Disease Prediction In Indian Women By Machine Learning Over Big Data From Healthcare Communities

[1]T Praveena, [2]J Swami Naik
[1]A PG Scholar, [2]Associate Professor
[1]Department of CSE,
G. Pulla Reddy Engineering College, Kurnool. 518007. Andhra Pradesh, India

Abstract : Disease prediction has long been thought to be a crucial topic. Artificial intelligence and machine learning techniques have already been developed to resolve this kind of medical care problem. Recently, neural network ensembles have been successfully used in a reasonably applications together with to help in diagnosing. Artificial intelligence with NN (neural network) can significantly improve the generalization ability of learning systems through training a finite vary of neural networks so combining their results. However, the performance of multiple classifiers in disease prediction is not whole understood. The key purpose of this study is to analyze the assorted factors those impact the Indian woman using CNN-MDP and using totally different classifiers additionally, we have a tendency to use varied evaluation criteria to look at the performance of those classifiers with real-life datasets. We have a tendency to experiment the changed prediction models over real-life hospital data collected from major states of India in 2014, 2015, 2016 and 2017.

IndexTerms **- Big data analytics, machine learning, healthcare, Disease Prediction in Indian Woman**

1 INTRODUCTION

The idea of big data is not new; however, the approach it's outlined is systematically ever-changing. varied tries at defining big data primarily characterize it as a group of data parts whose size, speed, type, and/or quality want one to hunt, adopt, and invent new hardware and software package mechanisms for archiving, analyzing and displaying data successfully. attention might be a prime-rate} example of but the three V's of data first is velocity, second is variety, and third one is volume are an innate side of the data it produces. This data is unfold among multiple healthcare systems, health insurers, researchers, government entities, so forth. Additionally, all of these repositories of data is siloed and more incapable of providing a platform for global data transparency. To feature to the three V's, the veracity of healthcare data is to boot very important for its meaningful use towards developing modification of location analysis. With the event of big data technology, extra attention has been paid to sickness prediction from the angle of big data analysis; varied researches square measure conducted by selecting the characteristics automatically from an oversized type of data to boost the accuracy of risk classification rather than the antecedently chosen characteristics. However, those existing work mostly thought of structured data. For unstructured data, for instance, exploitation convolutional neural network (CNN) to extract text characteristics automatically has already attracted wide attention and to boot achieved excellent results.

However, to the only of our information, none of previous work handles Indian ladies medical text data by CNN. moreover, there is a giant distinction between diseases in numerous regions, primarily due to the varied climate and living habits within the region. Thus, risk classification supported huge data analysis, the subsequent challenges remain: however ought to the missing data be addressed? however ought to the most chronic sicknesss during a bound region and therefore the main characteristics of the sickness within the region be determined? however will big data analysis technology be accustomed analyze the disease and make a stronger model? to unravel these issues, we tend to mix the structured and unstructured data in healthcare field to assess the danger of disease.
Patients' statistical information, test results and disease history are recorded within the EHR, sanctionative U.S.A. to spot potential data-centric solutions to cut back the prices of medical case studies. Wang [9] projected associate degree economical estimating algorithmic rule for the telehealth cloud system and designed a data coherence protocol for the Personal Health Records (PHRs)-based distributed system. Bates. [10] projected six applications of big data within the field of health-care. Qiu et al. [11] projected associate degree optimum big data sharing algorithmic rule to handle the complicate data set in telehealth with cloud techniques. one among the applications is to spot unsound patients, which might be used to cut back medical value since unsound patients usually need costly attention. Moreover, within the 1st paper proposing health-care cyber-physical system [12], it innovatively brought for-ward the idea of prediction-based attention applications, together with health risk assessment. Prediction exploitation ancient sickness risk models sometimes involves a machine learning algorithmic rule (e.g., multivariate analysis and supplying regression, etc.), and particularly a supervised learning algorithmic rule by the employment of coaching data with labels to coach the model [13], [14]. within the take a look at set, patients will be classified into teams of either unsound or low-risk. These models square measure valuable in clinical things and square measure wide studied [15], [16]. However, these schemes have the subsequent characteristics and defects. the info set is usually little, for patients and diseases with specific conditions [17], the characteristics square measure chosen through expertise. However, these pre-selected characteristics perhaps not satisfy the changes within the sickness and its influencing factors.

To solve these issues, we tend to mix the structured and unstructured data in attention field to assess the danger of sickness. First, we tend to used latent issue model to reconstruct the missing data from the medical records collected from a hospital in central China. Second, by exploitation applied math information, we tend to may verify the main chronic diseases within the region. Third, to handle structured data, we tend to consult hospital specialists to extract helpful options. For unstructured text data, we tend to choose the options mechanically exploitation Convolutional Neural Network algorithmic rule.

This paper is explained in below section as stated. we tend to describe the dataset and model in 2nd Section. The strategies utilized in this paper square measure represented in 3rd Section.

2 DATASET AND MODEL DESCRIPTION

In this section, we describe the hospital datasets which we use in the study. In addition, we are explaining disease risk prediction model and evaluation methods. In this study, we are using the data set (health records) of patients admitted to the hospital in 2014, 2015, 2016 and 2017. Furthermore, the hospitalization of Indian woman resulted by chronic diseases has always been continuously occupying a large part in this area through the statistics of the data. For example, the number of Indian woman patients hospitalized with the chronic diseases of cerebral infarction, hypertension, and diabetes accounted for 6.52% of the total number of patients admitted to the hospital in 2016, while the other diseases occupied a small proportion.

2.1 Hospital Data

The hospital dataset used in this study contains real-life hospital data, and therefore the data are stored within the data center. the data provided by the hospital embrace medication, diagnosing and illness varieties. we have a tendency to use a four year data set from 2014 to 2017. Our data concentrate on Indian women, including 1M (One Million) hospitalized patients with 10,000,000 records in total. The inmate department data is especially composed of structured and unstructured text data. The structured data includes laboratory data and therefore the patient's basic info like the patient's age, gender and life habits, etc. whereas the unstructured text data includes the patient's narration of his/her illness. As shown in Table I, the real-life hospital data collected from major states of india are classified into 2 classes, i.e., structured data and unstructured text data.

In order to present out the most disease that have an effect on this region, we've created a statistics on the amount of patients, the sex ratio of patients and therefore the major disease during this region once a year from the structured and unstructured text data, the applied math results are as shown in Table II. From Table II, we are able to get that the proportion of rural and urban patients hospitalized annually have very little distinction and additional patients

2.2 Hospital Data Attributes

In our data set, there are 203 various attributes consisting age, address, pills, illness type, diagnosed for, disease etc.
We can broadly categorize the attributes into three types.

1. Geography details
    a. Address
    b. Rural/Urban
2. Medication Details
    a. Diagnosis details
    b. Illness details
    c. Disability details
3. Food Habits
    a. Type of Food
    b. Number of times per day
4. Education or Professional details
    a. Education
    b. Job details
    c. Commute information
5. Social Details
    a. Cast/Religion
6. Pregnancy details
    a. No.of Kids
    b. Maternity type

Among these attributes, there are many unused or irrelevant attributes. Therefore, in cleansing stage, we have removed all the irrelevant attributes. We will introduce machine learning and deep learning algorithms to classify these data sets. For S-data, we use three conventional machine-learning algorithms, i.e., Naive Bayesian (NB), K-nearest Neighbor (K-NN), 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 [26]. For T-data, we pro-pose CNN-based unimodal disease risk prediction (CNN-UDP) algorithm to predict the risk of cerebral infarction disease. In the remaining of the paper, let CNN-UDP (T-data) denote the K-NN algorithm used for T-data. For S&T data, we predict the risk of cerebral infarction disease by the use of CNN-MDP algorithm, which is denoted by CNN-MDP(S&T-data) for the sake of simplicity. In the following section, the details about CNN-UDP (T-data) and CNN-MDP(S&T data) will be given.

Statistics	2014	2015	2016	2017
Number of patients	1,425,000	1,769,000	3,418,000	3,388,000
Andhra Pradesh	118,750	147,417	284,833	282,333

Table-I: Data Analysis

2.3 Disease Risk Prediction

From Table I, we have a tendency to get the most chronic disease during this region. The goal of this study is to predict whether or not a woman patient is amongst the cerebral infarct speculative population in line with their anamnesis. a lot of formally, we have a tendency to regard the chance prediction model for cerebral infarct because the supervised learning strategies of machine learning.

For dataset, in line with the various characteristics of the patient and also the discussion with doctors, we'll target the subsequent 3 datasets to achieve a conclusion.
Structured information (S-data): use the patient's structured information to predict whether or not the patient is at speculative of cerebral infarct.
Text information (T-data): use the patient's unstructured text information to predict whether or not the patient is at speculative of cerebral infarct.
In the experiment setting and dataset characteristics, we have a tendency to choose 1,000,000 patients in total because the experiment information and at random divided the information into coaching information and check data.

3 RELATED WORK

For the performance analysis within the experiment. First, we have a tendency to denote TP, FP, American state and FN as true positive (the variety of instances properly foretold as required), false positive (the variety of instances incorrectly foretold as required), true negative (the variety of instances properly foretold as not required) and false negative (the variety of instances incorrectly foretold as not required), severally.

In this section, we have a tendency to introduce the info imputation, CNN-based unimodal malady risk prediction (CNN-UDP) algorithmic program and CNN-based unimodal malady risk prediction (CNN-MDP) algorithmic program.

3.1 K-nearest neighbors

The K-nearest neighbors (K-NN) algorithmic program utilized in this project is additionally from R Library, that provides each unsupervised and supervised neighbors-based learning strategies.
Despite the simplicity of the algorithmic program, K-NN has been booming during a sizable amount of classification and regression issues.
Attribute scaling is additionally performed before victimisation K-NN, to confirm that the gap live accords equal weight to every variable.

3.2 NaiveBayes Model

Naive Bayes could be a easy multiclass classification algorithmic program with the idea of independence between each combine of options. Naive Bayes are often trained terribly with efficiency. among one pass to the coaching information, it computes the probability distribution of every feature given label, so it applies Bayes' theorem to calculate the probability distribution of label given associate degree observation and use it for prediction. These models ar generally used for document classification. among that context, every observation could be a document and every feature represents a term whose price is that the frequency of the term (in multinomial naive Bayes) or a zero or one indicating whether or not the term was found within the document (in Bernoulli naive Bayes). Feature values should be plus. The model sort is chosen with associate degree ex gratia parameter "multinomial" or "bernoulli" with "multinomial" because the default. Additive smoothing are often employed by setting the parameter λλ (default to one.01.0). For image classification, the input feature vectors ar sometimes thin, and thin vectors ought to be provided as input to require advantage of exiguity. Since the coaching information is barely used once, it's not necessary to cache it.

4 DATA IMPUTATION

For patient's examination information, there's an oversized variety of missing information thanks to human error. Thus, we want to fill the structured information. Before information imputation, we have a tendency to 1st determine unsure or incomplete medical information so modify or delete them to boost the info quality. Then, we have a tendency to use information integration for information pre-processing. we are able to integrate the medical information to ensure information atomicity: i.e., we have a tendency to integrated the peak and weight to get body mass index (BMI). For information imputation, we have a tendency to use the latent issue model [27] that is given to elucidate the evident variables in terms of the latent variables. consequently, assume that Rm n is that the information matrix in our health care model. The row designation, m represents the entire variety of the patients, and also the column

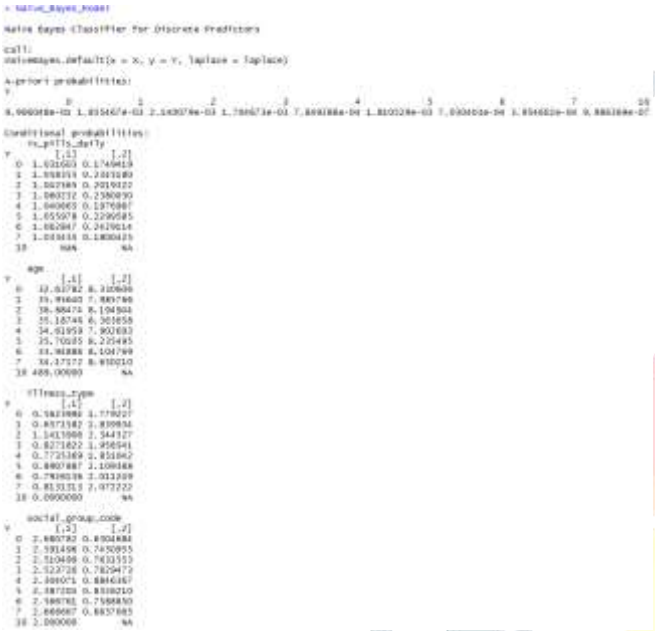
5 EXPERIMENTAL RESULTS

In this section, we discuss the performance of Naïve Bayes, K-NN and Random forest algorithms.

5.1 Run Time Comparison

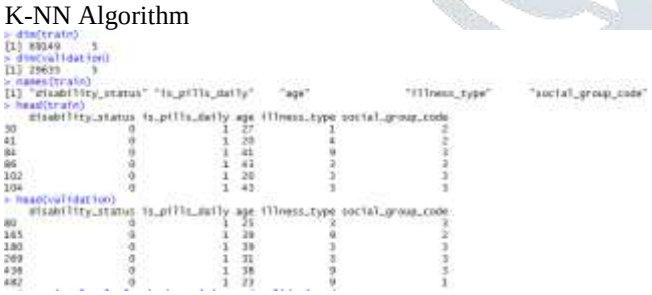
We compare the running time of K-NN (T-data) and CNN-MDP (S&T-data) algorithms in personal computer (2core CPU, 8.00G RAM) and data center (6core*2*7D84core CPU, 48*7D336G RAM). Here, we set the same Convolutional Neural Network iterations, which are 100 and extract the same 100 text features. As shown in Fig. 2, for K-NN (T-data) algorithm, the running time in data center is 178.5s while the time in personal computer is 1646.4s. For CNN-MDP (S&T-data) algorithm, its running time in data center is 178.2s while the time in personal computer is 1637.2s. That is, the running speed of the data center is 9.18 times on the personal computer. Moreover, we can see the running time of K-NN (T-data) and CNN-MDP (S&T-data) are basically the same from the figure, i.e. although the number of CNN-MDP (S&T-data) features increase after adding structured data, it does not make a significant change in time. The later experiments are based on the running results of the data center.

5.2 Naïve Bayes Algorithm:

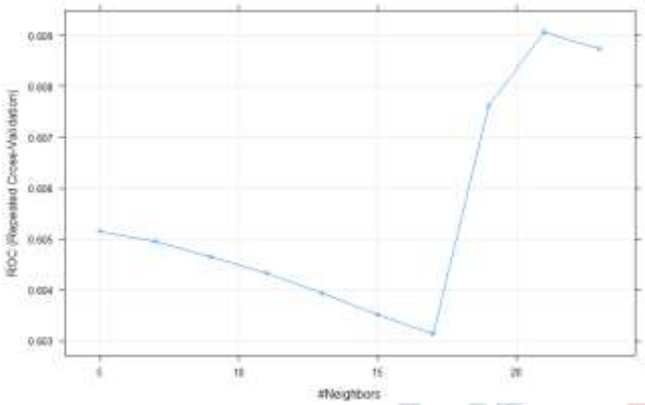


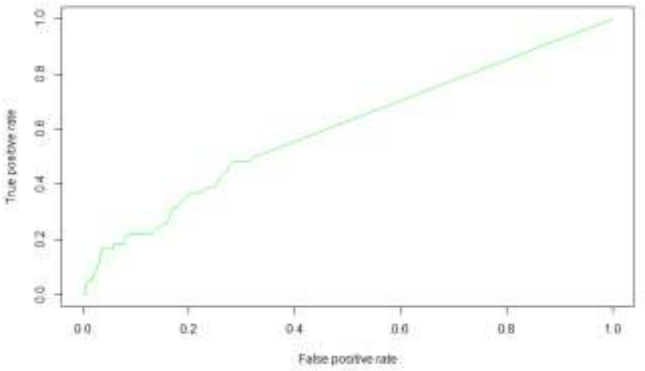
```
> Naive_Bayes_Model

Naive Bayes Classifier for Discrete Predictors

Call:
naiveBayes.default(x = X, y = Y, laplace = laplace)

A-priori probabilities:
Y
           0            1            2           27
6.757276e-02 1.826047e-01 7.498215e-01 9.986369e-07

Conditional probabilities:
    is_pills_daily
Y        [,1]      [,2]
  0  1.033994 0.1812177
  1  1.030723 0.1725662
  2  1.031987 0.1759654
  27       NaN        NA

    age
Y        [,1]     [,2]
  0   31.15654 8.969267
  1   26.67365 5.490665
  2   34.26140 8.158293
  27 489.00000       NA

    illness_type
Y        [,1]     [,2]
  0  0.4608291 1.609351
  1  0.5476008 1.786014
  2  0.5790077 1.795935
  27 0.0000000       NA

    social_group_code
Y        [,1]      [,2]
  0  2.666074 0.6466198
  1  2.677583 0.6304534
  2  2.680603 0.6315503
  27 2.000000        NA


> NB_Predictions=predict(Naive_Bayes_Model,dfa1)
> table(NB_Predictions,dfa1$outcome_pregnancy)

NB_Predictions      0      1      2     27
             0      0      0      0      0
             1     34    140    197      0
             2  67631 182714 750648      1
             27     0      0      0      0
```

K-NN Algorithm



```
> model1
k-Nearest Neighbors

69149 samples
    4 predictor
    2 classes: 'X0', 'X1'

Pre-processing: centered (4), scaled (4)
Resampling: Cross-Validated (10 fold, repeated 3 times)
Summary of sample sizes: 62234, 62235, 62235, 62233, 62234, 62233, ...
Resampling results across tuning parameters:

  k   ROC        Sens  Spec
   5  0.6051488  1     0
   7  0.6049478  1     0
   9  0.6046427  1     0
  11  0.6043311  1     0
  13  0.6039352  1     0
  15  0.6035111  1     0
  17  0.6031279  1     0
  19  0.6076119  1     0
  21  0.6090699  1     0
  23  0.6087303  1     0

ROC was used to select the optimal model using the largest value.
The final value used for the model was k = 21.
```






```
> ks <- max(attr(perf_val, "y.values")[[1]] - (attr(perf_val, "x.values")[[1]]))
> ks
[1] 0.2001522
```

5.3 Effect Of Sliding Window

When taking convolution of CNN, we need to confirm the number of words for sliding window first. In this experiment, the selected number of words for the sliding window are 1, 3, 5, 7 and 9. The iterations of Convolutional Neural Network are 200 and the size of convolution kernel is 100. As shown in Fig. 3, when the number of words for the sliding window are 7, the accuracy and recall of K-NN (T-data) algorithm are 0.95 and 0.98, respectively. And the accuracy and recall of CNN-MDP (S&T-data) algorithm are 0.95 and 1.00. These results are all higher than we choose other number of words for sliding window. Thus, in this paper, we choose the number of words for sliding window are 7.

5.4 Effect Of Iterations

We give out the change of the training error rate and test accuracy along with the number of iterations. As shown in Fig. 4, with the increase of the number of iterations, the training error rate of the K-NN (T-data) algorithm decreases gradually, while test accuracy of this method increases. The CNN-MDP (S&T-data) algorithm have the similar trend in terms of the training error rate and test accuracy. In Fig. 4, we can also obtain when the number of iterations are 70, the training process of CNN-MDP (S&T-data) algorithm is already stable while the K-NN (T-data) algorithm is still not stable. In other words, the training time of MDP(S&T data) algorithm is shorter, i.e. the convergence speed of CNN-MDP (S&T-data) algorithm is faster.

5.5 Effect of Text Features

The number of features extracted from structured data is certain, i.e. 79 features. However, the feature number of unstructured text data extracted by Convolutional Neural Network is uncertain. Thus, we research the effect of text feature number on accuracy and recall of K-NN (T-data) and CNN-MDP (S&T-data) algorithms. We extract 10; 20; ; 120 features from text by using CNN. Fig. 5 shows the accuracy and recall of each feature after it go through 200 times of iteration. From the Fig. 5(a) and Fig. 5(b), when the feature number of text is smaller than 30, the accuracy and recall of K-NN (T-data) and CNN-MDP (S&T-data) algorithms are smaller than the feature number of text is bigger than 30 obviously. This is because it is not able to describe a large number of useful information contained in the text when the text feature number is relatively small. Moreover, in the Fig. 5(a), the accuracy of CNN-MDP (S&T-data) algorithm is more stable than K-NN (T-data) algorithm, i.e. the CNN-MDP (S&T-data) algorithm is reduced fluctuation after adding structured data. As shown in Fig. 5(b), after adding structured data, the recall of CNN-MDP (S&T-data) algorithm is higher than K-NN (T-data) algorithm obviously. This shows that the recall of algorithm is iMPRoved after adding structured data.

6 Analysis Of Overall Results

In this section, we describe the overall results about S-data and S&T-data.

6.1 Structured Data (S-Data)

For S-data, we use traditional machine learning algorithms, i.e., NB, K-NN and DT algorithm to predict the risk of cerebral infarction disease. NB classification is a simple probabilistic classifier. It requires to calculate the probability of feature attributes. In this experiment, we use conditional probability formula to estimate discrete feature attributes and Gaussian distribution to estimate continuous feature attributes. The K-NN classification is given a training data set, and the closest k instance in the training data set is found. For K-NN, it is required to determine the measurement of distance and the selection of k value. In the experiment, the data is normalized at first. Then we use the Euclidean distance to measure the distance. As for the selection of parameters k, we find that the model is the best when k D 10. Thus, we choose k D 10. We choose classification and regression tree (CART) algorithm among several decision tree (DT) algorithms.

To determine the best classifier and improve the accuracy of the model, the 10-fold cross-validation method is used for the training set, and data from the test set are not used in the training phase. The model's basic framework is shown in Fig. 6. The results are shown in Fig. 7(a) and Fig. 7(b). From Fig. 7(a), we can see that the accuracy of the three machine learning algorithms are roughly around 50%. Among them, the accuracy of DT which is 63% is highest, followed by NB and K-NN. The recall of NB is 0.80 which is the highest, followed by DT and K-NN. We can also draw from Fig. 7(b) that the corresponding AUC of NB, K-NN and DB are 0.4950, 0.4536 and 0.6463, respectively. In summary, for S-data,

6.2 Structured and Text Data (S&T-Data)

According to the discussion in Section IV, we give out the accuracy, precision, recall, F1-measure and ROC curve under K-NN (T-data) and CNN-MDP (S&T-data) algorithms. In this experiment, the selected number of words is 7 and the text feature is 100. As for K-NN (T-data) and CNN-MDP (S&T-data) algorithms, we both run 5 times and seek the average of their evaluation indexes. From the Fig. 8, the accuracy is 0.9420 and the recall is 0.9808 under K-NN (T-data) algorithm while the accuracy is 0.9480 and the recall is 0.99923 under CNN-MDP (S&T-data) algorithm. Thus, we can draw the conclusion that the accuracy of K-NN (T-data) and CNN-MDP (S&T-data) algorithms have little difference but the recall of CNN-MDP (S&T-data) algorithm is higher and its convergence speed is faster. In summary, the performance of CNN-MDP (S&T-data) is better than K-NN (T-data).

7 Conclusion

In this paper, we analyzed Indian woman patients data to predict the pregnancy risk for long term diseases using structured and unstructured data from Indian hospitals. To the best of our data, none of the existing work focused on Indian Woman disease prediction with respect to pregnancy. Compared to several typical prediction algorithms, the prediction accuracy of our proposed algorithm reaches 93.8% with a convergence speed which is faster than that of the CNN-based unimodal disease risk prediction (CNN-UDP) algorithm

References

[1] P. Groves, B. Kayyali, D. Knott, and S. van Kuiken, The`Big Data'Revolution in Healthcare: Accelerating Value and Innovation. USA: Center for US Health System Reform Business Technology Of ce, 2016.
[2] M. Chen, S. Mao, and Y. Liu, ``Big data: A survey," Mobile Netw. Appl., vol. 19, no. 2, pp. 171 209, Apr. 2014.
[3] P. B. Jensen, L. J. Jensen, and S. Brunak, ``Mining electronic health records: Towards better research applications and clinical care," Nature Rev. Genet., vol. 13, no. 6, pp. 395 405, 2012.
[4] D. Tian, J. Zhou, Y. Wang, Y. Lu, H. Xia, and Z. Yi, ``A dynamic and self-adaptive network selection method for multimode communications in heterogeneous vehicular telematics," IEEE Trans. Intell. Transp. Syst., vol. 16, no. 6, pp. 3033 3049, Dec. 2015.
[5] M. Chen, Y. Ma, Y. Li, D. Wu, Y. Zhang, and C. Youn, ``Wearable 2.0: Enable human-cloud integration in next generation healthcare system," IEEE Commun., vol. 55, no. 1, pp. 54 61, Jan. 2017.
[6] M. Chen, Y. Ma, J. Song, C. Lai, and B. Hu, ``Smart clothing: Con-necting human with clouds and big data for sustainable health monitor-ing," ACM/Springer Mobile Netw. Appl., vol. 21, no. 5, pp. 825 845, 2016.
[7] M. Chen, P. Zhou, and G. Fortino, ``Emotion communi-
[8] cation system," IEEE Access, vol. 5, pp. 326 337, 2017, doi: 10.1109/ACCESS.2016.2641480.
[9] M. Qiu and E. H.-M. Sha, ``Cost minimization while satisfying hard/soft timing constraints for heterogeneous embedded systems," ACM Trans. Design Autom. Electron. Syst., vol. 14, no. 2, p. 25, 2009
[10] J. Wang, M. Qiu, and B. Guo, ``Enabling real-time information service on telehealth system over cloud-based big data platform," J. Syst. Archit., vol. 72, pp. 69 79, Jan. 2017.

[11] D. W. Bates, S. Saria, L. Ohno-Machado, A. Shah, and G. Escobar, ``Big data in health care: Using analytics to identify and manage high-risk and high-cost patients," Health Affairs, vol. 33, no. 7, pp. 1123 1131, 2014.
[12] L. Qiu, K. Gai, and M. Qiu, ``Optimal big data sharing approach for tele-health in cloud computing," in Proc. IEEE Int. Conf. Smart Cloud (Smart-Cloud), Nov. 2016, pp. 184 189.
[13] Y. Zhang, M. Qiu, C.-W. Tsai, M. M. Hassan, and A. Alamri, ``Health-CPS: Healthcare cyber-physical system assisted by cloud and big data," IEEE Syst. J., vol. 11, no. 1, pp. 88 95, Mar. 2017.
[14] K. Lin, J. Luo, L. Hu, M. S. Hossain, and A. Ghoneim, ``Localization based on social big data analysis in the vehicular networks," IEEE Trans. Ind. Informat., to be published, doi: 10.1109/TII.2016.2641467.
[15] K. Lin, M. Chen, J. Deng, M. M. Hassan, and G. Fortino, ``Enhanced ngerprinting and trajectory prediction for iot localization in smart build-ings," IEEE Trans. Autom. Sci. Eng., vol. 13, no. 3, pp. 1294 1307, Jul. 2016.
[16] D. Oliver, F. Daly, F. C. Martin, and M. E. McMurdo, ``Risk factors and risk assessment tools for falls in hospital in-patients: A systematic review," Age Ageing, vol. 33, no. 2, pp. 122 130, 2004.
[17] S. Marcoon, A. M. Chang, B. Lee, R. Salhi, and J. E. Hollander, ``Heart score to further risk stratify patients with low TIMI scores," Critical Pathways Cardiol., vol. 12, no. 1, pp. 1 5, 2013.
[18] S. Bandyopadhyay et al., ``Data mining for censored time-to-event data: A Bayesian network model for predicting cardiovascular risk from elec-tronic health record data," Data Mining Knowl. Discovery, vol. 29, no. 4, pp. 1033 1069, 2015.
[19] B. Qian, X. Wang, N. Cao, H. Li, and Y.-G. Jiang, ``A relative similarity based method for interactive patient risk prediction," Data Mining Knowl. Discovery, vol. 29, no. 4, pp. 1070 1093, 2015.
[20] A. Singh, G. Nadkarni, O. Gottesman, S. B. Ellis, E. P. Bottinger, and J. V. Guttag, ``Incorporating temporal EHR data in predictive models for risk strati cation of renal function deterioration," J. Biomed. Inform., vol. 53, pp. 220 228, Feb. 2015.
[21] J. Wan et al., ``A manufacturing big data solution for active preven-tive maintenance," IEEE Trans. Ind. Informat., to be published, doi: 10.1109/TII.2017.2670505.
[22] [21] W. Yin and H. Schutze, ``Convolutional neural network for paraphrase identi cation," in Proc. HLT-NAACL, 2015,901 911.
[23] N. Nori, H. Kashima, K. Yamashita, H. Ikai, and Y. Imanaka, ``Simul-taneous modeling of multiple diseases for mortality prediction in acute hospital care," in Proc. 21th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2015, pp. 855 864.
[24] S. Zhai, K.-H. Chang, R. Zhang, and Z. M. Zhang, ``Deepintent: Learning attentions for online advertising with recurrent neural networks," in Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2016,1295 1304.
[25] K. Hwang and M. Chen, Big Data Analytics for Cloud/IoT and Cognitive Computing. Hoboken, NJ, USA: Wiley, 2017.
[26] H. Chen, R. H. Chiang, and V. C. Storey, ``Business intelligence and analytics: From big data to big impact," MIS Quart., vol. 36, no. 4,
[27] 1165 1188, 2012.
[28] S. Basu Roy et al., ``Dynamic hierarchical classi cation for patient risk-of-readmission," in Proc. 21th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2015, pp. 1691 1700
[29] J. C. Ho, C. H. Lee, and J. Ghosh, ``Septic shock prediction for patients with missing data," ACM Trans. Manage. Inf. Syst., vol. 5, no. 1, p. 1, 2014. (2015). Ictclas. [Online]. Available: http://ictclas.nlpir.org/(2013). Word2vec. [Online]. Available: https://code.google.com/p/ word2vec/
[30] Y.-D. Zhang et al., ``Fractal dimension estimation for developing patholog-ical brain detection system based on minkowski-bouligand method," IEEE Access, vol. 4, pp. 5937 5947, 2016.
[31] Y.-D. Zhang et al., ``Facial emotion recognition based on biorthogonal wavelet entropy, fuzzy support vector machine, and strati ed cross vali-dation," IEEE Access, vol. 4, pp. 8375 8385, 2016.
[32] S.-H. Wang et al., ``Multiple sclerosis detection based on biorthogonal wavelet transform, RBF kernel principal component analysis, and logistic regression," IEEE Access, vol. 4, pp. 7567 7576, 2016.
[33] S.-M. Chu, W.-T. Shih, Y.-H. Yang, P.-C. Chen, and Y.-H. Chu, ``Use of tra-ditional Chinese medicine in patients with hyperlipidemia: A population-based study in Taiwan," J. Ethnopharmacol., vol. 168, pp. 129 135, Jun. 2015.
[34] Min Chen, Yixue Hao, Kai Hwang, Lu Wang, and Lin Wang `` Disease Prediction by Machine Learning Over Big Data From Healthcare Communities," IEEE Access,., vol.5, pp. 8869 8879, June 28, 2017.

AUTHORS

T. Praveena, B.Tech in CSE in 2011 form Vishnu Institute of Technology, JNTUK, Kakinada. Now pursuing M.Tech in CSE in G. Pulla Reddy Engineering College. Her research interests include parallel processing and big data analysis.

J. Swami Naik, B.Tech in CSE in 2001 with first class from G. Pulla Reddy Engg. College, Sri Krishna Devaraya University, Anatapurramu. M.Tech in CSE in 2003 with first class from IIT Guwahati. Selected for Ph.D in CSE from JNTUA, Anatapuramu. Has 14.5 years of teaching experience as associate professor, GPREC from July 2008 to till date