International Journal of Communication Networks and Information Security

ISSN: 2073-607X, 2076-0930
*Volume **14** Issue **1s** Year **2022** Page **167 : 176***

An Investigation on Disease Diagnosis and Prediction by Using Modified KMean clustering and Combined CNN and ELM Classification Techniques

Saiyed Faiayaz Waris
Research Scholar, Department of Computer Science and Engineering,
VelTech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Chennai, India.
saiyed.cse@gmail.com
S. Koteeswaran
Professor, Department of Computer Science and Engineering,
VelTech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Chennai, India
saiyed.cse@gmail.com

Article History

Received: 13 July 2022
Revised: 20 September 2022
Accepted: 26 October 2022

Abstract

Data analysis is important for managing a lot of knowledge in the healthcare industry. The older medical study favored prediction over processing and assimilating a massive volume of hospital data. The precise research of health data becomes advantageous for early disease identification and patient treatment as a result of the tremendous knowledge expansion in the biological and healthcare fields. But when there are gaps in the medical data, the accuracy suffers. The use of K-means algorithm is modest and efficient to perform. It is appropriate for processing vast quantities of continuous, high-dimensional numerical data. However, the number of clusters in the given dataset must be predetermined for this technique, and choosing the right K is frequently challenging. The cluster centers chosen in the first phase have an impact on the clustering results as well. To overcome this drawback in k-means to modify the initialization and centroid steps in classification technique with combining (Convolutional neural network) CNN and ELM (extreme learning machine) technique is used. To increase this work, disease risk prediction using repository dataset is proposed. We use different types of machine learning algorithm for predicting disease using structured data. The prediction accuracy of using proposed hybrid model is 99.8% which is more than SVM (support vector machine), KNN (k-nearest neighbors), AB (AdaBoost algorithm) and CKNCNN (consensus K-nearest neighbor algorithm and convolution neural network).

Keywords: Heart Disease Prediction, K means, (CNN) Convolutional neural network, (ELM) extreme learning machine

1. Introduction

Heart disease is the primary cause of death in the world today. Heart condition may be a favorite problem from the planet. Heart condition is leading explanation for death within the world. Heart disease has been on the rise during the past several years. There are many different types of reasons, causes, and factors that raise the risk of heart disease. There are regarded as a significant contributing factor to heart disease. The majority of hospitals admit patients with heart conditions. Because of smoking habits, men are more impacted by this condition. This research uses categorization algorithms to analyze the numerous types of cardiac conditions.

Heart is vital a organ of human body. The proper functioning of a heart is entwined with life itself. When the body's blood circulation is ineffective, organs including the brain suffer, and the heart may not function effectively. More than just a heart disease attack. Many hospitals today do not provide sufficient care however, increasing bill payment. Many hospitals treat patients on average, which leads to better outcomes.

Heart condition is that the leading explanation for per annum death within the world. Some sorts of disease create heart attack, they are coronary heart condition, angina, congestive coronary failure, cardiomyopathy, congenital heart condition, rhythmias, myocarditis, heart attack; heart cancer etc. Convolutional Neural Networks distinguish between objects in a picture by first assigning each one a weight based on its individual components. Compared to other deep learning methods, CNN requires a remarkably minimal amount of data pre-processing. One of CNN's main advantages is that it trains its classifiers with straightforward methods, allowing it to learn the characteristics of the target object. Convolution is a mathematical operation that is used to two functions to provide an output in the form of a third function that illustrates how the shape of one function is being affected or modified by the other function. It has three layers: convolution layer, pooling layer and fully connected layer. The first layer output is given to second layer as an input, which extract other important features of the input image such as corners and edge combinations. ELM is a feedforward neural networks that is different from standard neural network. It doesn't need gradient based back propagation to perform. The training of Single Hidden Layer Feedforward Neural Networks is made easier by the use of the ELM, a supervised learning framework (SLFN).

The driving force behind this is frequently to manage enormous amounts of heart condition data and, consequently, to forecast the threat of heart conditions. Information cleaning and data restoration are essential when medical data are not the proper format. The inability to execute disease prediction due to poorly prepared data can occasionally result in inaccurate disease prediction [1]. The prediction of disease based on symptoms is already done with the aid of structured data and predicted disease using the naive Bayes procedure. This article performs an operation on structured medical data. A deep learning concept called a convolutional neural network automatically extracts features from a sizable dataset to produce the required outcome. CNNUDRP is used for structured data to extract the required feature variables from the dataset and to administer illness prediction based on that dataset. The main intention of this article is to take the concept of structured data to forecast the risk of both gut disease and cardiac condition. The second goal is dealing with lacking values to discover the precise hazard of cardiac problem.

2. Literature Survey

[1] Multimodal illness risk prediction technique using hospital structured and unstructured data proposed using convolution neural networks. For several regions, authors developed disease prediction systems. Additionally, foresee that whether a patient encounter from the high risk of cerebral localized necrosis or low risk of cerebral localized necrosis. The exactness of illness expectation comes to up to 94.8% with faster than Convolutional neural organization based unimodal infection hazard forecast calculation.

In paper [2], the Alzheimer illness hazard forecast framework designed by the author with the help of EHR information of the patient. Here, they made use of a dynamic learning environment to sort out a real problem the patient was dealing with. The experts are aware of the similar conditions affecting the two patients, and they effectively analyze the patient risk.

Cloud-based wellbeing designed – Cps framework in [3], which deal with the gigantic measure of biomedical information. This framework performed different process on cloud-like information investigation, detecting and prediction of information. With the help of this strategy, a private gets massive amount of data about to comply with the gigantic measure of biomedical information inside the cloud. Likewise, the differed administrations identified with medical services familiar by this framework.

In paper [4], the author projected wearable 2.0 framework during which configuration shrewd launder able garments that improves the QoE and QoS of the cutting edge medical services framework. This information was used to capture the patient's physiological state. Additionally, this information is used for examination purposes. Discuss the concerns which face while planned the wearable 2.0 design. There are various uses discussed in this, including continued infection checking, old individuals care, feeling care and so forth.

[5] Proposed telehealth framework that examines the best approach to deal with an outsized measure of medical clinic information inside the cloud platform. A new ideal large information sharing calculation is considered in this paper. Clients determine the best layout for maintaining biomedical information using this approach.

[6] Proposed a substitution mixture method called the Genetic Algorithms (GAs) and Backing Vector Machines (SVMs). From the developmental measure, they need demonstrated that hereditary calculation is ideal appropriate for the get-together of trait informational index. It achieved a high normal exactness of 76.20%.

[7] Have made is coronary illness in contrast with raised precision. Numerous issues gather from heart. Specific same sickness information contrast with 60 calculations. Normal precision for diabetes. They accomplished from the projected RF the exactness of 74.47%, 80.49%and 87.13% individually. The heart disease forecasting includes heart or blood vessels. In order to lower the risk of illness, it is crucial to diagnose and anticipate heart diseases [9]. Extreme study is being done in the field of healthcare to forecast heart disorders.

For the prediction of heart condition, CNN and genetic algorithms are typically used. Additionally, age, family history, diabetes, cholesterol, smoking, hypertension, alcohol consumption, obesity, inactivity, and other significant factors are taken into account [10].

In paper [11], the authors used machine learning methods such Naive Bayes, Neural networks, and Decision tree algorithms to create the gastrointestinal disease prediction system.

In their study, Alexander, C.A. and Wang [12] looked through national and international databases to find studies about big data analytics in healthcare, heart attack prevention and prediction, big data technology, and privacy issues. According to the research that were examined, big data analytics can be used to predict heart attacks, and are crucial for handling and tailoring cardiovascular disease treatment. Furthermore, as the usage of big data in healthcare increases, patients will have accessible to better customized treatments. Patient privacy is protected in this improvement and application of big data, sensor use, mobile or tablet use, and landline use, as well as to protect the patient's capacity to control and determine how his/her health info is used.

Manogaran G *et al*. [13] has introduced tools of handling big data in healthcare where we have different sources of information this tool helps the advancement of big data in providing the more awareness of diseases and way of preventing them. In his research he has proposed different tools such as EHRs (Electronics Health Records) to gather and store the patient data. Social health is also seen as one of the big data holders where a patient is connected to the doctor on the far side of clinic with involvement of social networks. Here patients can communicate on their health and this social health supply big data.

The function of mobile cloud computing and big data analytics in allowing networked healthcare has been examined by Lo'ai AT et al. [14]. The adoption of cloud computing in healthcare is presented together with the inspiration for and creation of networked healthcare programs and infrastructure. A

mobile cloud computing structure based on cloudlets that will be used for big data applications in healthcare is described. Review of big data analytics methods, equipment, and applications. The use of big data and mobile cloud computing technologies is used to derive conclusions about the design of networked healthcare systems. It will be necessary to develop such intricate networked healthcare systems using modeling techniques that make use of high performance computing and big data technology.

In order to provide trustworthy disease-oriented monitoring and projection in this information age, Wong ZS. *et al.* [15] has created Artificial Intelligence (AI) technologies. With trustworthy data management platforms, AI technologies will make it possible to analyze vast amounts of surveillance and infectious disease data in an efficient manner to support government organizations, healthcare providers, and medical experts in the future in their response to disease. Governmental organizations, healthcare providers, and medical professionals will be interested in the improvements in syndromic surveillance that result from this for risk analysis and resource allocation..

3. Proposed Methodology

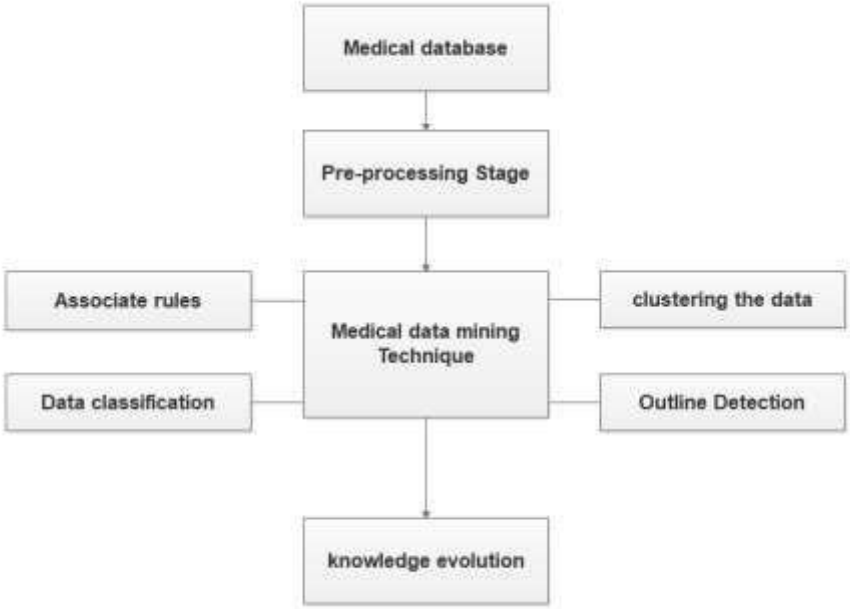

Figure 1. Framework for medical data mining

The framework is illustrated in figure.1 above, which begins with a specific medical issue. Before mining the data with one of the various data mining tools, the dataset should be cleaned up and preprocessed. Knowledge evaluation comes last, and medical professionals should be included.

In above model based, we proposed to modify the clustering and classification section. The K-means method has the advantages of simplicity, usability, and execution efficiency. It is appropriate for processing a sizable volume of continuous and high-dimensional numerical data. The data attributed to the same cluster are also very similar to one another. The K-means technique has several downsides, too. Users must predetermine the number of clusters in the data set, and choosing the right K is frequently challenging and the cluster centers chosen in the first stage frequently have an impact on clustering results. To overcome these drawbacks in k-means to modify the initialization and centroid steps.

And also in classification technique we make to combining Convolutional neural network and extreme learning machine technique. This makes to increase prediction and diagnose of diseases accuracy in medical sector. In here below figure 2 display an above modified proposed classification and clustering system model.

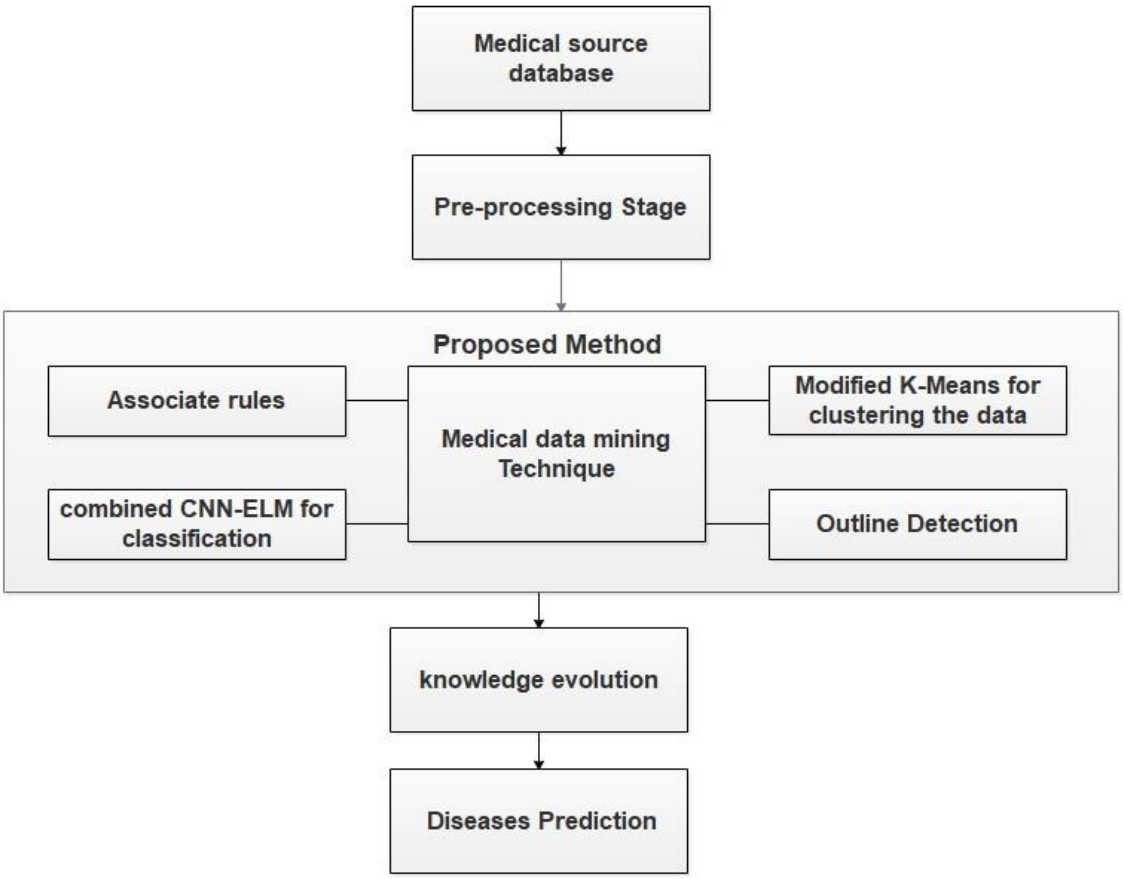

Figure 2. Proposed modified framework for medical data mining

4. Modified K-Means Clustering

The modified k-mean clustering is an unsupervised learning algorithm and its goal is to classify the dataset that is unlabeled and then the functions are clustered according to their similar features by using Euclidian distance formula. This algorithm aims to identify groups that are comparable to the ones represented by the variable k. Since there are k clusters in this scenario, there are a fixed number of centroids for each cluster. The cluster quality is improved by the enhancement of Euclidian formula. Based on the normalization, the enhancement is performed. The feature that is added is for calculating normal distance metrics on the basis of normalization, which is a pre-processing method that will increase the precision and potency of clusters by calculating best distances from the dataset that will result in more precise center points and as a result best clusters are formed [23].

Working Principle:
Step 1: In initial stage load the given dataset
Step 2: Dataset is distributed and plotted
Step 3: Now, modified k-means clustering is given on the dataset. For calculating centroids, Euclidian formula is used.
Step 4: Time and accuracy of centers is identified.
Step 5: For classifying the generated dataset Combined CNN and ELM technique is applied.
Step 6: After the normalization, iteration process is begin and more closer and exact centers are calculated to plot data.
Step 7: Time and accuracy is again calculated in which best accuracy of clusters are obtained. Let X = {d1,d2,d3,……dn} be the set of data points and Y = {z1,z2,z3,…….zc} be the set of centers.

1. Randomly select cluster centers “c”.

2. Determine the distance between the cluster centers and each data point.
3. Choose the cluster center from among all the cluster centers whose distance to the data point is the shortest.
4. For recalculating the cluster center, below formula is used,

   $Y_i = \left(\frac{1}{c_i}\right) \sum_{j=1}^{c_i} d_i$

   Where,

   c_i – Amount of data points in i^{th} cluster
5. Reevaluate the distance between every data point and the recently found cluster centers.
6. Stop if data was not changed, otherwise, start over at 3rd step.

5. Combined CNN-ELM Classification Technique

Convolution neural network is a very famous image recognition technique and one of the important deep learning methods. It is also one of the artificial neural networks, which mostly used in image analysis sectors. Figure 3 represents the basic structure of combined CNN and ELM classification model for the proposed work. Accuracy is calculated after the successful completion of CNN based feature extraction and ELM based classification.

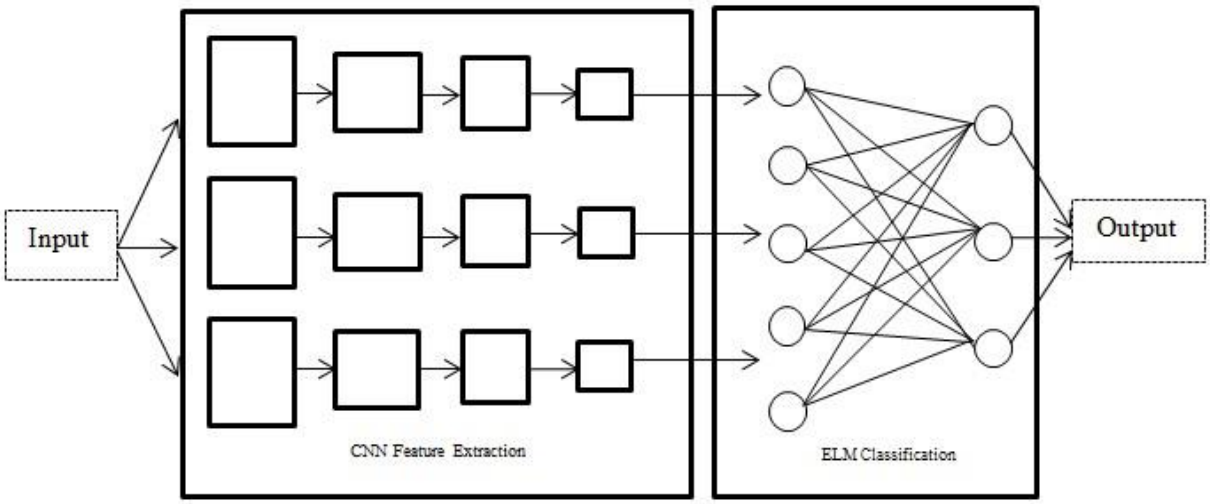

Figure 3. Structure of CNN-ELM Classification Model

The quick learning algorithm known as the "extreme learning machine" was developed for an image recognition system with a single hidden layer. No need to change or update the parameters during training; just change the hidden layer nodes to discover the optimal solution [24]. Compared to traditional classification techniques like CNN and SVM, ELM is more powerful since it learns quickly, has a strong generalization capacity, and requires less parameter changes.

Formally, consider a collection of "n" randomly different samples (s_i, D_i) for a single hidden layer neural network. Where, $s_i = \{s_{i1}, s_{i2}, \ldots . s_{in}\}^t \in R^p$, $D_i = \{D_{i1}, D_{i2}, \ldots . D_{im}\}^t \in R^q$ for N number of hidden nodes, a single layer may be expressed as,

$$L_i = \sum_{j=1}^{l} \alpha_j \; G\left(W_j . s_i + b_j\right) \qquad i = 1,2,3, \ldots \ldots, N \tag{1}$$

Where, activation function represented by $G(s)$, $W_j = \{W_{j1}, W_{j2}, \ldots \ldots, W_{jn}\}^t$ constitutes weights of input. α_j is the weight of output, bias is represented by b_j. The inner combination of the input weights and the input samples is denoted by $(W_j . s_i)$. to lower output error, only one hidden layer is used. The ELM

technique does not need any parameter adjustments. The α of the hidden layer and outcome matrices H are selected uniquely once the input weights and bias have been determined at random.

6. Result

For heart disease risk diagnosis and prediction, dataset from Kaggle is used in this paper. The dataset contain patient's health information with different age. From the overall dataset 10 datasets are shown in table 1.

Table 1. Dataset of Patients with Different Age

Dataset	**Age**	**Sex**	**cp**	**trtbps**	**chol**	**fbs**	**Rest_ecg**	**thalach**	**exang**	**caa**	**output**
1	63	1	3	145	233	1	0	150	0	0	1
2	37	1	2	130	250	0	1	187	0	0	1
3	41	0	1	130	204	0	0	172	0	0	1
4	56	1	1	120	236	0	1	178	0	0	1
5	57	0	0	120	354	0	1	163	1	0	1
6	57	1	0	140	192	0	1	148	0	0	1
7	67	1	0	160	286	0	0	108	1	3	0
8	63	1	0	130	254	0	0	147	0	1	0
9	53	1	0	140	203	1	0	155	1	0	0
10	48	1	1	110	229	0	1	168	0	0	0

The clinical feature representation of the entire dataset is illustrated in table 2. All the data is collected from Kaggle heart disease prediction dataset.

Table 2. Clinical Feature Representation

Age	Age of the patient in years
Sex	Patient gender
Cp	Chest pain type
Trtbps	Resting blood pressure (mmHg)
Chol (mg/dl)	Serum cholesterol (mg/dl)
FBS	Fasting Blood Sugar
Rest_ecg	Resting electrocardiographic result
Thalach	Maximum heart rate achieved
Exang	Exercise induced angina
Caa	Significant number of vessels (0-3)
Output	1 = Yes; 0 = No

This section provides an overview of medical data mining methods used for the diagnosis and prognosis of various heart conditions. Unless an important study from before that should be included, only literature studies from 2016 and beyond are considered. Processing is done on heart disease symptoms. Each feature has a prescribed range of values, and standard datasets are created for every cardiac illness. Such datasets have been employed in numerous research investigations to learn more about cardiac issues.

In this section, we'll examine several data mining techniques used generally in the field of healthcare after looking at various heart artery disorders and symptoms that, when they exhibit certain characteristics, indicate a heart ailment. Such vast amounts of data can be used for knowledge discovery in the healthcare industry, using elements like symptoms and patient records. The highest level of accuracy for data mining approaches used for heart disease diagnosis and prediction is shown in Table 3.

Table 3. Data Mining Approaches for Diagnosis and Prediction of Heart Disease

Data Mining Techniques	**Purpose of study**	**Maximum Accuracy level**
Neural Network	To diagnos e the presence of coronary heart diseases.	91%
Naïve Bayes Classifier +Genetic Algorithm Feature Reduction	To diagnose the presence of cardiovascular disease.	96.5%
Decision Tree +Genetic Algorithm Feature Reduction	To diagnose the presence of cardiovascular disease.	99.2%
RIPPER Classifier	To diagnose the presence of cardiovascular disease.	81.08%
C5 Classifier	To diagnose the presence of coronary heart diseases.	89.6%
Hybrid Genetic Neural Network	To diagnose the presence of cardiovascular disease	89%
Consensus K-nearest neighbor algorithm and convolution neural network	Prediction of heart condition	99.1%
Proposed Combined CNN-ELM Technique	To diagnose and prediction of Cardiovascular diseases (CVDs) disease risk	99.8%

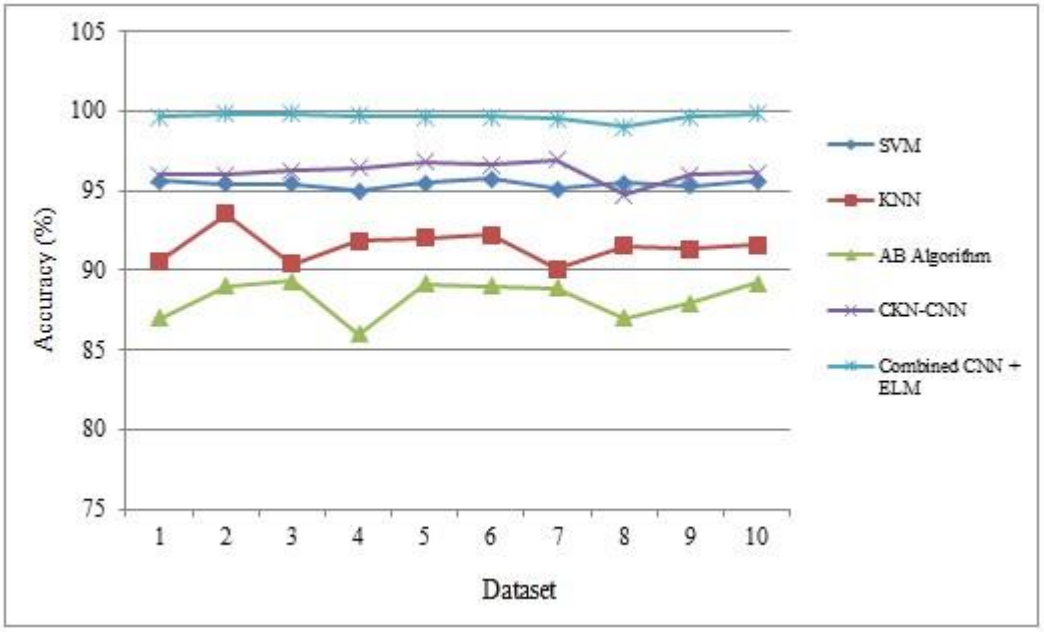

Figure 4. Classification Accuracy Results

In above describe some important data mining models used for heart disease diagnosis and prediction to calculate the accuracy level. Initially To diagnose the presence of coronary heart diseases by using neural network, it attain the accuracy level of 91%. And also C5 Classifier achieved 89.6% of accuracy in diagnose of coronary heart diseases. And then diagnose of cardiovascular disease by using decision tree +genetic algorithm feature reduction, Ripper classifier and hybrid genetic neural network. In which C5 classifier achieved of 89.6% accuracy and Ripper classifier achieved the accuracy of 81.08%., Decision Tree +Genetic Algorithm Feature Reduction achieved 99.2% of accuracy and Consensus K-

nearest neighbor algorithm and convolution neural network achieved the accuracy of 99.1%. It is a highest accuracy by comparing other techniques. And also lowest accuracy was attained by Ripper classifier. It is not enough level accuracy in medical sector.
In this paper, hybrid clustering and classification technique is implemented in data mining to achieve better prediction accuracy level. The classification accuracy result of CNN and ELM is compared with previous effective classification method and the results shows that the proposed combined method provides better accuracy. The classification accuracy of outcomes is represented in figure 4.

7. Conclusion

In this paper, a combined CNN and ELM model for heart disease diagnosis and prediction is presented. From the ten datasets, a comprehensive analysis is performed and compared against SVM, KNN AB Algorithm, and CKN-CNN. CNN-ELM technique is simple and able to provide better prediction accuracy than other complex machine learning models with less time. With the purpose of enhancing the classification accuracy and training speed of the classifier, combined CNN-ELM model is build. Experimental results on various datasets confirm the effectiveness, efficiency, and prediction accuracy of the proposed CNN-ELM classifier.

References

[1] M. Chen, Y. Hao, K. Hwang, L. Wang, and L. Wang, "Disease prediction by machine learning over big data from healthcare communities," IEEE Access, vol. 5, no. 1, pp. 8869– 8879, 2017.

[2] B. Qian, X. Wang, N. Cao, H. Li, and Y.-G. Jiang, " A relative similarity based method for interactive patient risk prediction," DataMiningKnowl.Discovery, vol. 29, no. 4, pp. 1070–1093, 2015.

[3] IM. Chen, Y. Ma, Y. Li, D. Wu, Y. Zhang, and C. Youn, " Wearable 2.0: Enable humancloud integration in next generation healthcare system," IEEE Commun. , vol. 55, no. 1, pp. 54–61, Jan. 2017.

[4] Y. Zhang, M. Qiu, C.-W. Tsai, M. M. Hassan, and A. Alamri,"HealthCPS: Healthcare cyberphysical system assisted by cloud and big data," IEEE Syst. J., vol. 11, no. 1, pp. 88–95, Mar. 2017.

[5] L. Qiu, K. Gai, and M. Qiu, "Optimal big data sharing approach for telehealth in cloud computing," in Proc. IEEE Int. Conf. Smart Cloud (Smart Cloud), Nov. 2016, pp. 184–189.

[6] Keyue Ding and Kent R Bailey et al, " Genotypeinformed estimation of risk of coronary heart disease based on genome-wideassociation data linked to the electronic medical record", International journal of BMC cardiovascular Disorders, Vol.11,2011

[7] Shou-En Lu and Gloria L Beckleset al, "Evaluation of risk equations for prediction of shortterm coronary heart disease events inpatients with long-standing type 2 diabetes: the Translating Research into Action for Diabetes", International Journal of BMC Endocrine Disorders, Vol.12, 2012

[8] Lee CH, Yoon HJ. Medical big data: promise and challenges. Kidney research and clinical practice. 2017 Mar;36(1):3.

[9] Dinov ID. Methodological challenges and analytic opportunities for modeling and interpreting Big Healthcare Data. Gigascience. 2016;

[10] Dharayani, Ramanti, et al. "Genomic Anomaly Searching with BLAST Algorithm using MapReduce Framework in Big Data Platform." 2019 International Workshop on Big Data and Information Security (IWBIS). IEEE, 2019.

[11] Siuly S, Zhang Y. Medical big data: neurological diseases diagnosis through medical data analysis. Data Science and Engineering. 2016 Jun 1;1(2):54-64.

[12] Rodger, J.A., 2015. Discovery of medical Big Data analytics: Improving the prediction of traumatic brain injury survival rates by data mining Patient Informatics Processing Software Hybrid Hadoop Hive. Informatics in Medicine Unlocked, 1, pp.17-26.

[13] Karthik R, Menaka R, Chellamuthu C. A comprehensive framework for classification of brain tumour images using SVM and curvelet transform. International Journal of Biomedical Engineering and Technology. 2015 Jan 1;17(2):168-77.
[14] Ta VD, Liu CM, Nkabinde GW. Big data stream computing in healthcare real-time analytics. In2016 IEEE International Conference on Cloud Computing and Big Data Analysis (ICCCBDA) 2016 Jul 5 (pp. 37-42). IEEE.
[15] Ristevski B, Stevanovska M, Kostovski B. Hadoop as a Platform for Big Data Analytics in Healthcare and Medicine.
[16] Selvakumar S, Parkavi R, Suganya R, Abirami AM. Big Data Analytics in Healthcare Sector. InMachine Learning Techniques for Improved Business Analytics 2019 (pp. 94-106). IGI Global.
[17] Kuo A, Chrimes D, Qin P, Zamani H. A Hadoop/MapReduce Based Platform for Supporting Health Big Data Analytics. InITCH 2019 Mar 26 (pp. 229-235).
[18] Vaishali, G. and Kalaivani, V., 2016, January. Big data analysis for heart disease detection system using map reduce technique. In 2016 International Conference on Computing Technologies and Intelligent Data Engineering (ICCTIDE'16) (pp. 1-6). IEEE.
[19] Alexander, C.A. and Wang, L., 2017. Big data analytics in heart attack prediction. J Nurs Care, 6(393), pp.2167-1168.
[20] Manogaran G, Thota C, Lopez D, Vijayakumar V, Abbas KM, Sundarsekar R. Big data knowledge system in healthcare. InInternet of things and big data technologies for next generation healthcare 2017 (pp. 133-157). Springer, Cham.
[21] Lo'ai AT, Mehmood R, Benkhlifa E, Song H. Mobile cloud computing model and big data analysis for healthcare applications. IEEE Access. 2016 Sep 26;4:6171-80.
[22] Wong ZS, Zhou J, Zhang Q. Artificial intelligence for infectious disease big data analytics. Infection, disease & health. 2019 Feb 1;24(1):44-8..
[23] Kumar, S., & Kaur, S. (2017). Modified K-Means Clustering Algorithm for Disease Prediction. *International Journal of Engineering and Technquesm*, *3*(3), 195-201.
[24] Wang, P.; Zhang, X.; Hao, Y. A Method Combining CNN and ELM for Feature Extraction and Classification of SAR Image. J. Sens. 2019, 2019, 1–9.