Turkish Online Journal of Qualitative Inquiry (TOJQI)
Volume 12, Issue 5, June 2021: 5070-5080

Research Article

Improving Heart Disease Prediction Accuracy Using Evolutionary Algorithms and CNN Models

R. Parthiban[1,] Dr. K. Santhosh Kumar[2]

Associate Professor[1], Department of CSE, IFET College of Engineering, Villupuram,
Assistant Professor [2], Department of IT, Annamalai University, Chidambaram,
parthineyveli@gmail.com[1], santhosh09539@gmail.com[2]

Abstract

Heart disease prediction is critical in the healthcare system due to the disease's high-risk factor. Data analysis is critical in predicting outcomes depending on medical history. Each factor must be taken into account to ensure that the prediction is accurate. Conventional methods rely on huge amounts of information rather than precise prediction. Data must be chosen carefully in order to achieve an earlier predictive process. If data collection is incomplete, analysis is harmed. As a result, the work proposed improving prediction accuracy by transforming missing information to indeed from the dataset. The method incorporates a CNN-MMEI classifier for previous accurate prediction (efficient multi - modality disease early identification utilizing convolutional neural networks (CNNs)) with an approach. Using data sets, the Nave Bayes algorithm is used to process the dataset. The suggested neuro-genetic approach identifies a feasible configuration for the optimal network. The results demonstrate that by combining an effective algorithm and a classifier, 85% percent accuracy is achieved. The performance indicators proposed will shed light on reliable variables.

Key terms: CNN-MMEI, Genetic algorithm

I. Introduction

A sudden growth in the fatality rate from heart disease has highlighted the importance of timely identification through effective methods. The disease is caused by a variety of factors, including stress, high cholesterol diets, and perhaps even biological predispositions [1]. Existing prediction methods rely on automatic diagnosis systems that are incapable of achieving a high degree of accuracy due to the inclusion of irrelevant features in the dataset. Numerous studies concentrate on feature preprocessing and feature selection. Early detection is critical because it affects a person gradually and he is unaware of the symptoms as they progress to a dangerous level[2]. A temperature and heart rate surveillance system were developed that notifies the patient if critical symptoms are predicted [3]. Deep learning has made a significant contribution to the medical field by enhancing the precision of disease prediction, detection, and diagnosis. It makes every effort to aid in early detection, but certain pitfalls are observed, resulting in irrelevant diagnosis. Automatic prediction is accomplished using structured data, which is data that is well-formed based on the data contained in a dataset. The ECG output, laboratory records, and hospital information from medical experts are all included in the prediction of heart disease. This work focuses on the data that is truly necessary for prognostication and eradicates irrelevant content via CNN-MDRP and the precision of selected features via an evolutionary algorithm. To extract essential patterns from multifaceted databases, an algorithm was developed that uses the Apriorism algorithm to determine the intensity and highest exchange length of a principles remain [4]. Previously published research used an evolutionary approach to select the most precise features for prediction. Numerous studies demonstrate that using an evolutionary algorithm to initialize the weights in an artificial neural network enables optimization and overcomes the disadvantages of slow convergence [5]. Unsupervised learning was implemented using an inter-genetic algorithm [6]. The obtained results by using a genetic algorithm to select features and maximizing the parameters haven't ever done nothing to improve the classification process's accuracy [7]. The limitations in predictive accuracy associated with traditional methodologies contribute very little to cardio disease as determined by a multidimensional database. As a result, we have been motivated to improve classification through the use of CNNs and evolutionary approach on large datasets.

II. Related Works

This section examines and characterizes the approaches developed for the detection of chronic heart disease and discusses the impact of various optimization techniques that excel at data mining. The Top K High-Utility pattern mining trend is used to determine the regularity of frequently used features. This is accomplished through the use of a tree-based data structure called the RP-Tree [8].

Extreme Machine Learning on the basis of fuzzy logic was developed for the purpose of analyzing risk factors for heart disease [9]. The Particle Swarm optimization algorithm is used to select features for risk analysis in this case. The results demonstrate that now the prognosis method is the most advanced in recent years. It was stimulated using MATLAB and the data set was obtained from the Alizadeh Sani repository.

We provided an experimental review of different traditional clustering approaches used in data mining techniques, which aided in extracting only the most important features from large datasets [10]. The extraction procedure entails extracting only precise features in the data and converting them to consequential data for subsequent data retrieval. Classification and clustering are the two primary data mining approaches, with data retrieval being facilitated by neural networks, decision trees, and Bayesian networks. A survey of supervised and semi-supervised detection techniques in real-time synthetic data for both category-based and unlabeled datasets.

Information clustering in ambiguous streaming data is a difficult concept to grasp when it comes to relational data. An Adaptive Connection-based Clustering (ACCA) approach was proposed in which conventional matrix formations were defined and matrix formations were highlighted based on their similarity to various clusters via various attributes [11]. A connection-based algorithm was used to determine which assessments are comparable from categorical data in order to generate final clusters based on similar attributes.

In the context of heart disease prediction, knowledge discovery and data analysis are investigated in a variety of ways. It’s because each practitioner has a number of unique features that must be taken into account when making a guess. For data that are uncertain in their indexing approach, the attributes must be organized in a systematic manner. A novel Fuzzy Partitioned Genetic algorithm (NFPGA) has been developed specifically for data with uncertain categories [12]. To begin, the splitting data set with the greatest number of clusters is combined. The procedure is iterated until the clusters match the predefined clusters that are relevant to the dataset. The UCI repository dataset was used to generate novel fitness functions; parallel partitioning was used to evaluate cross-over and mutation operations.

III. PROPOSED SYSTEM

This section explains a CNN-based algorithm for detecting unimodal disease risk. The activity helps determine whether an individual is beginning to experience vital clinical symptoms based on information contained in their medical record. The input contains critical factors such as gender, indication factor, and additional information in the form of (f1,f2,f3,...,fn). Another attribute that indicates the degree of a risk factor's intensity is R0-high risk and R1-minimal risk. The following section discusses the range of data and its characteristics as determined by experts. The data can be structured in a particular way or be simply textual information. Textual data appears to be an unstructured collection of data. The task of identifying important features for prognostication is complicated in this case. Un - organized data are combined to create a multidimensional database that determines the risk factor for each individual in the dataset. The unimodal illness risk prediction algorithm is based on the patient's textual data. Generally, both structured and unstructured data are insufficient for examining disease prediction. While the preceding process consumes time, we overcome this limitation by leveraging the CNN-MMEI algorithm. A neural network algorithm can be used to predict risk factors for heart disease using both strictly regimented and non - structured textual data. The use of convolutional neural networks (CNNs) extracts the required information from the available data automatically. CNN has the possibility to extract information from massive medical records that is currently unavailable or inaccessible. For the purpose of controlling general illness, statistical knowledge is utilized. Healthcare experts who care for those patients provide organized data in a standardized format for reusability. These steps eliminate superfluous data and provide precise information for predicting critical importance [13, 14]. CNN is divided into phases, which are detailed below.

(i) Data Analysis

Data analysis is the process of having obtained information that is contained in textual data in an indirect manner. When data is replaced, the term "attribute unit" is used; when an aspect of data is replaced, the term "attribute assignment" is used. Collection of primary data, we must eliminate any ambiguous or irrelevant data from patient records. Through this data segregation process, uniqueness in prognostication is achieved. After acquiring an aggregate index, the dataset is classified as multidimensional. Let Rxy be a data dimension, with x representing the total amount of information required and y chosen to represent the distinct qualities of each patient.

(ii) Genetic approach

The section explains how to properly assess fitness in a multi-objective function for the purpose of selecting a feature subset using a genetic algorithm. The subsets are represented using binary strings, where value respondents agree the incorporation of a feature during network training and boolean value 0 symbolises the concealing of a characteristic during network training.

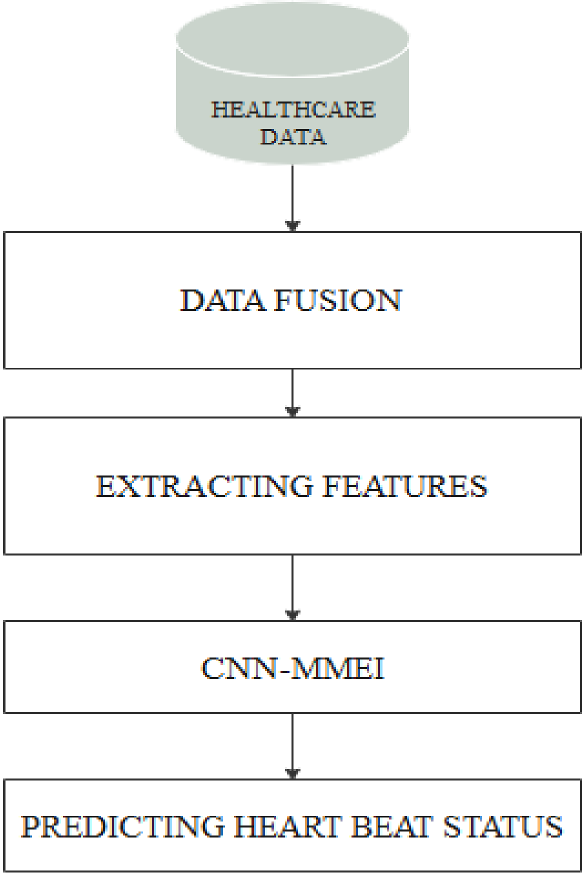

Fig 1: CNN-MMEI architecture

a. Optimization Function

Whereas the extrapolating the optimal feature from the subset, a CNN model is created for each subset for the purpose of assessing the fitness value. The CNN-MDRP has the same number of inputs as the binary string's number of 1's for representing a subset of features. During the CNN training process, each feature subset produces an error value F(b) for categorization of heart disease, where b is a binary string and 0≤F(b)≤1 is a zero. We think of the cost of having trained the features in the subset when formulating the optimization function. To compute the cost function, we assume that the cost of educating the feature is equal to one and is proportional to the amount of input neurons. C(b) of a binary string is calculated by dividing the number of ones in the string by the number of attributes. The cost function has a range of 0 to 1. By evaluating the cost function, one can formulate the erroneous value and optimization component as

$$O(b) = (2-F(b)) - \left(\frac{C(b)}{(2-F(b))}\right)\text{------}(1)$$

One such enhancement diminishes the convolutional neural network's error and cost of training. Through use of trivial solutions is avoided, and solutions with a high degree of accuracy and an acceptable cost are provided. To reduce the complexity of CNN-MDRP, we implement a neuro-genetic approach for feature optimization. The following section addresses the algorithm's steps. In broad sense, GA is an incremental technique that involves chromosomes to represent a sequence of genomes in order to find the optimal remedy in a computational complexity. The evolutionary algorithms are a complex search process that traverses a large search space in search of the optimal solution. The algorithm's steps are as follows:

(i) GA's chromosome contains candidate solutions.
(ii) The fitness function is used to determine the performance of each candidate.
(iii) GA operators are applied iteratively until an optimal solution is found.

The optimal parameter subsets are used to train the CNN for the purpose of predicting chronic heart disease. This hybrid methodology does not compensate for the lack of precision. The suggested framework begins by instantiating a random population using GA operators such as population size, generation number, crossover fraction, and mutation rate. The fitness of the optimization function is determined by the cost of having trained the deep neural network and the classification error rate. If the optimization algorithm is not optimal, a new number of criteria is used to generate a new generation. Iteration is used to obtain the optimal subset of features. The CNN is trained using a small set of data and its results are evaluated using the available dataset. For accurate prediction, the CNN model implemented a machine learning algorithm via multiple hidden layers.

IV. Results and Discussions

The information was analysed from 400 people who exhibit heart problems. Initially, the data is analysed to remove any missing or irrelevant type values. The table below lists the parameters for predicting chronic heart disease.

S.No	Attributes
1.	Age of the patient(in yrs)
2.	Sex(1=Female,0=Male)
3.	Chest pain categories: 1=atypical angina, 2=typical angina, 3=Non-angina pain, 4= asymptomatic
4.	BP in mm of Hg
5.	Cholesterol Serum(mg/dl)
6.	Fasting blood sugar>120mg/dl 1=True,0=false
7.	ECG results: 0=Normal; 1=ST elevation; 2= left ventricular Hypertrophy
8.	Maximum heart rate

9.	Exercise induced angina(Yes/No)
10.	ST depression induction through exercise
11.	Slope of peak exercise ST segment, 1= Sloping up, 2= flat,3=down sloping
12.	Number of vessels ranging from 0-3
13.	4=Normal, 6=fixed defect, 7= reversible defect
14.	Heart disease diagnosis Value 0: <50% diameter narrowing Value 1: >50% diameter narrowing

Table 1: Data set description

The data is derived from the Cleveland Clinic Foundation Database, which is instantly accessible. The performance of the proposed model is obtained by

computing the Sensitivity percentage (SE), Specificity Percentage (SP) and. Accuracy (AC). These validation parameters are defined as:

(i) $$Sensitivity = \frac{True\ positive}{True\ Positive + False\ Negative}$$

(ii) $$Specificity = \frac{True\ Negative}{True\ Negative + False\ Positive}$$

(iii) $$Accuracy = \frac{True\ positive + true\ negative}{True\ positive + True\ negative + False\ positive + False\ Negative}$$

Where True Positive is the number of attributes correctly classified as healthy, True Negatives is the number of subjects as abnormal; False Negatives is the number of subjects misclassified as abnormal when actually normal, and FP (False Positives) is the number of subjects misclassified as normal when actually abnormal.

Novel Pruning method[15]	68.47%
Kernel Difference weighted[16]	70.66%
SVM with Gaussian Kernel[17]	76.10%
Learning Vector Optimization[18]	76.92%
Fuzzy Weighted AIRS[19]	80.71%
NN with fuzzy membership function[20]	81.32%
MLP with two hidden layers[21]	85.55%
Neuro-Genetic approach[22]	89.58%
Evolutionary approach with CNN-MMEI	93.52%

TABLE 2: Comparison table with accuracy rate based on other methodologies in literature

The above table produces a clear vision on reduced complexity indicating highest accuracy rate compared to other methods in literature.

Performance Evaluation

By exploring datasets, furthermore, the influence of the training dataset on the validating and training accuracy, as well as the problem associated with it, is discussed in this section. It is put to the test by training a model with different training datasets containing 200, 400, 600, and 800 images, and then verifying the verification data set.

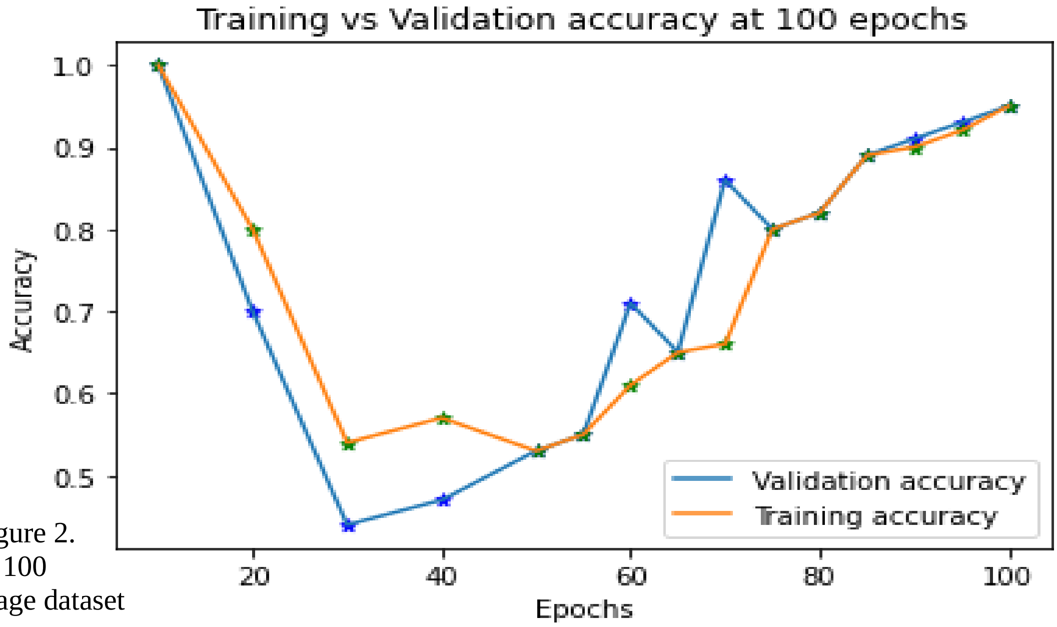

Figure 2. For 100 image dataset

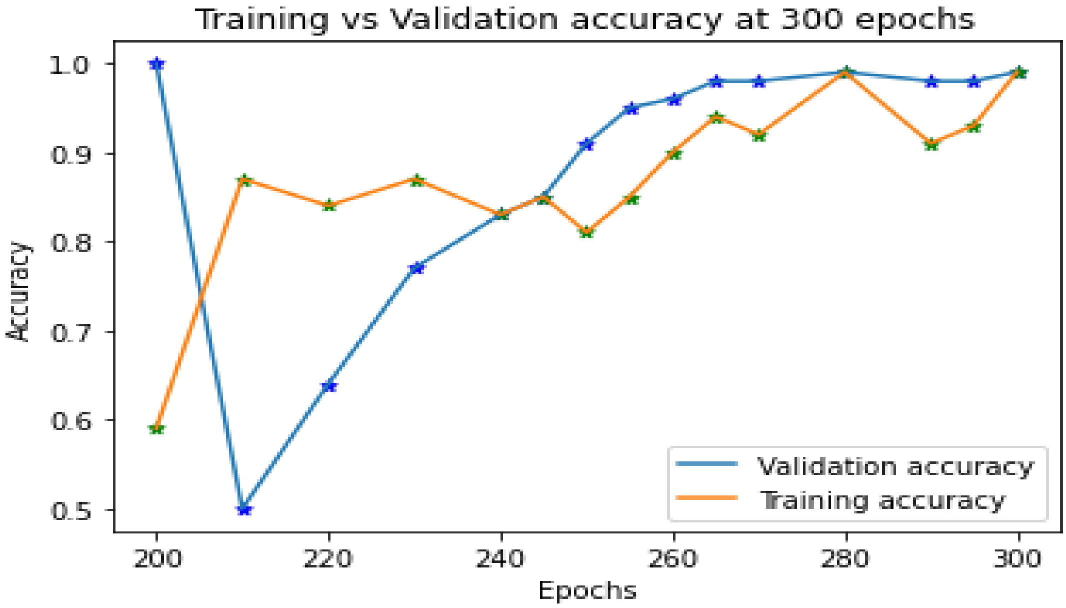

Figure 3. For 300 image dataset

As a result, as shown in figures 2 to 4, the train and test accuracy is displayed across the epochs. When the model has been trained with image database (Figure 2), the validation accuracy appears to be very poor, but it grows exponentially when the images is increased to 300 (Figure 3). Similarly, as the number of photos in the dataset grows from 500 to 700, the validation accuracy improves (Figure 4 and Figure 5).

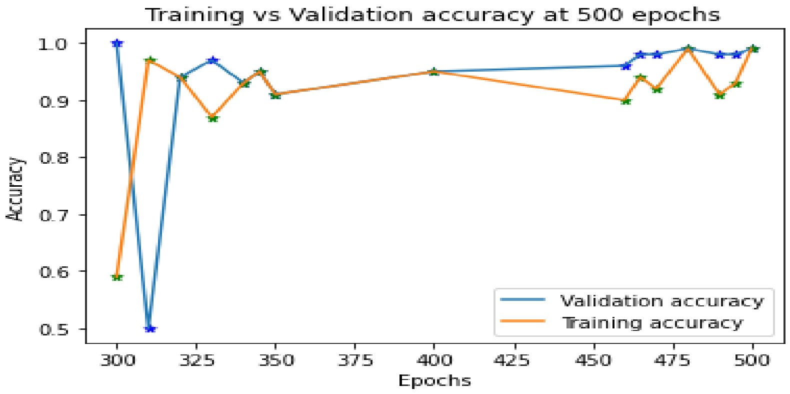

Figure 4. For 500 image dataset

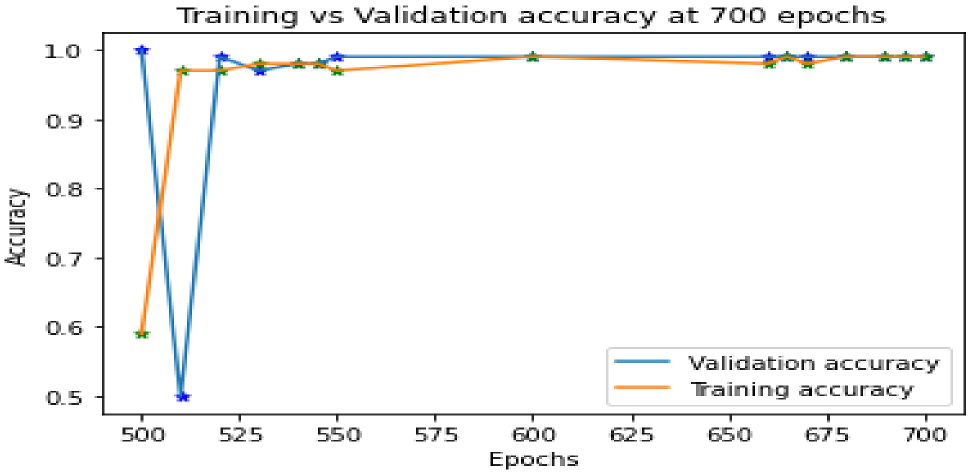

Figure 5. For 700 image dataset

It can be seen that if the model has been trained with 100 photos, the validation loss is quite high as illustrated in figure 6 , however when the model has been trained with 300 images, the validation loss is reduced as in figure b, but the training and validation losses are somewhat raised throughout epochs as shown in figure 7 ,8 and 9.

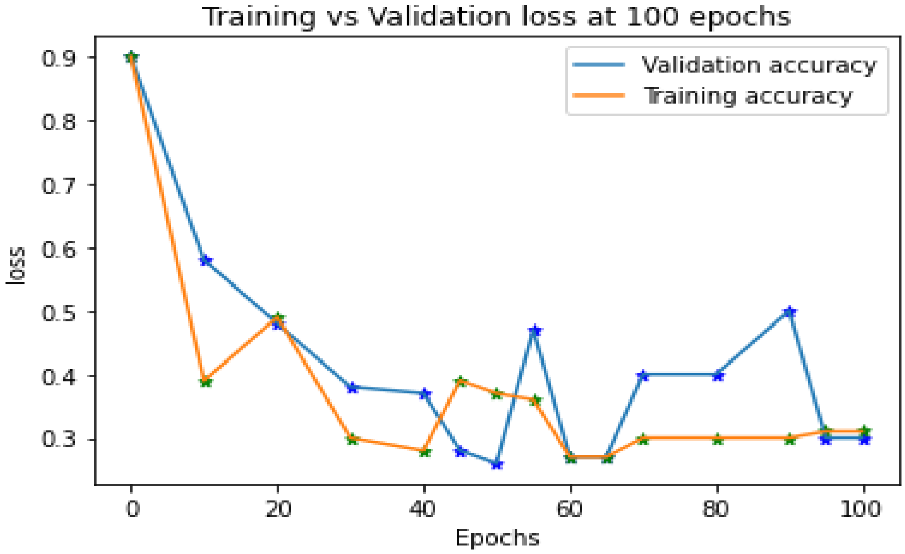

Figure 6. For 100 image dataset

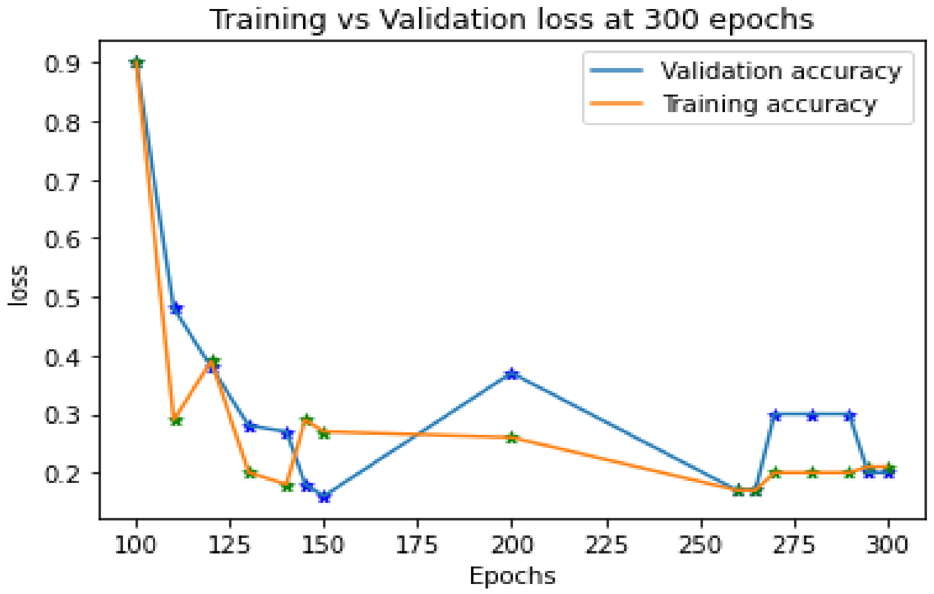

Figure 7. For 300 image dataset

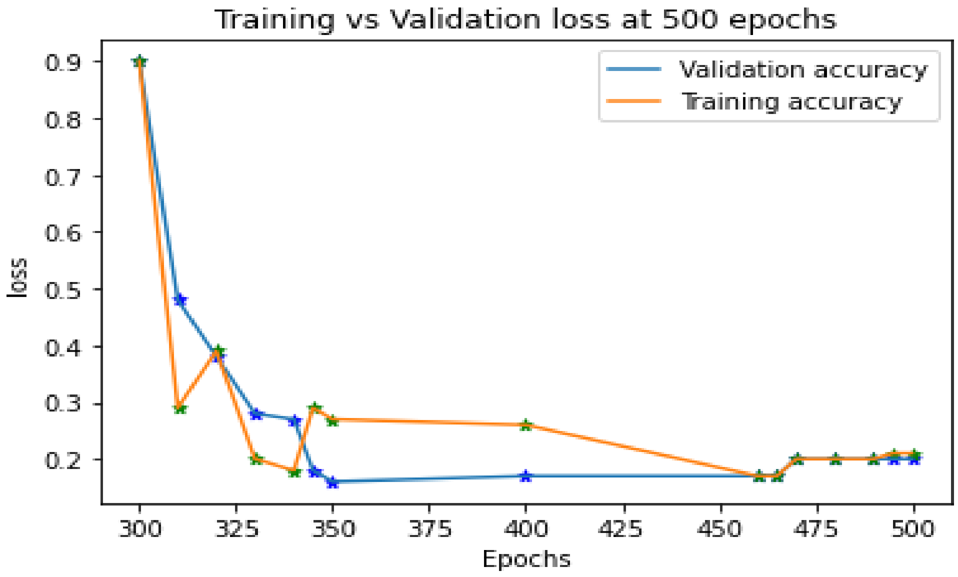

Figure 8. For 500 image dataset

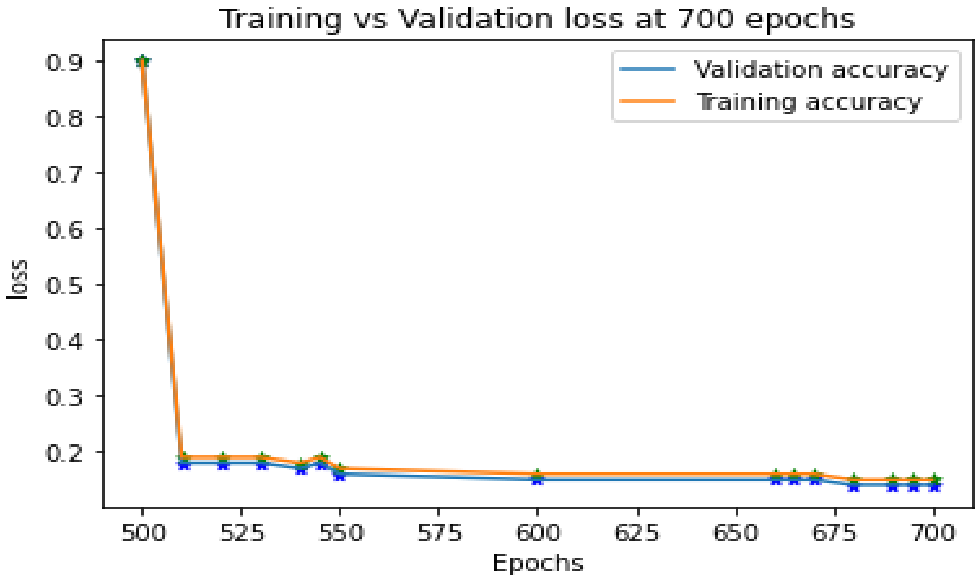

Figure 9 . For 500 image dataset

Conclusion

CNN-MMEI along with evolutionary approach produces higher level of accuracy in chronic heart disease detection. The task performed for preprocessing is through CNN-MMEI classifier which contributes its best in prediction of the disease through various convolutional hidden layers. The features extracted undergo genetic algorithm which is evolutionary based approach which reduces the complexity of features present in the subset. The genetic algorithm is used for generating an optimized function among the random population. It proves its efficiency by providing accuracy of 93.52%. For a clear picture of accuracy , the percentage of conventional methods are depicted in Table 2. The work has demonstrated improved accuracy value by combining evolutionary algorithm with CNN-MMEI.

Performance Evaluation

References

[1] Disease Risk Prediction by Using Convolutional Neural Network, Sayali Ambekar and Rashmi Phalnikar 978-1-5386-5257-2/18/$31.00 ©2018 IEEE

[2] U. R. Acharya, H. Fujita, O. S. Lih, M. Adam, J. H. Tan, and C. K. Chua,``Automated detection of coronary artery disease using different durations of ECG segments with convolutional neural network," Knowl.-Based Syst.,vol. 132, pp. 6271, Sep. 2017.

[3] Vijay Kumar, G., Bharadwaja, A., Nikhil Sai, N, "Temperature and heart beat monitoring system using IOT ",Proceedings - International Conference on Trends in Electronics and Informatics, ICEI 2017

[4] Mining popular patterns from multidimensional database Parallel and distributed frequent-regular patternmining using vertical format in large databases, Vijay Kumar G, krishna chaitanya T, Pratap Indian Journal of Science and Technology (2016).

[5] D.Shanthi, G.Sahoo and N. Saravanan, "Evolving connection Weights of ANN using GA with Application to the Prediction of Stroke Disease",International Journal of Soft computing 4(2): 95-102, ©Medwell Journals, 2009.

[6] M.Morita, R,.Sabourin, F.Bortolozzi, and C.Y.Suen, "Unsupervised feature selection using multi-objective genetic algorithms for handwritten word recognition. In proceedings of the 7th ICDAR, pages
666-670, IEEE Computer Society, 2003.

[7] C.L.Huang, C.J Wang, "A GA based feature selection and parameter optimization for SVM", Expert Systems with Applications, pp.231-240,2006.

[8] Mining high utility regular patterns in transactional database , "Vijay Kumar G., Vishnu Sravya S, Satish G.", International Journal of Engineering and Technology(UAE), 2018.

[9] Bhaskaru, O., Sreedevi, M."Risk feature aware accurate heart disease prediction system using fuzzy extreme learning machine", Journal of Advanced Research in Dynamical and Control Systems.

[10] Srinivas Kolli M. Sreedevi, "PROTOTYPE ANALYSIS OF DIFFERENT DATA MINING CLASSIFICATION AND CLUSTERING APPROACHES", ARPN Journal of Engineering and Applied Sciences.

[11] Srinivas Kolli M. Sreedevi, "Adaptive Clustering Approach to Handle Multi Similarity Index for Uncertain Categorical Data Streams", Journal of Adv Research in Dynamical & Control Systems, Vol. 10, 04-Special Issue, 2018.

[12] Srinivas Kolli M. Sreedevi,"A novel index based procedure to explore similar attribute similarity in uncertain categorical data", ARPN Journal of Engineering and Applied Sciences.

[13] M. Chen, Y. Hao, K. Hwang, L. Wang and L. Wang,"Disease Prediction by Machine Learning Over Big Data From Healthcare Communities," in IEEE Access, vol. 5,pp. 8869-8879, 2017.doi: 0.1109/ACCESS.2017.2694446.

[14] Sayali Ambekar and Dr.Rashmi Phalnikar, "Disease prediction by using machine learning", International

Journal of Computer Engineering and Applications,Volume XII, Special Issue, May 2018.

[15] Ali Mirza Mehmood and Mrithyunjaya Rao Kuppa, "A novel pruning approach using expert knowledge for data-specific pruning", Engineering with Computers pp.21-30,2012.

[16] Zuo.W.M, Lu. W.G, Wang K.Q, Zhang.H, "Diagnosis of cardiac arrhythmia using kernel difference weighted KNN classifier" Computers in Cardiology, pp.253-256, 2008.

[17] Uyar A, Gurgen F, "Arrhythmia Classification Using Serial Fusion of Support Vector Machine and Logistic Regression", Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications, pp.560, 6-8 Sept. 2007.

[18] Alaa M. Elsayad, "Classification of ECG arrhythmia using Learning Vector Quantization Neural Networks" (978-1-4244-5844-8/09/©2009 IEEE), Manuscript received July 30, 2009: revised 1 October 2010

[19] Kemel Polat, Seral Sahan, Salih Gunes, "A new method to medical diagnosis; Artificial immune recognition system (AIRS) with fuzzy weighted pre-processing and application to ECG arrhythmia", Expert systems with applications, Vol.31, Issue 2, pp.264-269, 2006.

[20] San-Hong Lee, Jung-Kwon Uhm, Lim J.S, "Extracting Input Features and Fuzzy Rules for Detecting ECG arrhythmia based on NEWFM", International Conference on Intelligent and Advanced systems, Division of Software, Kyungwon University, Korea.

[21] M.Meenakshi, H.S.Niranjana Murthy, "Comparison of ANN based Heart stroke classifiers using Varied folds dataset cross validation", SPRINGER proceedings of International conference on Intelligent computing, communication & devices (ICCD-2104), 18-19th April 2014.

[22] H.S.Niranjana Murthy, M.Meenakshi, Dimensionality Reduction using Neuro-Genetic approach for Early Prediction of coronary heart disease, Proceedings of International Conference on Circuits, Communication, Control and Computing, 2014.

[23] R.Parthiban, Dr.K.Santhosh Kumar, Dr.R.Sathya, D.Saravanan," A Secure Data Transmission And Effective Heart Disease Monitoring Scheme Using Mecc And Dlmnn In The Cloud With The Help Of Iot", International Journal of Grid and Distributed Computing, ISSN: 2005 – 4262, Vol. 13, No. 2, (2020), pp. 834 – 856.

[24] R.Bhavya, G.I.Archanaa, D.Karthika, D.Saravanan," Reflex Recognition of Tb Via Shade Duplicate Separation Built on Geometric Routine", International Journal of Pure and Applied Mathematics 119 (14), 831-836.

[25] D Saravanan, R Bhavya, GI Archanaa, D Karthika, R Subban," Research on Detection of Mycobacterium Tuberculosis from Microscopic Sputum Smear Images Using Image Segmentation", 2017 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC).

[26] D.Raghu Raman, D.Saravanan, R.Parthiban, Dr.U.Palani, Dr.D.Stalin David, S.Usharani, D.Jayakumar, " A Study On Application Of Various Artificial Intelligence Techniques On

Internet Of Things", European Journal of Molecular & Clinical Medicine ISSN 2515-8260 7 (9), 2531-2557.

[27] D Saravanan, J Feroskhan, R Parthiban, S Usharani, "Secure Violent Detection in Android Application with Trust Analysis in Google Play", Journal of Physics: Conference Series 1717 (1), 012055.

[28] D Saravanan, E Racheal Anni Perianayaki, R Pavithra, R Parthiban, " Barcode System for Hotel Food Order with Delivery Robot", Journal of Physics: Conference Series 1717 (1), 012054.

[29] D Raghu Raman, S Gowsalya Devi, D Saravanan, "Locality based violation vigilant system using mobile application", 2020 International Conference on System, Computation, Automation and Networking (ICSCAN-IEEE).

[30] R Parthiban, R Ezhilarasi, D Saravanan, "Optical Character Recognition for English Handwritten Text Using Recurrent Neural Network", 2020 International Conference on System, Computation, Automation and Networking (ICSCAN-IEEE).

[31] R Parthiban, V Abarna, M Banupriya, S Keerthana, D Saravanan, " Web Folder Phishing Discovery and Prevention with Customer Image Verification", 2020 International Conference on System, Computation, Automation and Networking (ICSCAN-IEEE).