A Genetic Algorithm based Hybrid Deep Learning Approach for Heart Disease Prediction

Khushbu Verma[1*] • Ankit Singh Bartwal[2] • Mathura Prasad Thapliyal[1]

[1]*Department of Computer Science and Engineering, HNB Garhwal University (A Central University) Srinagar Garhwal, Uttarakhand, India-246174*
[2]*Department of Chemistry, HNB Garhwal University (A Central University) Srinagar Garhwal, Uttarakhand, India-246174*

*Corresponding author: khushbuverma422@gmail.com

Abstract: People nowadays suffer from a variety of heart ailments as a result of the environment and their lifestyle choices. As a result, analyzing sickness at an early stage becomes a critical responsibility. Data mining uses disease data to uncover important knowledge. In this research paper, we employ the hybrid combination of a Genetic Algorithm based Feature selection and Ensemble Deep Neural Network Model for Heart Disease prediction. In this algorithm, we used a 0.04 learning rate and Adam optimizer was used for enhancement of the proposed model. The proposed algorithm has come to 98% accuracy of heart disease prediction, which is higher than the past approaches. Other exist models such as random forest, logistic regression, support vector machine, Decision tree algorithms have taken a higher time and give less accuracy compare to the proposed hybrid deep learning-based approach.

Keywords: Heart disease; deep neural networks; genetic algorithms; data analysis; cardio disease analysis.

Introduction

Heart disease prediction is an important task nowadays as well as health challenges in the actual world for automated analysis of heart disease. Heart disease impairs blood vessel function and results in coronary artery infections, which weaken the patient's body, particularly in adults and the elderly. Heart illnesses are the leading causes of mortality in the world (Acharya, 2017; Oh et.al., 2018). As a result, early detection of heart disease is critical for properly treating cardiac patients before they have a heart attack or stroke. Artificial intelligence has made computers smarter and has given them the ability to reason (Yildirim et.al., 2019; Sharma, 2017). Different experts believe that insight cannot be gained without learning. These learning are used to quickly categorize large amounts of data for quick categorization of huge data and accurate illness analysis. Because medical data is growing at an exponential rate, using it to predict the correct disease is becoming an important (Acharya, 2011). This paper, use a hybrid method of GA and DNN to perform data analysis and prediction of heart diseases. We performed data pre-processing and test with our Hybrid Approach of Neural Network and GA based heart disease analysis model and our approaches perform better results in the comparison of all the past approaches. In the present study we have discussed about the algorithm for the prediction of heart disease by using hybrid deep learning approach, proposed and operation of ensemble Deep Neural

Network (DNN), and dataset pre-processing study.

Literature Review

Authors introduced a risk analysis method (Chen et.al., 2017) by combining structured and unstructured hospital data. They devised a disease analysis system for a variety of areas. They looked at diabetes, cerebral infractions, and heart illness to see if they may anticipate illness. On structured data, the illness analysis is made. The Decision Tree method outperforms both the Nave Bayes and the KNN algorithms in terms of results. The accuracy of CNN-based unimodal illness risk analysis is compared to the accuracy of CNN-based multimodal disease risk analysis. This research considers both organized and unstructured data in its analysis. The author had to work only dealt with structured data, and none of the authors had dealt with unstructured data. However, this paper relies on both organized and unstructured data with the use of the patient's EHR data (Qian et.al., 2015) created a disease risk analysis system. The active patient risk model was constructed as a result of this. The Alzheimer's disease is used for that active risk analysis method. Based on Wearable sensor data of the 2.0 system presented by (Zhang et.al., 2017). It was possible to capture the patient's physiological status using this fabric. This information is also utilized for analysis. The problems that were encountered when designing the wearable 2.0 architecture were discussed. Many uses are explored in this,

including chronic illness monitoring and emotional care, among others. (Kunjir et.al., 2017) created cloud-based health–Cps system to handle massive amounts of biological data. The health-Cps system is developed in such a way that two technologies are preferred: cloud and big data. This system carried out a variety of tasks related to cloud-like data processing, monitoring, and analysis. This system also provides a variety of healthcare services. (Qiu et.al., 2016) the author of this study proposed a technological advancement in the health system, which is primarily focused on data exchange across all telehealth providers via the cloud. This paper proposes a cloud-based data sharing solution for enhanced data sharing using data sharing ideas. The best telehealth sharing model is shown here. This author devised a novel optimum massive data sharing method for this purpose. By using this approach, users can acquire the best answers for processing biological data. (Mendis et.al., 2011) suggested the best, that detect illness based on patient history. Multiple illnesses were anticipated, as well as an undiscovered pattern of patient state. They developed the most effective illness analysis model based on health data. In addition, the notion of numerous illnesses was discovered with some unknown pattern. (Leoni et.al., 2017) compared machine learning algorithms in terms of classification accuracy in a liver illness dataset and achieved 91 percent accuracy. The authors have utilized in a variety of application areas and has proven to be capable of doing classification efficiently

and with high results. (Sunny et.al., 2018) described particularly beneficial for illness detection on the supplied data set, according to the authors. Small amounts of data, such as symptoms or prior information gleaned from a medical diagnostic, are utilized to make an analysis. The limitations of this work were that they were unable to evaluate a huge dataset; as medical data continues to expand, it is necessary to categorize it, which is a difficult task. (Vinitha et.al., 2018) suggested a CNN-MDRP method for illness analysis using heart data from a medical examination. In this paper, we perform data analysis of heart patients using GA and neural networks. (Chen et.al., 2017) compare and outperform other machine learning algorithms in terms of classification accuracy in a liver illness dataset by 91 percent. The author has utilized in a variety of application areas and has proven to be capable of doing classification efficiently and with high results. (Amin et.al., 2013) described particularly beneficial for illness detection on the supplied data set, according to the authors. Small amounts of data, such as symptoms or prior information gleaned from a medical diagnostic, are utilized to make an analysis. (Palaniappan et.al., 2008) suggested a CNN-MDRP method for illness analysis using heart data from a medical examination. In this paper, we perform data analysis of heart patients using GA and neural networks.

Dataset

For analysis and monitoring system of heart diseases, we prepare heart diseases dataset that

taken raw form on the UCI machine learning. This heart disease dataset contain 54 features attribute and 900 patient records, which also consist of target attribute and rest of all have feature attribute. The target attribute has two numeric value 0 and 1, 0 means have not heart disease and 1 means have heart disease. First of all, clean it for analysis at each stage (Heart diseases dataset; symptoms).

Missing-data filtering

Wearable sensor data and data retrieved from EMRs both contain worthless and inaccurate information. Noisy wearable sensor of heart disease data produced an inaccurate result (Acharya, 2018). Furthermore, when data retrieved from EMRs does not contain at least one value, it is assumed that the data is missing. Text mining techniques may have failed to recognise a FRF value, or FRF values may not have been recorded, resulting in information being lost in the extracted data. We use a well-known filtering technique called Kalman filtering technique to clean the data. This filter eliminates noise, duplicate records, and inconsistencies from the data. The parameters X, ct_j, i, $X_i^{ct_j}$, and $Mean\ (X^{ct_j})$ in equation 1 shows features such that = {"age", "chol", "sex", "heart rate….}.

$$Mean\ (X^{ct_j}) = \frac{1}{n \sum X_i^{ct_j}} \qquad (1)$$

Feature selection

This section, Genetic Algorithm is used for feature selection depending on Many irrelevant characteristics are commonly included in patient data which reducing the

accuracy of heart disease forecast. It is critical to eliminate noisy data, pick important characteristics that aid in accurate findings, and decrease the dataset's complexity and dimensionality before using any analysis model. As a result, feature selection is a crucial step in improving data clarity in healthcare datasets. The genetic algorithms have fundamental functions such as inheritance and mutation, which use a variety of operators. According to the fitness norm, Selection operator chooses a member of an age group to play a part in the reproduction process. The most compatible members of the present age group are more likely to create the future population in this manner. The mutation operator is used to make modifications to a current-generation member's gene in order to generate a new generation member. After feature selection using genetic algorithm we have 20 important features that further used

for classification. There are 32 features have been reduced. Among of one have target class and rest of all have feature class.

Normalization

The heart illness dataset has a lot of characteristics, and each feature has a distinct numerical value, which makes the computing process more complicated. As a result, a normalisation approach is utilised to normalise datasets in the range of 0 to 1, as well as to reduce numerical complexity. Data normalisation may be done in a variety of ways. The well-known min-max normalisation approach is employed in the proposed system. Using the following equation, this technique maps a numerical value. The parameters $DV_{norm}, DV^{hd}, DV_{min}, DV_{max}$ in the following equation, are normalized data value, original data value, minimum data value and maximum data value, respectively.

$$DV_{norm} = \frac{DV^{hd} - DV_{min}}{DV_{max} - DV_{min}} * (new_{max} - new_{min}) + new\_\min \qquad (2)$$

Proposes Ensemble Deep Neural Network Model for Heart Disease Prediction

The connection between input and output variables determines the behaviour of the model. The inner layer of the neural network is the one that is not visible to the naked eye. Several hidden layers may exist. Inputs and their associations with other hidden layers affect how well a hidden layer performs. This research paper determines four numbers of

layers and 12 neurons in each layer except the output layer by trial and error. The output layer has only two nodes that produce a binary outcome of the proposed model. In comparison to the AdaBoost method, the boosting approach is better at managing noisy data, according to our testing and data analysis (Kumar, 2017). The input data, weight and bias are represented in the following equation.

$$Infonet_i = \sum_{i}^{n} x_i * w_{i,j} + b_j \qquad (3)$$

During the time of the training, the model modified weight by equation 4.

$$\text{Calculate } \Delta \qquad =- \alpha \qquad \text{and} \qquad (4)$$

-α represent learning rate and E represent the error rate which is represented by equation 5. Output layer has only two-node because we used binary classification for the detection of breast cancer disease. Equation 2 represents the difference between T and A.

$$E = \frac{1}{2} \sum_{s=1}^{x} \sum_{o=1}^{y} (T - A)^2 \qquad (5)$$

Error rate, number of samples, number of outputs, Target output, Actual output are all represented by E, S, O, T, A, respectively. The entire calculation was repeated until the high accuracy was not obtained as well as we got less difference between the actual output and target output. In this algorithm, we used a 0.04 learning rate and for the enhancement of the proposed model we have used Adam optimizer. After that, we obtained the accurate performance of the Proposed Deep Learning Model for Heart Disease prediction classifiers on the Heart Disease diseases dataset finding the best training iterations our model is built in such a manner that determines the precise number of iterations required to reach the maximum degree of accuracy (Sharma et.al., 2019; Bhurane et.al., 2019). Algorithm 1 briefly explains the working strategy of proposed deep learning model.

Algorithm 1. Ensemble Deep Neural Network Model for Heart Disease prediction.

Input: Number of Attribute A = a1, a2, a3, a4 ……. a20 for Heart Disease Prediction.
Output: The Diagnosis of Heart Disease Present or the absence of Heart Disease.
Begin
1.   For input all features of Heart Disease, do iterations.
2.   Computed the weighted sum and add bias in each.

$$\text{hidden layer node by DNN} = \sum_{k}^{m} X_k * W_{kl} + B_k$$

3.   Calculate $\Delta W_{kl} =- \alpha \frac{\partial E}{\partial W_{kl}}$ and $E = \frac{1}{2} \sum_{s=1}^{x} \sum_{o=1}^{y} (T - A)^2$
4.   Choose α and Update $W_{k,l}$
5.   Review the difference is as small as possible between T and A.
6.   **End for**
7.   To apply ReLU activation function to diagnosis heart disease.
8.   For enhance performance of algorithm use AdaBoost model.
If obtained high accuracy and less error than stop.

Results and Discussion

Table-1, Fig.1 and Fig.2 shows results of experiments, comparative analysis of algorithms and ROC graph of our proposed model and existed machine learning algorithms. The accuracy is inversely proportional to the error. We observe that, as the unpredictability of the dataset increases, the error rate decreases and the solution gets more accurate. When additional cases are reported and data from other sources become available, the network predictor quickly adapts the new situations with perfect accuracy.

Table-1: Experimental results of models

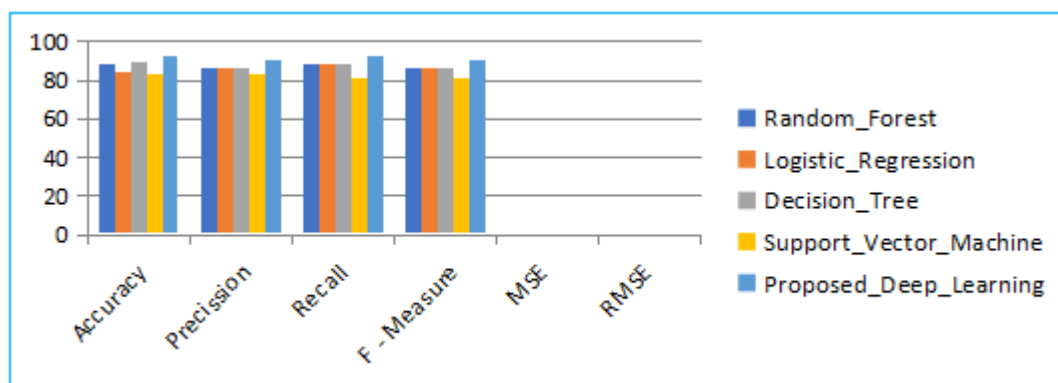|	Algorithms				
	Random Forest	Logistic Regression	Decision Tree	Support Vector Machine	Proposed Deep Learning
Accuracy	88	84	90	83	98
Precision	86	86	86	83	91
Recall	88	88	88	81	93
F-Measure	86	86	86	81	91
MSE	0.11	0.15	0.12	0.16	0.07
RMSE	0.34	0.39	0.32	0.43	0.26

Fig.1: Represent evaluation criteria of models.

The precision of the proposed data analysis is measured It is visible from Fig.1 that the proposed work has a low variation and a high level of precision. Precision can be seen as another metric to measure accuracy. Fig.1 represents that our proposed deep learning algorithm shows high accuracy as well another parameter of experimented outcome compares to existing machine learning algorithms. Table-2 show different exist models and our proposes Ensemble Deep Neural Network Model accuracy whereas we can clearly present that our model gives better result

among of other existing models for Heart Disease prediction.

High accuracy with the smallest error margin shows that the network predicted the importance of events very accurately, while medium accuracy with the smallest error margin suggests that the network properly predicted the form of the growth curve. From Fig.2, we can see that the proposed algorithm acquired high accuracy in the comparison of existing machine learning algorithms.

Table-2: A research that compares it to other recent studies.

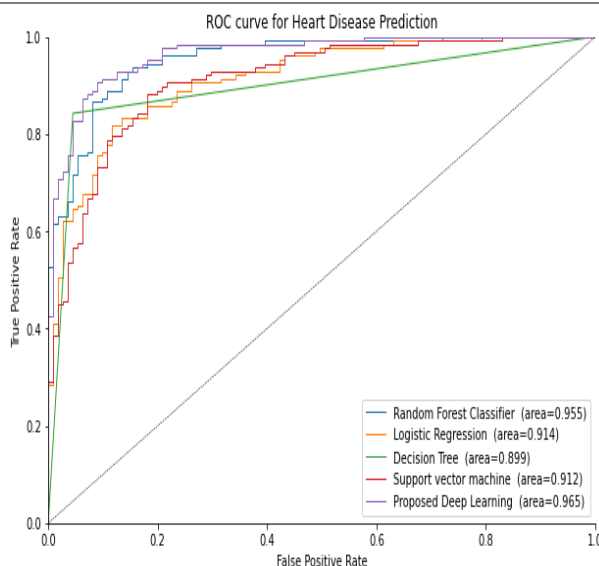| Year & Reference no.	Method	Accuracy
2017, (Chen et.al., 2017)	Convolutional neural network (CNN)-based model	94%
2017, (Sharmila et.al., 2017)	Fuzzy Neutral Network	91%
2018, (Vinitha et.al., 2018)	Random Forest algorithm is selected to develop CKDPS.	95%
2018, (Sunny et.al., 2018)	Logistic regression	97%
Proposed Algorithm	**Hybrid Approach of Neural Network and GA Model**	**98%**

Fig.2: Comparison of a ROC curves between different models for predicting heart Disease.

Conclusion

This paper proposes a genetic algorithm (GA) based feature selection and Deep Neural Network (DNN) based ensemble model for heart disease prediction of heart patients. The data is categorized into different groups, sex, heart rate and other symptoms. Based on the patient's symptoms, the study created a wide heart disease analysis. In this paper, a blend of Genetic Algorithm and ensemble model of

Neural Network. Exist machine learning methods is used for comparatively study of accuracy with our proposed model. The study identifies heart patients features in terms of age, sex, illness, blood pressure, heart rate, demography, and other disorders. In this paper, we conclude that our projected ensemble deep learning model achieved 98 % accuracy among existing machine learning algorithms.

Acknowledgement

Conflict of interest

There is no conflict of interest. This research work is not funded by any funding agencies.

References

Acharya UR, Fujita H, Sudarshan VK, Oh SL, Adam M, Tan, JH, ... & Chua KC (2017).Automated characterization of coronary artery disease, myocardial infarction, and congestive heart failure using contourlet and shearlet transforms of electrocardiogram signal;*Knowledge-Based Systems*, *132*: 156-166.

Acharya UR, Fujita H, Oh SL, Raghavendra U, Tan J, Adam M ... & Hagiwara Y (2018). Automated identification of shockable and non-shockable life-threatening ventricular arrhythmias using convolutional neural network. *Future Generation Computer Systems*, *79*: 952-959.

Acharya UR, Hagiwara Y, Koh JEW, Oh SL, Tan JH, Adam M & San-Tan R (2018). Entropies for automated detection of coronary artery disease using ECG signals: A review. *Biocybernetics and Biomedical Engineering*, *38*(2): 373-384.

Amin SU, Agarwal K & Beg R (2013, April). Genetic neural network-based data mining in prediction of heart disease using risk factors. In *2013 IEEE Conference on Information & Communication Technologies*, 1227-1231.

Bhurane AA, Sharma M, San-Tan R & Acharya UR (2019). An efficient detection of congestive heart failure using frequency localized filter banks for the diagnosis with ECG signals. *Cognitive Systems Research*, *55*: 82-94.

Chen M, Hao H, wang K, Wan L & Wang L (2017). Disease prediction by machine learning over big data from healthcare communities. *Ieee Access*, *5*: 8869-8879.

Chen M, Ma Y, Li Y, Wu D, Zhang Y & Youn CH (2017). Wearable 2.0: Enabling human-cloud integration in next generation healthcare systems. *IEEE Communications Magazine*, *55*(1): 54-61.

Disease and symptoms Dataset; www.github.com.

Heart disease Dataset; www.uci repository.com.

Kumar M, Pachori RB, & Acharya UR (2017). Characterization of coronary artery disease using flexible analytic wavelet transform applied on ECG signals. *Biomedical signal processing and control*, *31*: 301-308.

Kunjir A, Sawant H & Shaikh NF (2017). Data mining and visualization for prediction of multiple diseases in healthcare. In *2017 International Conference on Big Data Analytics and Computational Intelligence (ICBDAC)*, 329-334.

Mendis S, Puska P, Norrving B & World Health Organization. (2011). *Global atlas on cardiovascular disease*

prevention and control. World Health Organization.

Nithya B & Ilango V (2017). Predictive analytics in health care using machine learning tools and techniques. In *2017 International Conference on Intelligent Computing and Control Systems (ICICCS),* 492-499.

Oh SL, Ng EY, San Tan R & Acharya, UR (2018). Automated diagnosis of arrhythmia using combination of CNN and LSTM techniques with variable length heart beats. *Computers in biology and medicine*, *102*: 278-287.

Palaniappan S & Awang R (2008). Intelligent heart disease prediction system using data mining techniques. In *2008 IEEE/ACS international conference on computer systems and applications*, 108-115.

Qian B, Wang X, CaN Li H & Jiang Y G (2015). A relative similarity-based method for interactive patient risk prediction. *Data Mining and Knowledge Discovery*, *29*(4): 1070-1093.

Qiu L, Gai K & Qiu M (2016). Optimal big data sharing approach for tele-health in cloud computing. In *2016 IEEE International Conference on Smart Cloud (Smart Cloud)*, 184-189).

Sharma M, & Acharya UR (2019). A new method to identify coronary artery disease with ECG signals and time-Frequency concentrated antisymmetric biorthogonal wavelet filter bank. *Pattern Recognition Letters*, *125*: 235-240.

Sharma M, San-Tan R & Acharya UR (2018). A novel automated diagnostic system for classification of myocardial infarction ECG signals using an optimal biorthogonal filter bank. *Computers in biology and medicine*, *102*: 341-356.

Sharmila SL, Dharuman, C & Venkatesan P (2017). Disease classification using machine learning algorithms-a comparative study. *International Journal of Pure and Applied Mathematics*, *114*(6): 1-10.

Sunny AD, Kulshreshtha S, Singh S, Srinabh BM & Sarojadevi H (2018). Disease diagnosis system by exploring machine learning algorithms. *Int. J. Innov. Eng. Technol*, *10*(2): 14-21.

Vinitha S, Sweetlin S, Vinusha H & Sajini S (2018). Disease prediction using machine learning over big data. *Computer Science & Engineering: An International Journal (CSEIJ)*, *8*(1): 1-8.

Yildirim O, Baloglu UB, Tan RS, Ciaccio EJ & Acharya UR (2019). A new approach for arrhythmia classification using deep coded features and LSTM networks. *Computer methods and programs in biomedicine*, *176*: 121-133.

Zhang Y, Qiu M, Tsai CW, Hassan MM & Alamri A (2015). Health-CPS: Healthcare cyber-physical system assisted by cloud and big data. *IEEE Systems Journal*, *11*(1): 88-95.