Prediction of Chronic Disease by Machine Learning

Anandajayam.P
Department of computer science
Manakula Vinayagar Institute of Technology
Puducherry ,India
anandajayam@gmail.com

Aravindkumar.S
Department of computer science
Manakula Vinayagar Institute of Technology
Puducherry ,India
aravind.kumar368@gmail.com

Arun.P
Department of computer science
Manakula Vinayagar Institute of Technology
Puducherry ,India
arunpari19@gmail.com

Ajith.A
Department of computer science
Manakula Vinayagar Institute of Technology
Puducherry ,India
ajithkingmaker7@gmail.com

Abstract-- In the field of biomedical and healthcare communities the accurate prediction plays the major role to find out the risk of the disease in the patient[15]. The prediction gives the benefits of early disease detection. However the analysis accuracy as the relationship with the condition of the medical data, thus the poor condition of the medical data[4] leads to less accuracy of prediction. Here we use a certain machine learning algorithm to state the rate of disease. Prediction process is done using the dataset provided online from certain hospitals, the entire dataset will be preprocessed and the missing values will be reconstructed. Compared to several types of prediction algorithms, the RNN algorithm[21]gives the highest accuracy of prediction around 97.6% with a convergence speed.

Keywords-- Big data analytics, Machine Learning, Chronic Disease, Health care prediction.

I. Introduction

According to the medical report, the people death rate is high due to the chronic disease and over 70% of the patient's money is spent over chronic disease treatment. Chronic disease is a serious problem in healthcare all over the world[14]. According to the medical standard report, chronic disease is the main reason for death. Therefore the analysis of chronic disease is more important to reduce the risk of people life. With the growth in the field of medial, the patient's data collection[16] is more convenient. The patient's information like demographics, lab test result and other medical histories of the disease in form of EHR which acts as a central data, which provides [13] solution to reduce the cost of medical case studies. The electronic health records(EHR)[14] is the digital format of the patient's details, the EHR contains the test result, statistical data information, and the disease history record.

Since big data is a growing technology they are given some attention for disease prediction using big data analytics[3], where accuracy in prediction the threat of disease is improved[20]. Nowadays data from healthcare is large, thus from understanding the what kind of data it is further processing over is made. Big data analytics plays a major role handling and processing the data in a better manner. Disease in each region may vary based on their habitat of living conditions and threat over disease in such regions may also vary. Some of the challenges in big data are addressed. How the disease can be analyzed with the help of big data [14]analysis? How the correct model can be chosen in relation to the big data analysis?

To solve the problem both structured and the unstructured data is combined in the field of healthcare to determine the threat over disease. First, we pre-processed the data to find out the missing values and the missing values[3] are reconstructed that helps the prediction to define a high accuracy rate. Second, to handle the structured data the RNN algorithm is used and to handle the unstructured data the CNN algorithm is used. Finally, we processed multiple algorithms to explore the accuracy rate of the risk by chronic disease. The models are processed with the combination of both structured and unstructured data of hospital records[2]. From this experiment, we come to a conclusion based performance of RNN[12] algorithm comparatively with other existing algorithms.

II. DATASET AND MODEL DESCRIPTION

Here we are going to have a brief description of the dataset used in this experiment and also a brief study about the prediction model which is used for prevention from disease using this model

A. Dataset description

The data used in our experiment contains real-life data. To protect a patient's privacy we are not using the patient's personal information like name, id, location etc[2]. The dataset contains the structure data and in addition to that, the unstructured text data is included for the better accuracy rate in risk prediction. The dataset contains around 43400 records of data which is enough to get an accurate prediction result. The structured data includes the fundamental information about the patient such as the demographics of a patient and living habitat of a patient and laboratory data information. The unstructured data

@IEEE 978-1-7281-1524-5

consists of text data which is of patients narration over the doctor about the disease and symptoms.

In this experiment mainly focus on cerebral infarction from the provided dataset, where the dataset list of patients with all disease. By using supervised only patients with chronic disease are first segregated. We mainly focus on cerebral infarction since it is a fatal disease, which leads to high death rate all over the world when compared from other diseases.consists of text data which is of patients narration over the doctor about the disease and symptoms.

B. Prediction Model description

From Table II we can predict whether the patient having the high threat of cerebral infarction or not using the historical data of the patient. This can be done using the supervised learning one of the methods comes under machine learning. In supervised learning[13] where the input is known, such as demographics and other details of the patient, X= {X1, X2, X3...XN}.The output consists of patients threat over the disease is high or not such as O={O0, O1}. Here O0 shows that the patient has a high percentage of getting cerebral infarction, O1 shows that the patient has a high percentage of getting cerebral infarction[18]. Introduction about a dataset, experimental setting, characteristics of datasets and learning algorithms.

From the dataset, we came to know about the characteristics of the patient and doctor discussion with a patient from know information we will focus on two data to obtain a defined output.

STRUCTURED DATA: The structured data consists of defined data in a table format such as Demographics[3] and other details of the patient, from which the patient high percent of cerebral infarction is predicted. Normally data which are given in the format of row and column are known as structured data in terms of normal understanding.

UNSTRUCTURED DATA: The unstructured data consists of text data which may be of patient report or patients narration of his symptoms, from which the patient high percent of cerebral infarction is predicted. It is known for the additional information for the prediction model which helps to make prediction much better and accuracy rate can be increased.

In this experiment we select data of 80 percent for training is known as train data and other data of 20 percent for testing is known as test data, in a ratio of [8:2]. If there are a total of 100 records, 80 percent of record is taken as train data and rest 20 percent of data as test data for testing.

We will make use of certain machine learning as well as deep learning algorithms briefly. For structured data, we can use certain supervised learning algorithms such as Naive Bayes, Support Vector Machine[23], K-nearest neighbor [9][10], Random forest classifiers Decision Tree Classifiers[5][6], Logistic Regression[22] and Random Forest and some of the other algorithms are also used to predict the threat of cerebral infarction. Unstructured Text Data we use Natural Language Processing (NLP) to predict the high threat of cerebral infarction. As said before text data is the narration of patient over doctor which is processed using the natural language processing for understanding the words used by a normal human in his day to day life.

TABLE I

ATTRIBUTES SET IN HOSPITAL DATA

Data category	Item	Description
Structured data	Demographics of the patient Living habits Examination items and results Diseases	Patient's gender, age, height, weight, etc. Whether the patient smokes, has a genetic history, etc. Includes 682 items, such as blood, etc. Patient's disease, such as cerebral infarction, etc.
Unstructured text data	Patient's readme illness Doctor's records	Patient's readme illness and medical history Doctor's interrogation records

The above TABLE I represents the data format used for the prediction. The above information are obtained from the patients from his patients' history such as EHR.

B. Correlation of Dataset

The hospital dataset contains around 12 attributes which are related to the cause of disease and also it deals with the patients' information, thus the relation between the attributes plays the major role in the field of risk prediction. With the help of the relation among the attributes, the system model will learn the data easily and efficiently.

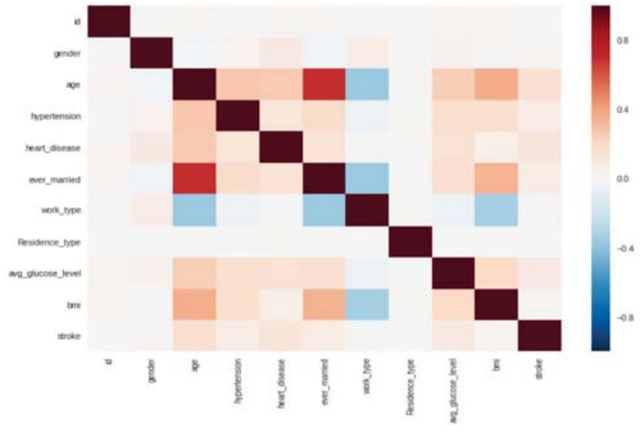


Fig.2 *Heat Map Correlation*

To know the relation between the attributes of the dataset the heatmap correlation graph and the pair plot graph are used to analyze the data correlation. The heatmap graph will explain the relation and correlation of the attributes and the pair plot graph will give us the compression of the attributes one another.

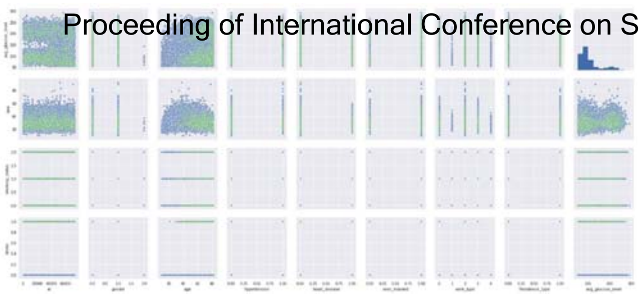


Fig.1 *Pairplot Graph*

C. Model Evaluation

Using this evaluation method accuracy of the model is known. performance metrics of a model is calculated using the confusion matrix, which consists of true positive(TP), false positive(FP), true negative(TN), false negative(FN). We have four metrics such as accuracy, precision, recall, and F1-score are given below.

- TP- how correct the value is predicted as needed
- FP-how incorrect the value is predicted as needed
- TN-how correct the value is predicted a not needed
- FN-how incorrect the value is predicted as needed

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \tag{1}$$

$$Precision = \frac{TP}{TP+FP} \tag{2}$$

$$Recall = \frac{TP}{TP+FN} \tag{3}$$

$$F1 - Score = \frac{2\ X\ Precision\ X\ Recall}{Precision + Recall} \tag{4}$$

We also use another performance metrics such as ROC (Receiver Operating Curve), which is the curve drawn between true positive (TP) and false positive(FP)

$$TPR = \frac{TP}{TP + FN} \tag{5}$$

$$TFR = \frac{FP}{FP + TN} \tag{6}$$

III.Proposed method

Here we will be experimenting some of the machine learning algorithms to get a detailed idea about the methods for the risk prediction.

A. Naive Bayesian (NB)

Bayes strategies are a lot of directed learning calculations dependent on applying Bayes' hypothesis with the "innocent" supposition of restrictive autonomy between each pair of highlights given the estimation of the class variable. Bayes' hypothesis expresses the accompanying relationship, given class variable A and B ward highlight vector x1 through xn.

$$P(A/B) = \frac{P(B/A)\ X\ P(A)}{P(B)} \tag{7}$$

B. K-nearest Neighbour (KNN)

As the name implies in knn algorithm[9] the input train data fixed over X and Y axis and then the test data is fitted over the plots in same X and Y axis .The nearest plotted points from test data with minimum distance are taken and majority of True or False values are taken as the desired output.The nearest K points[10] taken should always be an odd number.

Steps to be followed

- Plot the input
- Test data are given
- Find the distance of each point with each test data
- Take the 'K' nearest point

Assign the mode of the nearest point as output

$$D = \sqrt{(X1 - Y1)^2 + (X2 - Y2)^2 + ... + (Xn - Yn)^2} \tag{8}$$

C. Decision Tree (DT)

It is a Segregation of given input based upon their features. From the given input the output is classified based on the maximum depth of the problem[5]. The depth can be given manually, but to find the best depth they run as a loop and best depth with high accuracy is taken. Some times there is a chance of overfitting and underfitting of a model can happen according to the maximum depth of the model.

Under fitting[6] takes place when there is no proper information is provided.

Over fitting takes place when model try to learn more data from the input.

D. Support Vector Machine (SVM)

SVM algorithm makes an axis over the plots and separates different features obtained from the plain. According to a number of different features present they separated respectively. When the two different objects are placed over the plane then they are separated according to their features extraction.

Types Of SVM

- Linear
- Poly
- RBF

SVM Type 1

Linear method a linear line is placed from which different features are extracted using the given formula,

$$\tfrac{1}{2} w^T w + C \sum_{i=1}^{N} \varsigma_i \tag{9}$$

SVM Type 2

Poly method a circular or irregular shape is drawn over the x and y axis and features are extracted using the given formula,

$$\frac{1}{2}w^{T}w - \nu\rho + \frac{1}{N}\sum_{i=1}^{N}\varsigma_{i} \qquad (10)$$

E. Recurrent Neural Network (RNN)

The basic idea of recurrent networks is to have loops. These loops are allowed to use information from previous passes in a network[21]. The memory length depends on a number of factors but to note that it is not an indefinite loop. You can think of the memory as degrading, with older information[19] being less and less usable. For example, if we just want the network to do one thing they remember whether an input from earlier was 1, or 0. In a network[21] it is not difficult for a network to remember whether the previous output was 1 around the loop continuously. However every time you send in a 0, then the output can become somewhat lower[21]. After some number of passes the input loop which will be arbitrarily low, making the output of the network as 0.

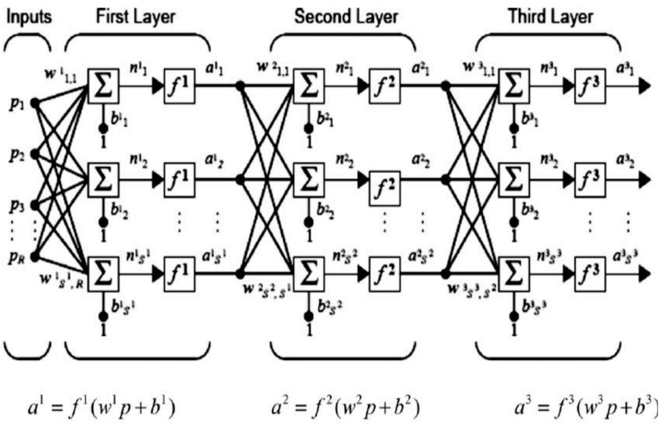
RNN Formula

$$h^{(t)} = f\left(h^{(t-1)}, x^{(t)}; 0\right) \qquad (11)$$

$$L\left(\left\{x^{(1)},...,x^{(T)}\right\}, \left\{y^{(1)},...,y^{(T)}\right\}\right) = \sum_{t} L^{(t)} = \sum_{t} - log\widehat{y}_{y^{(t)}} \qquad (12)$$

h(t) denotes current state
h(t-1) denotes previous state
x(t) denotes current input

Each time output from the previous state is given as the additional input to current state plus normal input is also given. Various hidden layers are given, based on constraints and how typical your model, number of hidden layers are decided.



$$a^1 = f^1(w^1p+b^1) \qquad a^2 = f^2(w^2p+b^2) \qquad a^3 = f^3(w^3p+b^3)$$

$$a^3 = f^3(w^3 f^2(w^2 f^1(w^1 p + b^1) + b^2) + b^3)$$

Fig.3 RNN

IV ANALYSIS OF OVERALL RESULT

In this section, the overall result of the algorithms will be evaluated and analysed with the visual representation.

A. Decision Tree

Splitting- The way toward apportioning the informational index into subsets. Parts are shaped on a specific variable and in a specific area. For each split, two conclusions are made: the indicator variable utilized for the split, called the part factor, and the arrangement of qualities for the indicator variable (which are part between the left kid hub and the correct youngster hub), called the split point. The split depends on a specific standard, for instance, aggregates of squares (for relapse) from the whole informational index. The leaf hub, additionally called a terminal hub, contains a little subset of the perceptions. Part proceeds until a leaf hub is built.

Pruning- The shortening of parts of the tree. Pruning is the way toward diminishing the measure of the tree by transforming some branch hubs into leaf hubs and expelling the leaf hubs under the first branch. Pruning is helpful in light of the fact that order trees may fit the preparation information well, yet may complete poor employment of arranging new qualities. Lower branches might be emphatically influenced by anomalies. Pruning empowers you to locate the following biggest tree and limit the issue. A less difficult tree frequently evades over-fitting.

Tree selection- The way toward finding the littlest tree that fits the information. Generally, this is the tree that yields the least cross-approved mistake. See the clarification for a cross-approval blunder in Output From Decision Trees.

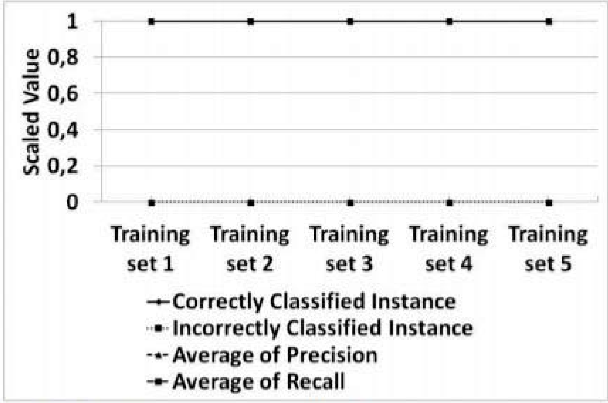


Fig.4 Decision Tree performance on training set

B. Naive Bayes

Naive Bayes classifiers have worked great in some genuine circumstances, broadly record characterization and spam separating. They require a little measure of preparing information to appraise the important parameters. (For hypothetical reasons why guileless Bayes functions admirably, and on which sorts of information it does, see the references underneath.)

Bayes classifiers can be incredibly quick contrasted with progressively complex strategies. The decoupling of the class contingent component conveyances implies that every dissemination can be autonomously assessed as a one dimensional circulation. This thus mitigates issues originating from the scourge of dimensionality.

In below fig.5 the performance the algorithm is given,where 5 set of training is done and results are compared respectively. Each time their true positive(TP) ,false positive(FP) ,precision and recall also calculated.

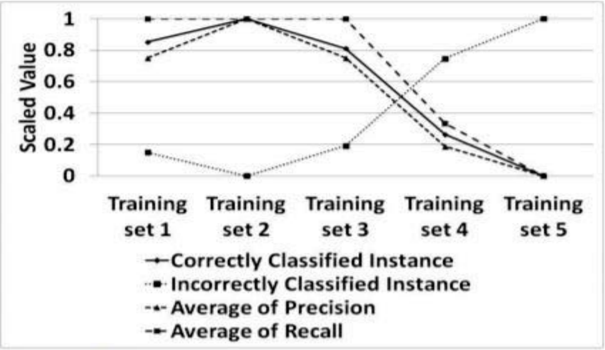


Fig.5 Naive Bayes performance on training set

C. K-Nearest Neighbor

In K-nearest neighbor (KNN) [9] k value is known based features similar nearest neighbor are chosen from which the mode of the nearest point is taken as the output. In KNN[10] algorithm K should always be odd. KNN algorithm map the data as a graph and also map the test data and return the result according to the training data. From the nearest neighbor, the majority of true and false are taken and the result is decided. In our work, we use Euclidean[11] distance to calculate the nearest distance. Comparatively, Euclidean provides better distance results compared to others. In this experiment, we use Euclidean distance[10]. Based on coordinates of the given sample they plotted over the x and y-axis and Euclidean distance is determined using the given formula,

$$x^{2} = (c-a)^{2} + (d-b)^{2} \qquad (13)$$

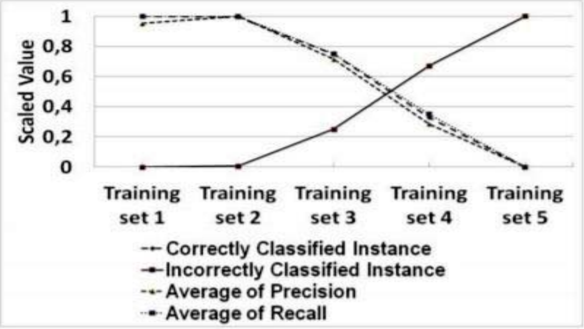


Fig.6 K-NN performance on training set

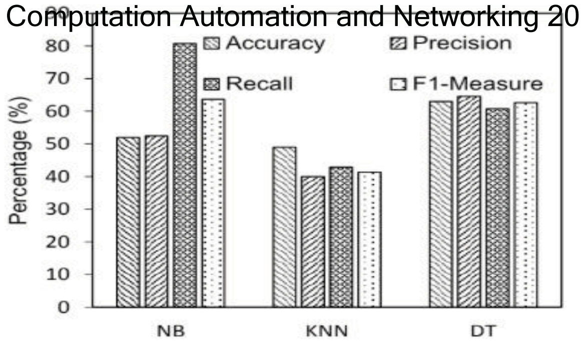


Fig.7 Other Algorithm Accuracy Calculation

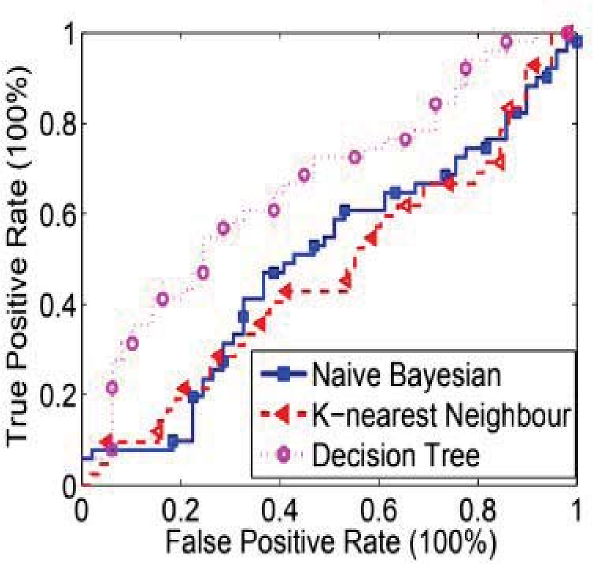


Fig.8 Other Algorithm accuracy graph

D. RNN Algorithm evaluation

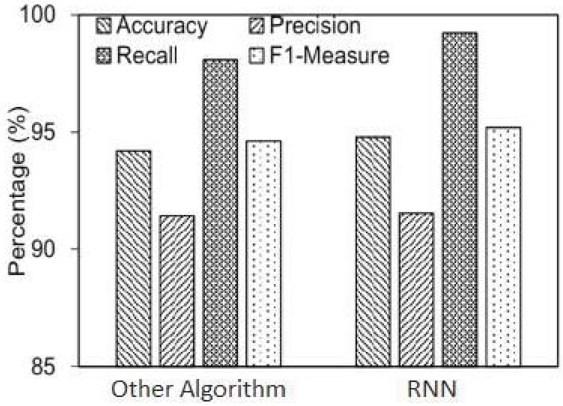


Fig.9 Other Algorithm vs RNN Accuracy Calculation

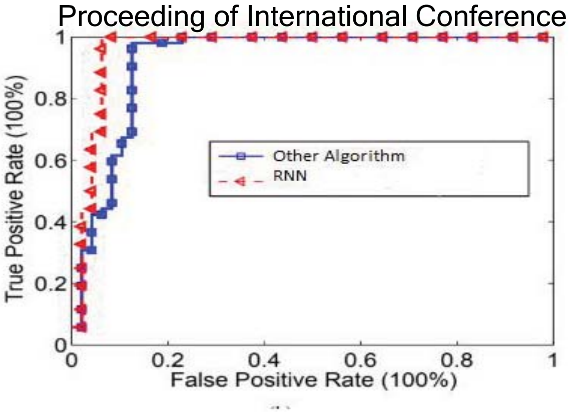


Fig.10 Other Algorithm vs RNN Accuracy Graph

V. Conclusion

According to the medical report, the people death rate is high due to the chronic disease and over 70% of the patients' money is spent over chronic disease treatment. Healthcare is a serious issue in major countries all over the world. According to the medical standard report, chronic disease is the main cause of death. Therefore the analysis over the chronic disease is more important to reduce the risk of people life. The obtained result comparatively from the predictive algorithms such as naive Bayes, K-Nearest Neighbour, Decision tree, support vector machine, and Recurrent neural network. Where Recurrent neural network provides us with the highest rate of accuracy is 97.62% since they follow the feed-forward loop. In the future, our work will include MRI scans, X-rays for better accuracy rate.

VI. Reference

[1] Min Chen, Yixue Hao, Kai Hwang, Fellow, IEEE, Lu Wang, and Lin Wang, “Disease Prediction by Machine Learning over Big Data from Healthcare Communities”, IEEE Access 2017.

[2] G. K. Gupta, Introduction to Data Mining with Case Studies. Prentice Hall of India, New Delhi, 2006.

[3] P-N. Tan, M. Steinbach, V. Kumar, Introduction to Data Mining. Addison Wesley Publishing, 2006.

[4] J. Han and M. Kamber, Data Mining: Concepts and Techniques. Morgan-Kaufmann Publishers, San Francisco, 2001.

[5] X. Niuniu and L. Yuxun, “Review of Decision Trees,” IEEE, 2010.

[6] V. Mohan, “Decision Trees: A comparison of various algorithms for building Decision Trees,” Available at: http://cs.jhu.edu/~vmohan3/document/ai_dt.pdf

[7] T. Miquelez, E. Bengoetxea, P. Larranaga, “Evolutionary Computation based on Bayesian Classifier,” Int. J. Appl. Math. Comput. Sci. vol. 14(3), p p. 335 – 349, 2004.

[8] M. K. Stern, J. E. Beck, and B. P. Woolf, “Naïve Bayes Classifiers for User Modeling,” Available at: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.118.979

[9] Wikipedia, “k-Nearest Neighbor Algorithm,” Available at: http://en.wikipedia.org/wiki/K-nearest_neighbor_algorithm

[10] V. Garcia, C. Debreuve, “Fast k Nearest Neighbor Search using GPU,” IEEE, 2008.

[11] J. Davis and M. Goadrich, “The Relationship Between Precision-Recall and ROC Curves,” Proceedings of the 23rd International Conference on Machine Learning, Pittsburgh, 2006.

[12] 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 Transactions on Intelligent Transportation Systems, vol. 16, no. 6, pp. 3033–3049, 2015.

[13] P. Groves, B. Kayyali, D. Knott, and S. V. Kuiken, “The big data revolution in healthcare: Accelerating value and innovation,” 2016.

[14] M. Chen, S. Mao, and Y. Liu, “Big data: A survey,” Mobile Networks and Applications, vol. 19, no. 2, pp. 171–209, 2014.

[15] P. B. Jensen, L. J. Jensen, and S. Brunak, “Mining electronic health records: towards better research applications and clinical care,” Nature Reviews Genetics, vol. 13, no. 6, pp. 395–405, 2012.

[16] J. C. Ho, C. H. Lee, and J. Ghosh, “Septic shock prediction for patients with missing data,” ACM Transactions on Management Information Systems (TMIS), vol. 5, no. 1, p. 1, 2014.

[17] “Ictclas,” http://ictclas.nlpir.org/.

[18] “word2vec,” https://code.google.com/p/word2vec/.

[19] Y.-D. Zhang, X.-Q. Chen, T.-M. Zhan, Z.-Q. Jiao, Y. Sun, Z.-M. Chen, Y. Yao, L.-T. Fang, Y.-D. Lv, and S.-H. Wang, “Fractal dimension estimation for developing pathological brain detection system based on minkowski-bouligand method,” IEEE Access, vol. 4, pp. 5937–5947, 2016.

[20] S. Basu Roy, A. Teredesai, K. Zolfaghar, R. Liu, D. Hazel, S. Newman, and A. Marinez, “Dynamic hierarchical classification for patient risk-of-readmission,” in Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining. ACM, 2015, pp. 1691–1700.

[21] Sak Haim, Andrew Senior, Franoise Beaufays, "Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition", 2014.

[22] D. W. Hosmer, S. Lemeshow, Applied Logistic Regression, Wiley Interscience, 2000.

[23] C.J.C. Burges, "Simplified Support Vector Decision Rules", *Proc. 13th Int'l Conf. Machine Learning*, pp. 71-77, 1996.