iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 7 **Issue:** VI **Month of publication:** June 2019

DOI: http://doi.org/10.22214/ijraset.2019.6114

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.177
Volume 7 Issue VI, June 2019- Available at www.ijraset.com

Disease Prediction by Machine Learning from Healthcare Communities

Sakshi Gupta[1], Pooja Navali[2], Amruta Sao[3], Savita Bhat[4], Prof. Yogesh Thorat[5]
[1, 2, 3, 4, 5]*Computer Department, Savitribai Phule Pune University*

Abstract: ***These days in biomedical field use of Mining and machine learning knowledge is expanding, genuine study of medicinal dataset advantages in early illness discovery, quiet care and organization administrations. The machine studying calculations are proposed for a success expectation of ceaseless infection to beat the trouble of deficient facts. Genetic set of rules will be applied to improve the lacking statistics. The set of data accommodates of structured facts and unstructured facts. RNN rules are applied for extraction of capabilities from unstructured information. For prediction of a disease, framework proposes SVM calculation and Naive Bayesian calculation for unstructured and structured statistics personally from medical data. Two calculation KNN and SVM are proposed to give the proper solution of the inquiry.Community Question Answering (CQA) gadget is moreover proposed which will also provide proper responses to the clients.KNN set of rules will carry out class on solutions and SVM calculation will carry out class on answers. it is going to assist patient to discover quality inquiries and answers identified with infections.***

Keywords: Data analytics; Machine Learning; Healthcare; Community Question Answering (CQA), K Nearest Neighbor, (KNN) and Support Vector Machine (SVM), electronic health records (EHR).

I. INTRODUCTION

With the development of residing requirements, the occurrence of chronic disease is growing. it's far important to carry out hazard assessments for chronic illnesses. With the boom in clinical information, amassing electronic health records (EHR) is increasingly convenient. Proposed a healthcare system the usage of smart clothing for sustainable health monitoring had thoroughly studied the heterogeneous systems and done the satisfactory results for price minimization on tree and easy direction cases for heterogeneous systems, patients' statistical data, take a look at outcomes and ailment records are recorded within the EHR, enabling us to discover ability statistics-centric solutions to reduce the charges of medical case studies. Proposed a green float estimating set of rules for the tele-fitness cloud machine and designed a statistics coherence protocol for the PHR (non-public fitness file)-primarily based distributed system. Cloud system and designed a statistics coherence protocol for the PHR (personal fitness document)-primarily based disbursed device. Proposed six packages of massive records inside the field of healthcare but those schemes have characteristics and defects also. The information set is commonly small, for patients and diseases with particular situations; the traits are selected thru revel in. but, these pre-selected characteristics perhaps now not fulfil the changes in the ailment and its influencing factors.[6] With the development of large data analytics technology, extra attention has been paid to disorder prediction from the angle of massive records evaluation, diverse researches were conducted by means of selecting the traits mechanically from a massive range of information to improve the accuracy of risk category, in preference to the formerly selected traits. however, the ones current work usually taken into consideration-based information. For unstructured records, as an example, the use of convolutional neural network (CNN) to extract textual content characteristics routinely has already attracted extensive interest and additionally achieved superb consequences. Moreover, there's a large distinction among sicknesses in different areas, more often than not because of the various weather and dwelling behaviour in the region. for that reason, hazard type based on massive information analysis, the subsequent demanding situations remain: How must the missing statistics be addressed? How need to the primary chronic sicknesses in a sure area and the primary characteristics of the sickness in the place be decided? How can huge records analysis generation be used to analyse the disease and create a higher model?

To remedy these problems, we integrate the established and unstructured data in healthcare subject to assess the threat of disorder. First, we used latent thing model to reconstruct the missing records from the medical data accumulated from a medical institution in imperative China. 2nd, by way of using statistical expertise, we ought to decide the important persistent illnesses within the location. 1/3, to deal with dependent facts, we discuss with health centre experts to extract beneficial capabilities. For unstructured textual content records, we select the functions robotically using CNN algorithm.

Subsequently, we propose a singular CNN-primarily based multimodal disease risk prediction (CNN-MDRP) algorithm for based and unstructured information. The disease hazard model is received through the mixture of structured and unstructured features. via the test, we draw at end that the overall performance of CNN-MDPR is higher than other current strategies. consumers.

II. PROBLEM STATEMENT

Existing scheme has some defects. Like the data set is typically small, for patients and diseases with specific conditions, the characteristics are selected through experience. However, these pre-selected characteristics maybe not satisfy the changes in the disease and its influencing factors. To overcome these problems, in this paper, this project proposes SVM algorithm & Naive Bayesian algorithm for disease prediction using unstructured and structured data, respectively from hospital.

III.RELATED WORK

Min Chen, Yixue Hao, Kai Hwang, Lu Wang, and Lin Wang, [1] In this paper, we streamline machine learning algorithms for effective prediction of chronic disease outbreak in disease-frequent communities. We experiment the modified prediction models over real life hospital data collected from central China in 2013-2015. To overcome the difficulty of incomplete data, we use a latent factor model to reconstruct the missing data. We experiment on a regional chronic disease of cerebral infarction. We propose a new convolutional neural network based multimodal disease risk prediction (CNN-MDRP) algorithm using structured and unstructured data from hospital.To the best of our knowledge, none of the existing work focused on both data types in the area of medical big data analytics. Compared to several typical prediction algorithms, the prediction accuracy of our proposed algorithm reaches 94.8% with a convergence speed which is faster than that of the CNN-based unimodal disease risk prediction (CNN-UDRP) algorithm.

W. Yin and H. Schutze, [2] In this the new deep learning architecture Bi-CNN-MI paraphrases identification (PI). The PI compares two sentences on multiple levels of granularity. In this BI-CNN means two CNN and MI is Multigranular interaction. They determine whether paraphrase roughly have the same mea

ning. They are closely related to NN for sentence representation and text matching. They are mainly based on Convolutional sentence model. The parameters of the entire model are optimized for PI. Use of language modelling task is to address the lack of training data. Results on the MSRP corpus surpass that of previous NN competitors. The Bi-CNN-MI can be used for sentence matching, question answering in future. The new deep learning architecture Bi-CNN-MI Paraphrase Identification (PI). The PI contemplates two sentences on various levels of granularity. They choose if rephrase by and large has a similar importance. The parameters of the considerable number of models are updated for PI. Usage of vernacular showing task is to address the nonattendance of planning data.

Seema sharma, Jitendra Agarwal, Shikha Agarwal, Sanjeev Sharma, [3] In this the clinical data demonstrate the categories and treatment of patients that represent the under used data sources which are much greater in research potential than the currently which is realized. The potential of EHR (Electronic Health Record) is for establishing the new patients by revealing the unknown disease correlation. In EHR and mining of it a broad range of ethical, legal and technical reasons may hinder the systematic deposition. The potential for the medical research and clinical health care by using EHR data and the challenges which can be overcome before this becomes a reality. The capacity of Electronic Health Record (EHR) is for setting up the new patients by revealing the dark sickness connection. In EHR and its mining a sweeping extent of good, honest to goodness and particular reasons may keep the systematic declaration. The tele-health administrations are being used which are known as the tele-health cautioning organizations. They are generally used as a piece of metropolitan urban communities.

Jensen PB, Jensen LJ, Brunak S, [4] In this the tele-health services are being used which are known as the telephone health advisory services. They are mostly used in metropolitan cities. Due to tele-health services the patients can get a help easily. Rapid increase in tele-health system has received various techniques like cloud computing and big data. They have proposed a dynamic programming to produce optimal solutions so that data sharing mechanisms can be handled. In this it considers the transmission probabilities, the timing constraints, and also the maximizing network capacities. Due to tele-health organizations the patients can get help effortlessly. A quick incremental in the tele-health structure has become diverse strategies like distributed computing and enormous information. They have a dynamic programming to make perfect game plans with the objective that data sharing frameworks can be dealt with. In this it contemplates the transmission probabilities, the arranging objectives, and moreover increasing as far as possible. L. Qiu, K. Gai, and M. Qiu, [5] In this for a content conclusion examination with jointed Convolutional Neural

Network (CNN) and Recurrent Neural Network (RNN) engineering, taking the upsides of both like course grained neighborhoods highlights features which are made by CNN and long-separate conditions learned by methods for the RNN. The provincial perpetual infection has been engaged.

IV. METHODOLOGY

A. CNN

A feature is an individual measurable property or characteristic of a phenomenon being observed. Choosing informative, discriminating and independent features is a crucial step for effective algorithms in pattern recognition, classification and regression. When the input data to an algorithm is too large to be processed and it is suspected to be redundant (e.g. the same measurement in both feet and meters, or the repetitiveness of images presented as pixels), then it can be transformed into a reduced set of features (also named a feature vector). Determining a subset of the initial features is called *feature selection*.[1] The selected features are expected to contain the relevant information from the input data, so that the desired task can be performed by using this reduced representation instead of the complete initial data. Feature extraction involves reducing the amount of resources required to describe a large set of data. When performing analysis of complex data one of the major problem's stems from the number of variables involved. Analysis with a large number of variables generally requires a large amount of memory and computation power, also it may cause a classification algorithm to overfit to training samples and generalize poorly to new samples. Feature extraction is a general term for methods of constructing combinations of the variables to get around these problems while still describing the data with sufficient accuracy.

1) Layers of CNN

a) Representation Of Text Data: As for each word in the medical text, we use the distributed representation of Word Embedding in natural language processing, i.e. the text is represented in the form of vector.

b) Convolution layer of text CNN: choose two words from the front and back of each word vector and Perform convolution operation on weight matrix and word vector. Then we get feature graph.

c) Pool layer of text CNN: Taking the output of convolution layer as the input of pooling layer, we use the max pooling (1-max pooling) operation.

i.e., select the max value of the *n* elements of each row in feature graph matrix.

After max pooling we obtain features. The reason of choosing max pooling operation is that the role of every word in the text is not completely equal, by maximum pooling we can choose the elements which play key role in the text. In spite of different length of the input training set samples, the text is converted into a fixed length vector after convolution layer and pooling layer, for example, in this experiment, after convolution and pooling, we get 100 features of the text.

d) Full connection layer of text CNN: Pooling layer is connected with a fully connected neural network.

e) CNN Classifier: The full connection layer links to a classifier.

B. Naive Bayes

Naive Bayes is a simple technique for constructing classifiers: models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set. It is not a single algorithm for training such classifiers, but a family of algorithms based on a common principle: all naive Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable. For example, a fruit may be considered to be an apple if it is red, round, and about 10 cm in diameter. A naive Bayes classifier considers each of these features to contribute independently to the probability that this fruit is an apple, regardless of any possible correlations between the colour, roundness, and diameter features.

C. KNN

In k-NN classification, the output is a class membership. An object is classified by a majority vote of its neighbours, with the object being assigned to the class most common among its k nearest neighbours (k is a positive integer, typically small). Ifk = 1, then the object is simply assigned to the class of that single nearest neighbour. In the classification phase, *k* is a user-defined constant, and an unlabelled vector (a query or test point) is classified by assigning the label which is most frequent among the *k* training samples nearest to that query point.

D. Decision Tree

Decision tree learning is a method commonly used in data mining. The goal is to create a model that predicts the value of a target variable based on several input variables. An example is shown in the diagram at right. Each interior node corresponds to one of the input variables; there are edges to children for each of the possible values of that input variable. Each leaf represents a value of the target variable given the values of the input variables represented by the path from the root to the leaf.

A decision tree is a simple representation for classifying examples. For this section, assume that all of the input features have finite discrete domains, and there is a single target feature called the "classification". Each element of the domain of the classification is called a *class*. A decision tree or a classification tree is a tree in which each internal (non-leaf) node is labeled with an input feature. The arcs coming from a node labeled with an input feature are labeled with each of the possible values of the target or output feature or the arc leads to a subordinate decision node on a different input feature. Each leaf of the tree is labeled with a class or a probability distribution over the classes.

Flow of the project

1) In this paper, for S-data, according to the discussion with doctors and Pearson's correlation analysis, we extract the patient's demographics characteristics and some of the characteristics associated with cerebral infarction and living habits (such as smoking).

2) Then, we obtain a total of patient's features.

3) For Unstructured -data, we first extract words in the text to learn Word Embedding. Then we utilize the independent feature extraction by CNN.

4) For S-data, we use three conventional machine learning algorithms, i.e., Naive Bayesian (NB), K-nearest Neighbour (KNN), and Decision Tree (DT) algorithm [24], [25] to predict the risk of cerebral infarction disease. This is because these three machine learning methods are widely used.

a) *Input:* Patient's data D, Set of Classifiers C

b) *Output:* Disease that a person can be infected with D_i

LebelSet←Φ

i) if (!D is structured)
ii) SD ← Φ // Initialization of structured data
iii) SD = extractFeatures(D) // Data is converted into structured format
iv) end if
v) Else
vi) SD←D
vii) end else
viii) foreach classifier c in
ix) LabelSet←c.classify(SD) //Classification for //disease prediction using each classifier
x) end foreach
xi) LableCount LC ← Φ
xii) foreach label l in lableSet
xiii) LC ← count(l,lableSet) //count of every lable in labelSet
xiv) end foeach
xv) D_i ← lableOf(max(count)) // lable that has max count
xvi) Return D_i

V. SYSTEM ARCHITECTURE

To help foresee whether a patient is experiencing chronic disease or not as indicated by his/her medical history. The input esteem is the attribute value of the patient, which incorporates the patient's close to home data, for example, age, sex, the pervasiveness of side effects, and living propensities (smoking or not) and other structured information and unstructured information. The yield esteem shows whether the patient is experiencing chronic disease or not. For disease hazard, demonstrating the precision of risk expectation relies upon the assorted variety highlight of the doctor's facility information, i.e., the better is the element depiction of the disease, the higher the exactness will be. For some straightforward sickness, e.g., hyperlipidemia, just a couple of highlights of organized information can get a decent depiction of the illness, bringing about genuinely great impact of disease expectation. Be that as it may, for an unpredictable disease, for example, a cerebral infarction, diabetes, hypertension and asthma just utilizing highlight of structured data isn't a decent method to depict the disease. In this way, use the structured data as well as the content information of patients in view of the Support Vector Machine and Naive Bayes (NB) algorithms. In fig. 1, the dataset contains patients' information related to chronic disease. The dataset is been collected from the hospital. With the help of dataset, the accurate prediction of disease can be done. In structured data the prediction of disease is done with the help of symptoms of each chronic disease. The disease prediction is done by NB algorithm. The NB algorithm is useful for predicting the probability of multiple classes based on various attributes. In this the prediction of disease is by 96% based on the symptoms

of chronic diseases like hypertension, diabetes, cerebral infraction and asthma. For Structured data, the system uses a traditional machine learning algorithm, i.e., NB algorithm to predict the disease. NB classification is a simple probabilistic classifier. It requires calculating the probability of feature attributes. For Structured information, framework utilizes conventional machine learning calculation, i.e., NB calculation to anticipate the sickness. NB characterization is a straight forward probabilistic classifier. It requires to figure the likelihood of highlight properties. A NB classifier is a simple probabilistic classifier based on applying Bayes' theorem with strong independent assumption. A more descriptive term for the underlying probability model would be the self-determining feature model. In basic terms, a NB classifier assumes that the presence of a particular feature of a class is unrelated to the presence of any other feature. The NB classifier performs reasonably well, even if the underlying assumption is not true. The advantage of the NB classifier is that it only requires a small amount of training data to estimate the means and variances of the variables necessary for classification. In order to train a Naive Bayes (NB) model for text classification, there is a need to prepare data set. Genetic algorithm includes process of initialization, and then it improves with a repetitive application of mutation, crossover, inversion and selection operations. It requires a genetic representation and fitness function. When some user's data is missing then it is been recovered by genetic algorithm. In unstructured data, if there is missing data which is caused by patient's mistake. Then missing data is been recovered with the genetic algorithm. The unstructured data mainly focuses on the case study and interrogation which are given by doctors. The Recurrent Neural Network (RNN) algorithm is used to extract features of the text. The stop words are been removed from the text data and the features are extracted successfully. After text feature extraction, SVM Classier performs classification on the data, it will predict whether the patient is suffering from chronic disease or not. With the help of RNN, unstructured data is been converted into structured and the prediction of chronic disease is been done. In a traditional neural system, it is expected that all inputs (and outputs) are autonomous of each other. On the off chance that you need to foresee the following word in a sentence you better know which words preceded it. RNNs are called recurrent on the grounds that they play out a similar undertaking for each component of a sequence, with the yield being relied upon the past calculations. Another approach to consider RNNs is that they have a "memory" which catches data about what has been figured up until now. In principle RNNs can make utilization of data in subjectively long arrangements. The textual features are extracted by RNN. In Fig 1. Xt is the input, Ht is the hidden state which is calculated based on the previous hidden state and input of the current state. V,W and U are weight matrices, gh is the activation function, bh is the bias function and Ot is the output.The basic storage architecture of the system is shown in fig.1.[1]

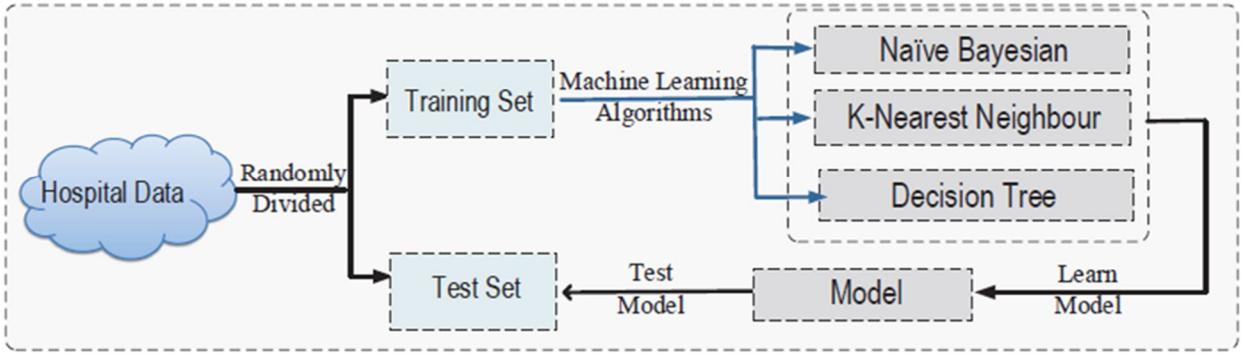


Fig.1 System Architecture

VI. RESULT

Data mining supports many different techniques for knowledge discovery and pre- diction such as classification, clustering, sequential pattern mining, association rule mining and analysis. Data mining is extensively used in business analysis, strategic decision making, financial forecasting, future sales prediction etc. machine learning algorithms are proposed for effective prediction of chronic disease. To extract feature from unstructured data RNN algorithm will be used. Here, user will upload the test file i.e previous health record. RNN algorithm extract the fea- tures from that file . and pass that features to the Nave SVM algorithm for disease predication. System proposes Naive Bayesian algorithm to predict the disease us- ing structured data. System allows user to select the symptoms. System passes that symptoms to the Navie Bayes algorithm to perform disease prediction.

Community question answering system (CQA) is also proposed in this paper. it pre- dicts the question and answers and provides appropriate answers to the users. For that two algorithms are proposed KNN and SVM. KNN algorithm performs feature extraction and classification on questions and SVM algorithm performs classifica- tion on answers. It will help user to find best questions and answers related to the chronic diseases.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.177
Volume 7 Issue VI, June 2019- Available at www.ijraset.com

VII. CONCLUSION

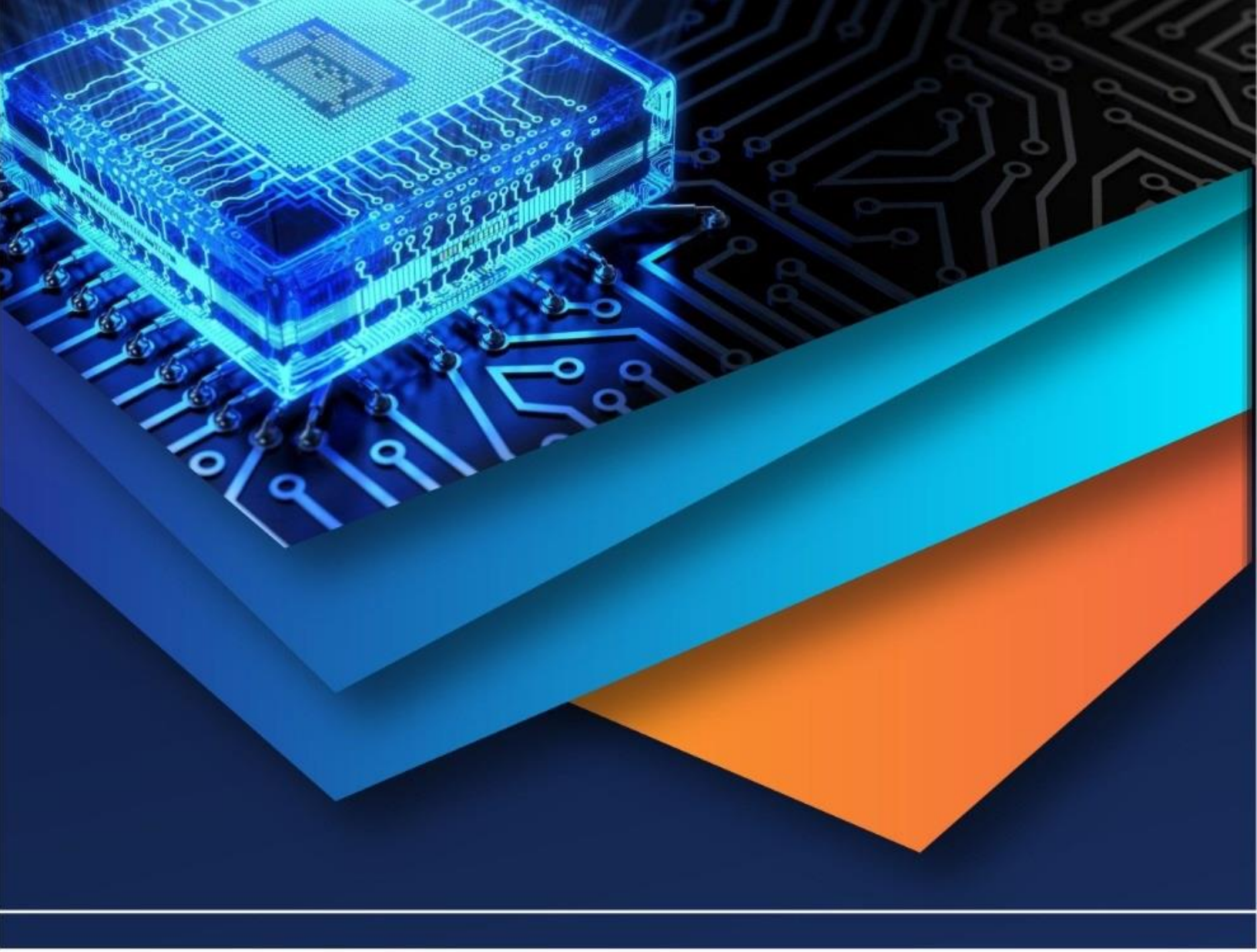
As disease has increased, a new conventional neural network based multimodal disease risk prediction (CNNMDRP) algorithm in which structured and unstructured data from hospital is being used. In this structured and unstructured data, the personal information and detail history of the patient is being stored. In this CNN-MDRP both data are being used for predicting the chronic disease in that particular patient. In unstructured data patients may have missing data. So, the missing data of that particular patient can also retrieve through the genetic algorithm. The featured from unstructured data are been extracted correctly. Then the extracted features are structured data. Both Structured data and extracted structured data are used for predicting the exact chronic disease with Naive Bayes classifier and the SVM classifier. Community question answering system (CQA) is also proposed to help user to post the questions and answers related to the disease. To propose CQA system KNN and SVM algorithms are used.

VIII. ACKNOWLEDGMENT

I would prefer to give thanks the researchers likewise publishers for creating their resources available. I'm conjointly grateful to guide, reviewer for their valuable suggestions and also thank the college authorities for providing the required infrastructure and support.

REFERENCES

[1] Min Chen, Yixue Hao, Kai Hwang, Lu Wang, and Lin Wang, "Disease Prediction by Machine Learning over Big Data from Healthcare Communities", IEEE transaction, 2017, pp 8869-8879.
[2] W. Yin and H. Schutze," Convolutional neural network for paraphrase identification", in HLTNAACL,2015, pp. 901-911.
[3] Seema sharma, Jitendra Agarwal, Shikha Agarwal, Sanjeev Sharma, "Machine Learning Techniques for Data Mining: A Survey", in Computational Intelligence and Computing Research, IEEE International Conference on. IEEE, 2013, pp.1-6.
[4] Jensen PB, Jensen LJ, Brunak S, "Mining electronic health records: towards better research applications and clinical care," Nat Rev Genet.2013 Jan; 14(1):75.
[5] L. Qiu, K. Gai, and M. Qiu, "Optimal big data sharing approach for tele-health in cloud computing", in Smart Cloud (Smart Cloud), IEEE International Conference on. IEEE, 2016, pp. 184-189.
[6] Sakshi Gupta,Pooja Navali,Amruta Sao, Savita Bhat"A Survey On Disease Prediction Using Machine Learning From Healthcare Communities", in JETIR,2018, pp. 2349-5162.
[7] Siwei Lai,Xu Kang Liu,Jun Zhao, "Recurrent Convolutional Neural Networks for Text Classification", in proceeding of the twenty-ninth AAAI Conference on Artificial Intelligence 2015.
[8] Xingyou Wang, Weijie Jiang, Zhiyong Luo, "Combination of Convolutional and Recurrent Neural Network for Sentimental Analysis of Short Texts", International Conference on Computational Linguistics: technical papers, 2016, pg 2428-2437



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY