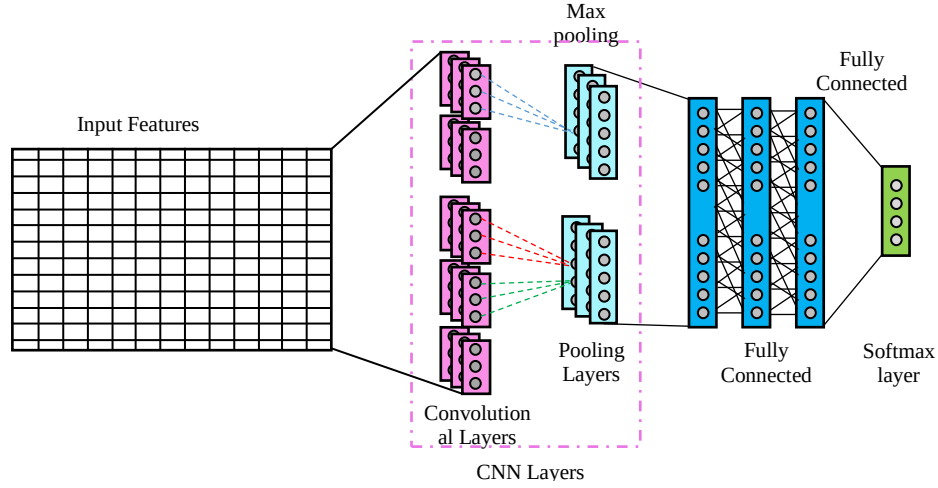


Figure 1: The architecture of CNN used for Prediction Model

B. Modified Backpropagation

Roughly speaking, a neural network is a set of connected input/output units in which each connection has a weight associated with it. During the learning phase, the network learns by adjusting the weights so as to be able to predict the correct class label of the input tuples. Backpropagation is a neural network learning algorithm. It is a fast matrix-based algorithm to compute the output from a neural network. CNN-based prediction model is trained using a modified backpropagation training method [12].

Output deviation of the k^{th} neuron in output layer O :

$$d(O_k^0) = y_k - t_k \quad (1)$$

Input deviation of the k^{th} neuron in the output layer:

$$d(I_k^0) = (y_k - t_k)\phi(v_k) = \phi(v_k)d(O_k^0) \quad (2)$$

Weight and bias variation of k^{th} neuron in the output O :

$$\Delta W_{k,x}^0 = d(I_k^0)y_{k,x} \quad (3)$$

$$\Delta Bias_k^0 = d(I_k^0) \quad (4)$$

Output bias of k^{th} neuron in hidden layer H :

$$d(O_k^H) = \sum_k^{i++} d(I_k^0)W_{i,k} \quad (5)$$

Input bias of k^{th} neuron in hidden layer H :

$$d(I_k^H) = \phi(v_k)d(O_k^H) \quad (6)$$

Weight and bias variation in row x , column y in the m^{th} feature pattern, a former layer in front of k neurons in hidden layer H .

$$\Delta W_{m,x,y}^{H,k} = d(I_k^H)y_{x,y}^m \quad (7)$$

$$\Delta Bias_k^H = d(I_k^H) \quad (8)$$

Output bias of row x , column y in m^{th} feature pattern, sub-sample layer S

$$d(O_{x,y}^{S,m}) = \sum_k^{i++} d(I_{m,x,y}^H)W_{m,x,y}^{H,k} \quad (9)$$

Input bias of row x , column y in m^{th} feature pattern, sub-sample layer S

$$d(I_{x,y}^{S,m}) = \phi(v_k)d(O_{x,y}^{S,m}) \quad (10)$$

Weight and bias variation of row x , column y in m^{th} feature pattern, sub sample layer S

$$\Delta W_{x,y}^{S,m} = \sum_{x=0}^{fh} \sum_{y=0}^{fw} d(I_{\lfloor \frac{x}{2} \rfloor, \lfloor \frac{y}{2} \rfloor}^{S,m}) O_{x,y}^{C,m} \quad (11)$$

Among them, C represents convolution layer

$$\Delta Bias_{x,y}^{S,m} = \sum_{x=0}^{fh} \sum_{y=0}^{fw} d(O_{x,y}^{S,m}) \quad (12)$$

Output bias of row x , column y in k^{th} feature pattern, convolution layer C

$$d(O_{x,y}^{C,k}) = d(I_{\lfloor \frac{x}{2} \rfloor, \lfloor \frac{y}{2} \rfloor}^{S,m}) W^k \quad (13)$$

Input bias of row x , column y in k^{th} feature pattern, convolution layer C

$$d(I_{x,y}^{C,k}) = \phi(v_k)d(O_{x,y}^{C,k}) \quad (14)$$

Weight variation of row r , column c in m^{th} convolution core, corresponding to k^{th} feature pattern in l^{th} layer, convolution C .

$$\Delta W_{r,c}^{k,m} = \sum_{x=0}^{fh} \sum_{y=0}^{fw} d(I_{x,y}^{C,k}) O_{x+r,y+c}^{i-1,m} \quad (15)$$

Total bias variation of the convolution core

$$\Delta Bias_{r,c}^{C,k} = \sum_{x=0}^{fh} \sum_{y=0}^{fw} d(I_{x,y}^{C,k}) \quad (16)$$

IV. Proposed Model

Heart Disease prediction system is created using the convolutional neural networks. The main benefit of CNN is that it has high discriminative power. The flowchart of the proposed model is shown in figure 2.

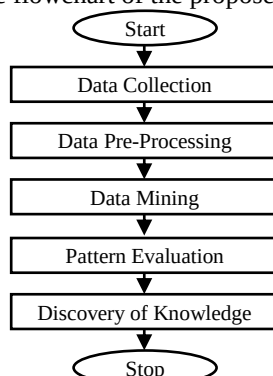


Figure 2: Flowchart of CNN based heart disease prediction system

- i) Data Collection: in this phase, the available dataset is supplied as input to the proposed model.
- ii) Data Pre-Processing: This phase mainly concentrates on noise reduction and feature selection. Feature selection is the process in which useful attributes are taken for the training purpose. In this model, the values of 13 different clinical attributes that are selected among the 78 available attributes are supplied to first convolutional layer.
- iii) Data Mining: This phase is very important for any system. In this, CNN is applied to the dataset.
- iv) Pattern Evaluation: In this phase, the performance analysis is performed to evaluate the results.
- v) Discovery of Knowledge: In this, we get the level of the heart disease problem.

A. Data Source

Cleveland dataset is widely used dataset for heart disease prediction system [13]. It is generally used by machine learning researchers. This dataset contains 303 instances and 76 attributes, but only 14 of them are referred in all published studies. The 'Num' field which has varying values from 0 – 4 (like absence to presence with varying degree) is taken as output attribute.

The dataset has some missing values in it. First, missing values are filled with interpolation values. This dataset is split into two parts: one for training (85%) and second for validation purpose (15%). It has total of 258 instances and 13 attributes for training purpose and validation set has 45 instances and 13 attributes. 14 attributes are used in this experiment. 13 out of 14 are taken as input parameters and remaining one i.e. num is taken as an output parameter for the network. Num is used to predict the output value and its value varies between 0 and 4. Here, 0 represents the absence of disease and rest values shows the various degrees of the presence of disease level. The description of the attributes is given below:

Table 1: Attributes used in the experiment

Clinical Features	Description
Age	Age
Ca	Number of major vessels (0-3) colored by fluoroscopy
Chol (mg/dl)	Serum Cholesterol
Cp	Chest Pain type
Exang	Exercise-induced angina
Fbs	Fasting blood sugar
Num	Diagnosis of heart disease
Oldpeak	ST depression induced by exercise relative to rest
Restecg	Resting electrocardiographic results
Sex	Gender
Slope	The slope of the peak exercise ST segment
Thal	3=normal; 6=fixed defect; 7= reversible defect
Thalach	Maximum heart rate achieved
Trestbps (mmHg)	Resting Blood Pressure

B. Performance Evaluation Criteria

Four different metrics i.e. accuracy, precision, recall, and F1-measures are used to evaluate the performance of the proposed system. First, we denote TP, FP, TN, and FN as true positive (the number of instances correctly

predicted as required), false positive (the number of instances incorrectly predicted as required), true negative (the number of instances correctly predicted as not required) and, false negative (the number of instances incorrectly predicted as not required) respectively. Then, we can obtain four metrics: accuracy, precision, recall, and F1-measure as follows:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1 - Measure = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

Accuracy, Precision, Recall, and F1-Measure are used to express the success of predicting heart disease system. Using only accuracy can be sometimes misleading. Sometimes selecting the model which has lower accuracy is desirable, because it provides more robust predictor for the problem. All predictions can be predicted as the value of majority class by model when the problem domain has a large class imbalance. Therefore, we prefer the four different factor to get more accurate results.

C. Experimental Results

Initially, CNN is trained with 2 convolutional layers with 91 feature maps and 4 fully connected layers of 1024 hidden units each. The performance of the system is also evaluated for varying number of convolutional layer. Later, we increase the convolutional layers from 2 to 5 with the same number of feature maps to note the effects of the convolutional layer. Note that no change is made in fully connected layers. The best performance i.e. 95% accuracy is achieved for 3 convolutional layers.

Table 2: performance comparison for a various number of convolutional layers

Number of Convolutional Layers	Accuracy	Recall	Precession	F1-Measure
2	80	89.28	87.20	88.23
3	95	96.59	97.70	97.14
4	95	97.62	96.47	97.04
5	92	94.11	96.38	95.24

V. Conclusion

In this paper, we proposed a new CNN-based heart disease prediction model. This model is evaluated on Cleveland dataset. The model is given 13 clinical features obtained from Cleveland dataset as input. The training of the proposed model is done with modified backpropagation 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. It is the first attempt when CNN is applied in this kind of model. When compared to existing systems, the prediction accuracy of our prediction system is 9% relative higher than them.

References

- [1] W. H. Organization, "New initiative launched to tackle cardiovascular disease, the world's number one killer. 2016," *Photos by Tommy Trenchard, IntraHealth International*, 2017.
- [2] S. N. Blair, J. B. Kampert, H. W. Kohl, C. E. Barlow, C. A. Macera, R. S. Paffenbarger, et al., "Influences of cardiorespiratory fitness and other precursors on cardiovascular disease and all-cause mortality in men and women," *Jama*, vol. 276, pp. 205-210, 1996.
- [3] Z. Shen, M. Clarke, R. Jones, and T. Alberti, "Detecting the risk factors of coronary heart disease by use of neural networks," in *Engineering in Medicine and Biology Society, 1993. Proceedings of the 15th Annual International Conference of the IEEE*, 1993, pp. 277-278.
- [4] A. 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.
- [5] 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.
- [6] 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.
- [7] 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.
- [8] 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.
- [9] T. U. Mane, "Smart heart disease prediction system using Improved K-means and ID3 on big data," in *Data Management, Analytics and Innovation (ICDMAI), 2017 International Conference on*, 2017, pp. 239-245.
- [10] K. Wagstaff, C. Cardie, S. Rogers, and S. Schrödl, "Constrained k-means clustering with background knowledge," in *ICML*, 2001, pp. 577-584.
- [11] P. Sermanet, S. Chintala, and Y. LeCun, "Convolutional neural networks applied to house numbers digit classification," in *Pattern Recognition (ICPR), 2012 21st International Conference on*, 2012, pp. 3288-3291.
- [12] S. Abid, F. Fnaiech, and M. Najim, "A fast feedforward training algorithm using a modified form of the standard backpropagation algorithm," *IEEE Transactions on neural networks*, vol. 12, pp. 424-430, 2001.
- [13] A. Asuncion and D. Newman, "UCI machine learning repository," ed, 2007.