Innovative Science and Technology Publications

International Journal of Future Innovative Science and Technology,
ISSN: 2454- 194X Volume-3, Issue-2, Jun - 2017

A SURVEY ON MACHINE LEARNING ALGORITHMS APPLIED ON HEALTH CARE DATA FOR DISEASE PREDICTION

V.Priyanga

PG Scholar
Computer Science and Engineering Department,
Vivekanandha College of Engineering for Women,
Namakkal, India.
Priyangadhakshi27@gmail.com

A.Thamaraiselvi

Assistant Professor
Computer Science and Engineering Department,
Vivekanandha College of Engineering for Women,
Namakkal, India.
thamaraiselviarjunan@gmail.com

Jun – 2017

www.istpublications.com

International Journal of Future Innovative Science and Technology, ISSN: 2454- 194X, Volume-3, Issue-2, Jun - 2017

A SURVEY ON MACHINE LEARNING ALGORITHMS APPLIED ON HEALTH CARE DATA FOR DISEASE PREDICTION

V.Priyanga

PG Scholar
Computer Science and Engineering Department,
Vivekanandha College of Engineering for Women,
Namakkal, India.
Priyangadhakshi27@gmail.com

A.Thamaraiselvi

Assistant Professor
Computer Science and Engineering Department,
Vivekanandha College of Engineering for Women,
Namakkal, India.
thamaraiselviarjunan@gmail.com

Abstract: With the revolution of big data in biomedical and healthcare communities, accurate analysis of medical data helps early disease detection, patient care and community services. Nowadays disease prediction is very important to reduce the mortality of human deaths. Using the datasets of different patients taken from the sources like data.gov.in, hospital etc we can analyze and predict the chronic diseases. Based on the datasets provided with patient's symptoms, past diagnosis and lifestyle we can predict the diseases like diabetes, increased blood pressure, breast cancer, heart disease and chronic kidney disease etc. Machine learning algorithms are used for processing the structured and unstructured data for effective prediction. Already many machine learning algorithms are applied in the healthcare for disease prediction which gives accurate accuracy. In this paper, we survey about various disease prediction methods using machine learning algorithms which mainly focus on their accuracy. Based on the result of analysis we propose a new algorithm for accurate disease prediction which has higher accuracy than the existing solutions.

Index Terms: *Big data analytics; Machine Learning Algorithms; Healthcare Data*

1. INTRODUCTION

Disease prediction plays an important role for predicting the diseases in the earlier stage. Most of the peoples from the rural side are not having the knowledge about the disease. So awareness of disease should be provided to the people to predict the diseases and reduce the mortality of humans. Due to the resource problem they are not able to meet the doctor and finally it leads to death. Healthcare is to place more emphasis on prevention and less on treatment the former is proactive while the latter is reactive. Not every health issue is preventable, but in many cases, early intervention can lead to better health outcomes and lower costs. One of the key elements of preventive healthcare is the disease screen. Through such screening, diseases can be diagnosed while still in the early, treatable stage. However, it is not feasible for everyone to get screened for every possible disease. A better solution is to have a cheap, scalable, and reliable means of measuring disease risk and predicting disease occurrence. Healthcare providers or insurers could then use this predictor to identify high-risk individuals and recommend tests or other interventions.

Conventional medicine requires doctors and other health care professionals to treat diseases using drugs, radiation and therapy. These professionals are well trained in the field of medicine. But it is not possible to remember all the information that they may need for every circumstances. Even if the professionals had access to all the data that they needed to treat the diseases they face, it would take a long time for them to analyze all of that data and come up with a suitable solution based on the patient's medical profile.

Health care industry generates a large amount of complex data about patients regarding clinical examination, treatment report, hospital resource management records, electronic patient records, medicine etc which has become cumbersome to organize properly. Predictive analytics is very important in healthcare field. For effective prediction of chronic diseases we use machine learning algorithms. The accurate diagnosis of life-threatening diseases such as breast cancer, heart disease, liver disease etc is a very crucial task in medical science. The humans and computers can be integrated together to achieve best results for correct diagnosis of diseases by balancing the knowledge of human experts in related domains with the vast search potential of computers. This kind of difficulty could be resolved with the help of machine learning techniques. Computer based decision support system can play an important role in correct diagnosis and cost effective treatment. The use of computers and information technology is being increasingly implemented in health care organization in order to help doctors in their day to day decision making activities. We aim to analyze the risks of diseases like diabetes, breast cancer, increased blood pressure, heart disease and chronic kidney disease in individuals based on their diagnostics history, symptoms experienced and current lifestyle using machine learning and to suggest preventive measures, assessment and for information to the patient. There has been a lot of work over the years to develop a model that can be used to predict the risk of various diseases in individuals in order to prevent them and reduce the risk.

Advantages of using machine learning in health care are

- More accurate diagnosis.
- Early involvement to prevent diseases.
- If the predicted risk is high, necessary steps can be taken to avoid the disease.
- Patients can use this system for information for self.

Machine Learning Algorithms:

Freed from the limitations of human scale thinking and analysis, machine learning is able to discover and display the patterns buried in the data. The goal of machine learning is to understand the structure of data so that accurate predictions can be made based on the properties of that data. Machine learning techniques can be classified into supervised learning techniques and unsupervised learning techniques.

Supervised Learning Techniques:

In case of supervised machine learning, algorithm induces a mapping function from given labeled training dataset to map new input data to its desired output. Labeled training dataset comprises of examples, which is a pair of input data and its output value. The problems solved by supervised learning techniques are basically categorized as regression and classification problems. In a regression problem, input variables are mapped to continuous output function whereas in a classification problem, input variables are mapped to discrete categories.

Unsupervised Learning Techniques:

In case of unsupervised machine learning, algorithm infers a mapping function to find hidden patterns and correlation between them from unlabelled input dataset. Input dataset comprises of examples, each example is an input data with no explicit output value. We have very little idea of the output values in this case. We can find correlations by clustering the data as there is no feedback or teacher available for correction. For example, we have to discover close-knit group of friends in face book. E.g. k-means clustering algorithm, hierarchical clustering.

2. LITERATURE REVIEW

This section gives a detailed review about various machine learning algorithms for disease prediction. The following papers are survived in this section.

In [1] authors Min Chen, Yixue Hao, Kai Hwang, Fellow, IEEE, Lu Wang, and Lin Wang in the paper "**Disease Prediction by Machine**

Learning over Big Data from Healthcare Communities" explains that accurate analysis of medical data benefits early disease detection, patient care and community services. Moreover, different regions exhibit unique characteristics of certain regional diseases, which may weaken the prediction of disease outbreaks. For effective disease prediction we use machine learning algorithms. Using the datasets of different patients taken from the sources like data.gov.in, hospital etc we can analyze and predict the diseases. The datasets are classified into two types; they are structured data and unstructured data. The structured data includes laboratory data and the patient's basic information such as the patient's age, gender and life habits, etc. While the unstructured text data includes the patient's narration of his/her illness, the doctor's interrogation records and diagnosis, etc. For S-data, we use three conventional machine learning algorithms, i.e., Naïve Bayesian (NB), K-nearest Neighbor (KNN), and Decision Tree (DT) algorithm to predict the risk of chronic disease. For unstructured data we use convolutional neural network (CNN) to extract text characteristics automatically. For simple diseases only a few features of structured data can resulting in fairly good effect of disease risk prediction. For complex diseases the features of structured data is not enough to predict the high risk of diseases.

In [2] authors Mohtaram Mohammadi, Mitra Hosseini, Hamid Tabatabaee in the paper "**Prediction of Diabetes Using Bayesian Network**" proposes a method for effective prediction of diabetes mellitus. Diabetes mellitus is a chronic disease and a major public health challenge worldwide. Using the datasets of different patients taken from the hospital we can predict the diabetes disease. Datasets contains all the details of person like fast gttvalue, casual gttvalue, number of time pregnant, diastolic blood pressure (mmhg), triceps skin fold thickness (mm), serum insulin(µU/ml), body mass index (kg/m)diabetes pedigree function, age of person. Data mining method is used to extract knowledge from information stored in dataset and generate clear and understandable description of patterns. Bayesian Network classifier was proposed to predict the persons whether diabetic or not. In this paper classification with Bayesian classifier shows the best accuracy for diagnosis of diabetes. This method has a drawback is that all risk factors have not been considered. Bayesian classifier is not sufficient if there are any missing values.

In [3] authors Yajuan Wang, Kenney Ng, Roy J. Byrd, Jianying Hu, Shahram Ebadollahi, Zahra Daar, Christopher deFilippi, Steven R. Steinhubl, Walter F. Stewar in the paper "**Early Detection of Heart Failure with Varying Prediction Windows by Structured and Unstructured Data in Electronic Health Records**" describes about the early detection of heart failure using structured and unstructured data taken from the Electronic Health Records. Early detection of HF would provide the means to test lifestyle and pharmacologic interventions that may slow disease progression and improve patient outcomes. Structured and unstructured data from electronic health records (EHR) to predict onset of HF with a particular focus on how prediction accuracy different in relation to time before diagnosis. The prediction window was assessed from 60 to 720 days prior to the diagnosis date and the predictive ability of clinical factors was examined to identify factors that are more effective for early detection. When the prediction window decreased, performance of the predictive HF models increased from 65% to 74% for the unstructured data. For structured data the predictive HF models increased from 73% to 81%. As a result the combination of the two data increased from 76% to 83%. The limitation of this method is that prediction accuracy of unstructured data is less than that of the structured data.

In [4] authors Prerana T H M, Shivaprakash N C , Swetha N in the paper **"Prediction of Heart Disease Using Machine Learning Algorithms- Naïve Bayes, Introduction to PAC Algorithm, Comparison of Algorithms"** explains the using of machine learning algorithms for the prediction of heart disease. Prediction of heart disease is a major challenge faced by hospitals and medical centers, especially when it comes to accuracy. Classification techniques of Data Mining and Machine Learning Algorithms play a significant

role in prediction and data exploration. The main objective of this paper is predicting the heart disease risk level of a patient using machine learning algorithms and creating a Centralized System for both doctors and patients to login and view the data on Cloud. Doctors may sometimes fail to take accurate decisions while diagnosing the heart disease of a patient, therefore heart disease prediction systems which use machine learning algorithms assist in such cases to get accurate results. There are many tools available which use prediction algorithms but they have some flaws. For analyzing the heart disease we use datasets with 13 attributes. Then the data set is fed into the classification model i.e. Naïve Bayes Classification and Probabilistic Analysis and Classification to predict the risk of heart disease. The advantage of using this algorithm is to decrease the execution time and process more data.

In [5] authors Maulik R. Kamdar in the paper **"Visualizing Personalized Cancer Risk Prediction"** explains different methods used for predicting the risk of cancer. It enables evidence-based personalized diagnosis for any patient, the location of where the tumor occurs (e.g. brain, liver, etc.) is less relevant than the underlying genetic signature that the cancer cells express. The genomic datasets are available for data analysis, and identifying the underlying patterns could aid in the diagnosis, prognosis and treatment on a personalized basis. The motivation of this project is to lead towards the development of a diagnostic framework, integrated with a cancer genome visualization tool, which enables a clinical researcher to visually predict cancer risk in new patients using machine learning (ML) classification models. Finally, we use Unsupervised Learning on the datasets for the discovery of new Bio-markers. The algorithms used for the prediction is Naive Bayes, SVM, Decision tree and Random forest. The advantage of this method gives lower test error and higher Sensitivity.

In [6] authors Tony Hao Wu, Grantham Kwok-Hung Pang, Enid Wai-Yung Kwong in the paper "**Predicting Systolic Blood Pressure Using Machine Learning**" describes about the prediction of systolic blood pressure by correlated variables (BMI, age, exercise, alcohol, smoke level etc) using artificial neural network. For prediction system, two neural network algorithms are back-propagation neural network and radial basis function network are used to construct and validate the prediction system. This method of predicting systolic blood pressure contributes to giving early warnings to young and middle-aged people who may not take regular blood pressure measurements. Based on a database the probabilities of the total difference between the measured and predicted value of systolic blood pressure under 10mm Hg are 51.9% for men and 52.5% for women using the back propagation neural network. With the same input variables and network status, the radial basis function networks are 51.8% and 49.9% for men and women respectively. Sometimes an isolated blood pressure measurement is not very accurate due to the daily oscillation and the predictor can provide another reference value to the medical staff. In this paper machine learning techniques are used as an efficient tool for analyzing the relationship between the lifestyle condition (BMI, age, stress and exercise level) and the systolic blood pressure of a person.

In [7] authors Hiba Asria, Hajar Mousannif, Hassan Al Moatassime, Thomas Noel in the paper "**Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis**" explains the performance comparison between different machine learning algorithms used for breast cancer prediction. Breast cancer is the main cause of women's deaths worldwide. So prediction is very important to save the humans life. For predicting the risk of breast cancer, the data are collected from the hospital. Classification and data mining methods are used
to classify the data and are widely used in analysis and diagnosis for making decisions. In this paper, the performances are compared with the different machine learning algorithms based on efficiency and effectiveness of each algorithm in terms of accuracy, precision, sensitivity and specificity. The goal is to achieve the best accuracy with the lowest error rate in analyzing data. The experimental results show that machine learning algorithms are better than the data mining methods for accurate predictions. In

machine learning algorithms, SVM achieves the highest accuracy (97.13%) with the lowest error rate (0.02%) compared to other algorithms.

In [8] author Manish Kumar in the paper "**Prediction of Chronic Kidney Disease Using Random Forest Machine Learning Algorithm**" describes the prediction of chronic kidney disease based on the performance of machine learning algorithms. Chronic kidney disease may be caused by diabetes, high blood pressure, hypertension, Coronary artery Disease, lupus, Anemia, Bacteria and albumin in urine, complications from some medications, Deficiency of Sodium and Potassium in blood and family history of kidney disease and many more. Chronic Kidney Disease prediction is one of the most central problems in medical decision making because it is one of the leading cause of death. For effective prediction of chronic kidney disease we use dataset which are collected from the hospital. The algorithms used in the prediction task are Random Forest (RF), Naïve Bayes, Sequential Minimum Optimization (SMO), Radial Basis Function (RBF Classifier), Multilayer Perceptron Classifier (MLPC) and Simple Logistic (SLG) using Weka. In this paper, performances of six algorithms were compared with each other. The experimental result shows that Random Forest has produced superior prediction performance in terms of classification accuracy, AUC and MCC respectively.

In [9] authors Faezeh Hosseinzadeh, Amir Hossein KayvanJoo, Mansuor Ebrahimi and Bahram Goliaei in the paper "**Prediction of lung tumor types based on protein attributes by machine learning algorithms**" explains the early diagnosis of lung cancers and distinction between the tumor types (Small Cell Lung Cancer (SCLC) and Non-Small Cell Lung Cancer (NSCLC) are very important to increase the survival rate of patients. Machine learning methods have been widely used in prediction especially in medical diagnosis. The algorithms performances in lung cancer tumor type prediction increased when they applied on datasets created by attribute weighting models rather than original dataset. Classification and prediction of lung tumors based on structural and physicochemical properties of associated proteins performed and several prediction models such as SVM, ANN and NB used. Seven prediction models of support vector machines (SVM, LibSVM, SVM Linear, SVM Evolutionary, SVM PSO, SVM Fast Large Margin and SVM Hyper) are applied on the datasets. The combination of protein features and attribute weighting models with machine learning algorithms can be effectively used to predict the type of lung cancer tumors (SCLC and NSCLC). Comparing the performances of three types of machine learning models (SVM , ANN and NB) to predict and detect the type of lung tumors based on structural and physicochemical attributes of proteins showed that the Neural Net model ran on SVM dataset gained the best accuracy (88%).

In [10] authors S.Florence, N.G.Bhuvaneswari Amma, G.Annapoorani, K.Malathi in the paper "**Predicting the Risk of Heart Attacks using Neural Network and Decision Tree**" describes the methods for predicting the risk of heart attacks. Heart attack is a common problem in all human beings with the age above 30. The cholesterol level is another one major problem which leads to heart attack. Medical diagnosis is an important it is a complicated task that needs to be sdone accurately and efficiently. In healthcare system to predict the heart attack perfectly, there are some techniques which are already in use. There is some lack of accuracy in the available techniques like Naïve Bayes. In this paper, algorithms like neural network and Decision tree (ID3) are used to predict the heart attacks. Using the dataset provided by the UCI machine learning repository we can analyze and predict the risk of heart attack. The dataset contains 6 attributes like age, sex, cardiac duration, signal, possibility of attack. Statistics provide a strong fundamental background for quantification and evaluation of results. The results of the prediction give more accurate output than the other techniques.

3. ANALYSIS TABLE

This section presents the study on machine learning algorithms for disease

predictions which we reviewed in previous section. Based on this study results we can find the more efficient algorithm used for predicting the diseases.

Table 1: COMPARITIVE STUDY ON MACHINE LEARNING ALGORITHMS FOR DISEASE PREDICTION

S.NO	TITLE	ALGORITHMS USED	MERITS	DEMERITS
1.	Disease Prediction by Machine Learning over Big Data from Healthcare Communities	1.Naïve Bayesian 2.K-nearest neighbor 3.Decision Tree 4.Convolutional Neural Network	Accuracy and speed is high	Uncertainty in processing the unstructured text data
2.	Prediction of Diabetes Using Bayesian Network	Bayesian Network Classifier	Casual Relationship and probabilistic semantics	All risk factors have not been considered. Bayesian classifier is not sufficient for any missing values.
3.	Early Detection of Heart Failure with Varying Prediction Windows by Structured and Unstructured Data in Electronic Health Records	Predictive HF models	Structured and unstructured data achieves superior performance	Early detection of HF is not effective due to slowing progression
4.	Prediction of Heart Disease Using Machine Learning Algorithms- Naïve Bayes, Introduction to PAC Algorithm, Comparison of Algorithms and HDPS	1.Naive Bayes classifier 2.Probabilistic Analysis and Classification	It decreases the execution time and process more data	Result obtained is partially accurate
5.	Visualizing Personalized Cancer Risk Prediction	1.Naive Bayes 2.SVM 3.Decision tree 4.Random forest	Lower test error and higher Sensitivity	Prediction is not exhaustive due to biomarkers
6.	Predicting Systolic Blood Pressure Using Machine Learning	1.Back-propagation neural network 2.Radial basis function network	Hidden nodes in the neural networks are giving reasonably good results	Blood pressure measurement is not very accurate due to daily fluctuation
7.	Using Machine Learning Algorithms for Breast Cancer Risk Prediction and	1.SVM 2.Naive Bayes 3.K-Nearest Neighbor	Best performance with low error rate	FP rate is low

	Diagnosis			
8.	Prediction of Chronic Kidney Disease Using Random Forest Machine Learning Algorithm	1.Random Forest 2.Naive Bayes 3.SMO 4.RBF	RF performs better than other classifiers	Yield poor classification accuracy
9.	Prediction of lung tumor types based on protein attributes by machine learning algorithms	1.SVM 2ANN 3.Naïve Bayes	Processing time and accurate results both are beneficiary.	It takes more time to process the data.
10.	Predicting the Risk of Heart Attacks using Neural Network and Decision Tree	1.Naïve Bayes 2.Neural Network 3.Decision tree	Strong fundamental background for quantification and evaluation of results	Risk occurs in different instances

4. POSSIBLE SOLUTION

To solve the above mentioned demerits, we combine the structured and unstructured data in healthcare field to assess the risk of disease.
First, we used latent factor model to reconstruct the missing data from the medical records collected from a hospital using Convolutional Neural Network. Then by using statistical knowledge, we could determine the major chronic diseases in the region. To handle structured data, we consult with hospital experts to extract useful features.
For unstructured text data, we select the features automatically using CNN algorithm. Finally, we proposed an improved three dimensional CNN-based multimodal disease risk prediction algorithm to process the structured and unstructured data to get more efficient predictions than the existing solution.

5. CONCLUSION

In this paper, we surveyed many research works focused on health care prediction and analyze the merits and demerits of every algorithm used work based on time taken for result and accuracy of prediction. And finally we proposed an improved three dimensional CNN-based multimodal disease risk prediction algorithm to process the structured and unstructured data to get more efficient predictions than the existing solution.

REFERENCES

[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" in IEEE access journal,2017.

[2]Mohtaram Mohammadi, Mitra Hosseini, Hamid Tabatabaee, "Prediction of Diabetes Using Bayesian Network" in International Journal of Computer Science and Information Technologies, Vol. 5 (4) , 2014, 5174-5178.

[3]Yajuan Wang, Kenney Ng, Roy J. Byrd, Jianying Hu, Shahram Ebadollahi, Zahra Daar, Christopher deFilippi, Steven R. Steinhubl, Walter F. Stewar, "Early Detection of Heart Failure with Varying Prediction Windows by Structured and Unstructured Data in Electronic Health Records" 978-1-4244-9270-1/15/$31.00 ©2015 IEEE.

[4] Prerana T H M, Shivaprakash N C , Swetha N, "Prediction of Heart Disease Using Machine Learning Algorithms- Naïve Bayes, Introduction to PAC Algorithm, Comparison of Algorithms" in International Journal of Science and

Engineering , Volume 3, Number 2 - 2015 PP: 90-99 ©IJSE Available.

[5]Maulik R. Kamdar, "Visualizing Personalized Cancer Risk Prediction", http://en.wikipedia.org/wiki/DNA_methylation#In_cancer.

[6]Tony Hao Wu, Grantham Kwok-Hung Pang, Enid Wai-Yung Kwong, "Predicting Systolic Blood Pressure Using Machine Learning", 978-1-4799-4598-6/14/$31.00 ©2014 IEEE.

[7]Hiba Asria, Hajar Mousannif, Hassan Al Moatassime, Thomas Noel in the paper "Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis" in 6th International Symposium on Frontiers in Ambient and Mobile Systems,2016.

[8]Manish Kumar in "Prediction of Chronic Kidney Disease Using Random Forest Machine Learning Algorithm" in International Journal of Computer Science and Mobile Computing, Vol. 5, Issue. 2, February 2016, pg.24 – 33.

[9]Faezeh Hosseinzadeh, Amir Hossein KayvanJoo, Mansuor Ebrahimi and Bahram Goliaei "Prediction of lung tumor types based on protein attributes by machine learning algorithms" in Springerplus, 2013.

[10]S.Florence, N.G.Bhuvaneswari Amma, G.Annapoorani, K.Malathi "Predicting the Risk of Heart Attacks using Neural Network and Decision Tree" in International Journal of Innovative Research in Computer and Communication Engineering, Vol. 2, Issue 11, November 2014.

[11]Apoorva Sharm, Pallavi Rawat, Kajal Pandey, Ravi Shankar Rai "Health Analytics Using Machine Learning: A Survey" in International Journal of Innovative Research in Computer and Communication Engineering, Vol. 5, Issue 4, April 2017.