PERFORMANCE ANALYSIS OF MACHINE LEARNING TECHNIQUE TO PREDICT DIABETES

Sumalatha Potteti, M.Tech(Ph.D), Assistant Professor, Department of CSE, sumalatha.po@gmail.com

G.Mamatha, BTech, Department of CSE, mamathayadav150@gmail.com

D.Navya, BTech, Department of CSE, dupellynavya@gmail.com

P.Priyanka, BTech, Department of CSE, priyankapachipala38@gmail.com

ABSTRACT: Diabetes Mellitus is among critical diseases and lots of people are suffering from this disease. Age, obesity, lack of exercise, hereditary diabetes, living style, bad diet, high blood pressure, etc. can cause Diabetes Mellitus. People having diabetes have high risk of diseases like heart disease, kidney disease, stroke, eye problem, nerve damage, etc. Current practice in hospital is to collect required information for diabetes diagnosis through various tests and appropriate treatment is provided based on diagnosis. Big Data Analytics plays an significant role in healthcare industries. Healthcare industries have large volume databases. Using big data analytics one can study huge datasets and find hidden information, hidden patterns to discover knowledge from the data and predict outcomes accordingly. In existing method, the classification and prediction accuracy is not so high. In this paper, we have proposed a diabetes prediction model for better classification of diabetes which includes few external factors responsible for diabetes along with regular factors like Glucose, BMI, Age, Insulin, etc. Classification accuracy is boosted with new dataset compared to existing dataset. Further with imposed a pipeline model for diabetes prediction intended towards improving the accuracy of classification.

1. INTRODUCTION

Healthcare sectors have large volume databases. Such databases may contain structured, semi-structured or unstructured data. Big data analytics is the process which analyses huge data sets and reveals hidden information, hidden patterns to discover knowledge from the given data. Considering the current scenario, in developing countries like India, Diabetic Mellitus (DM) has become a very severe disease. Diabetic Mellitus (DM) is classified as Non-Communicable Disease (NCB) and many people are suffering from it. Around 425 million people suffer from diabetes according to 2017 statistics. Approximately 2-5 million patients every year lose their lives due to diabetes. It is said that by 2045 this will rise to 629 million.[1] Diabetes Mellitus (DM) is classified asType-1 known as Insulin-Dependent Diabetes Mellitus (IDDM). Inability of human's body to generate sufficient insulin is the reason behind this type of DM and hence it is required to inject insulin to a patient. Type-2 also known as Non-Insulin-Dependent Diabetes Mellitus (NIDDM). This type of Diabetes is seen when body cells are not able to use insulin properly.Type-3 Gestational Diabetes, increase in blood sugar level in pregnant woman where diabetes is not detected earlier results in this type of diabetes. DM has long term complications associated with it. Also, there are high risks of various health problems for a diabetic person. A

technique called, Predictive Analysis, incorporates a variety of machine learning algorithms, data mining techniques and statistical methods that uses current and past data to find knowledge and predict future events. By applying predictive analysis on healthcare data, significant decisions can be taken and predictions can be made. Predictive analytics can be done using machine learning and regression technique.

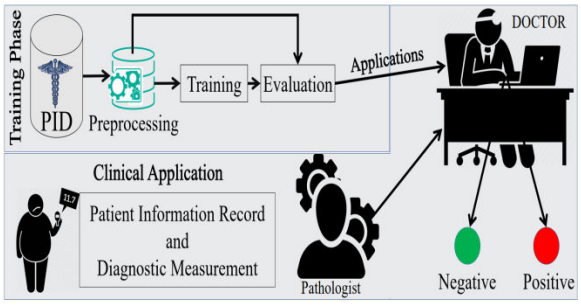

Fig.1: Example figure

Predictive analytics aims at diagnosing the disease with best possible accuracy, enhancing patient care, optimizing resources along with improving clinical outcomes.[1] Machine learning is considered to be one of the most important artificial intelligence features supports development of computer systems having the ability to acquire knowledge from past experiences with no need of programming for every case. Machine learning is considered to be a dire need of today's situation in order to eliminate human efforts by supporting automation with minimum flaws. Existing method for diabetes detection is uses lab tests such as fasting blood glucose and oral glucose tolerance. However, this method is time consuming. This paper focuses on building predictive model using machine learning algorithms and data mining techniques for diabetes prediction.

2. LITERATURE REVIEW

Predictive Analysis of Diabetic Patient Data Using Machine Learning and Hadoop

Now days from health care industries large volume of data is generating. It is necessary to collect, store and process this data to discover knowledge from it and utilize it to take significant decisions. Diabetic Mellitus (DM) is from the Non Communicable Diseases (NCD), and lots of people are suffering from it. Now days, for developing countries such as India, DM has become a big health issue. The DM is one of the critical diseases which has long term complications associated with it and also follows with various health problems. With the help of technology, it is necessary to build a system that store and analyze the diabetic data and predict possible risks accordingly. Predictive analysis is a method that integrates various data mining techniques, machine learning algorithms and statistics that use current and past data sets to gain insight and predict future risks. In this work machine learning algorithm in Hadoop MapReduce environment are implemented for Pima Indian diabetes data set to find out missing values in it and to discover patterns from it. This work will be able to predict types of diabetes are widespread, related future risks and according to the risk level of patient the type of treatment can be provided.

Prediction of Diabetes Based on Personal Lifestyle Indicators

Diabetes Mellitus or Diabetes has been portrayed as worse than Cancer and HIV (Human Immunodeficiency Virus). It develops when there are high blood sugar levels over a prolonged period. Recently, it has been quoted as a risk factor for developing Alzheimer, and a leading cause for

blindness & kidney failure. Prevention of the disease is a hot topic for research in the healthcare community. Many techniques have been discovered to find the causes of diabetes and cure it. This research paper is a discussion on establishing a relationship between diabetes risk likely to be developed from a person's daily lifestyle activities such as his/her eating habits, sleeping habits, physical activity along with other indicators like BMI (Body Mass Index), waist circumference etc. Initially, a Chi-Squared Test of Independence was performed followed by application of the CART (Classification and Regression Trees) machine learning algorithm on the data and finally using Cross-Validation, the bias in the results was removed.

Predictive Analytics in Health Care Using Machine Learning Tools and Techniques

When we have a huge data set on which we would like to perform predictive analysis or pattern recognition, machine learning is the way to go. Machine Learning (ML) is the fastest rising arena in computer science, and health informatics is of extreme challenge. The aim of Machine Learning is to develop algorithms which can learn and progress over time and can be used for predictions. Machine Learning practices are widely used in various fields and primarily health care industry has been benefitted a lot through machine learning prediction techniques. It offers a variety of alerting and risk management decision support tools, targeted at improving patients' safety and healthcare quality. With the need to reduce healthcare costs and the movement towards personalized healthcare, the healthcare industry faces challenges in the essential areas like, electronic record management, data integration, and computer aided diagnoses and disease predictions. Machine Learning offers a wide range of tools, techniques, and frameworks to address these challenges. This paper depicts the study on various prediction techniques and tools for Machine Learning in practice. A glimpse on the applications of Machine Learning in various domains are also discussed here by highlighting on its prominence role in health care industry.

Diagnosis of Diabetes Using Classification Mining Techniques

Diabetes has affected over 246 million people worldwide with a majority of them being women. According to the WHO report, by 2025 this number is expected to rise to over 380 million. The disease has been named the fifth deadliest disease in the United States with no imminent cure in sight. With the rise of information technology and its continued advent into the medical and healthcare sector, the cases of diabetes as well as their symptoms are well documented. This paper aims at finding solutions to diagnose the disease by analyzing the patterns found in the data through classification analysis by employing Decision Tree and Naïve Bayes algorithms. The research hopes to propose a quicker and more efficient technique of diagnosing the disease, leading to timely treatment of the patients.

3. IMPLEMENTATION

In existing method, the classification and prediction accuracy is not so high. Healthcare industries have large volume databases. Using big data analytics one can study huge datasets and find hidden information, hidden patterns to discover knowledge from the data and predict outcomes accordingly. Huge number of people is becoming its victim every day and many are unaware if they have it or not. There are two

types of diabetes. Diabetes mellitus and diabetes. The test conducted to detect diabetes is physical examination and blood sugar test. More research is required to stop this disease. Machine learning and cloud computing will play a major role in the research related work to detect and timely cure it for the people. Diabetes specially affects the elderly and obese people. Diabetes can cause other variety of health problems like heart attack, kidney failure, high blood pressure and diabetic foot syndrome.

Diabetes can be controlled if it is predicted earlier. To achieve this goal this project work we will do early prediction of Diabetes in a human body or a patient for a higher accuracy through applying, Various Machine Learning Techniques. Machine learning techniques Provide better result for prediction by con- structing models from datasets collected from patients. In this work we will use Machine Learning Classification and ensemble techniques on a dataset to predict diabetes. Which are K-Nearest Neighbor (KNN), Logistic Regression (LR), Decision Tree (DT), Support Vector Machine (SVM), Gradient Boosting (GB) and Random Forest (RF). The accuracy is different for every model when compared to other models. The Project work gives the accurate or higher accuracy model shows that the model is capa- ble of predicting diabetes effectively. Our Result shows that Random Forest achieved higher accuracy compared to other machine learning techniques.

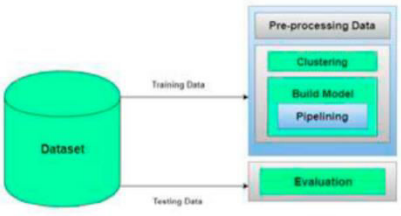

Fig.2: System architecture

MODULES:

- Dataset Collection
- Data Pre-processing
- Clustering
- Build Model
- Evaluation

MODULES DESCRIPTION:

i. **Dataset Collection:** This module includes data collection and understanding the data to study the patterns and trends which helps in prediction and evaluating the results. Dataset description is given belowThis Diabetes dataset contains 800 records and 10 attributes

ii. **Data Pre-processing**: This phase of model handles inconsistent data in order to get more accurate and precise results. This dataset contains missing values. So we imputed missing values for few selected attributes like Glucose level, Blood Pressure, Skin Thickness, BMI and Age because these attributes cannot have values zero. Then we scale the dataset to normalize all values.

iii. **Clustering**: In this phase, we have implemented K-means clustering on the dataset to classify each patient into either a diabetic or non-diabetic class. Before performing K-means clustering, highly correlated attributes were found

which were, Glucose and Age. K-means clustering was performed on these two attributes. After implementation of this clustering we got class labels (0 or 1) for each of our record.

iv. **Model Building:** This is most important phase which includes model building for prediction of diabetes. In this we have implemented various machine learning algorithms for diabetes prediction. These algorithms include Support Vector Classifier, Random Forest Classifier, Decision Tree Classifier, Extra Tree Classifier, Ada Boost algorithm, Perceptron, Linear Discriminant Analysis algorithm, Logistic Regression, K-Nearest Neighbour, Gaussian Naïve Bayes, Bagging algorithm, Gradient Boost Classifier.

v. **Evaluation:** This is the final step of prediction model. Here, we evaluate the prediction results using various evaluation metrics like classification accuracy, confusion matrix and f1-score.

4. METHODOLOGY

ALGORITHM USED:

K-Nearest Neighbor (KNN):

The k-nearest neighbors algorithm, also known as KNN or k-NN, is a non-parametric, supervised learning classifier, which uses proximity to make classifications or predictions about the grouping of an individual data point. The KNN algorithm can compete with the most accurate models because it makes highly accurate predictions. Therefore, you can use the KNN algorithm for applications that require high accuracy but that do not require a human-readable model. The quality of the predictions depends on the distance measure. It's used in many different areas, such as handwriting detection, image recognition, and video recognition. KNN is most useful when labeled data is too expensive or impossible to obtain, and it can achieve high accuracy in a wide variety of prediction-type problems.

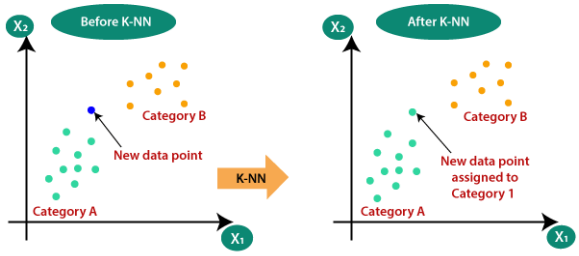

Fig.3: KNN model

Logistic Regression (LR):

Logistic regression is a supervised learning classification algorithm used to predict the probability of a target variable. The nature of target or dependent variable is dichotomous, which means there would be only two possible classes. Logistic regression is a statistical method used to predict the outcome of a dependent variable based on previous observations. It's a type of regression analysis and is a commonly used algorithm for solving binary classification problems.

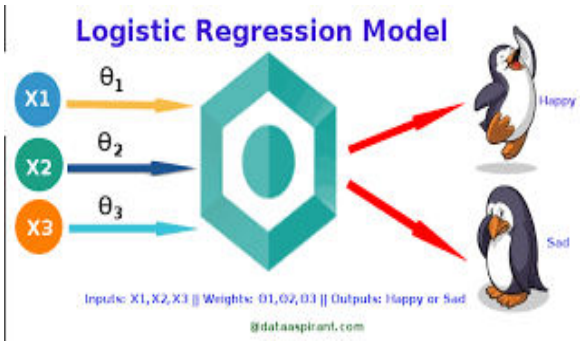

Fig.4: LR model

Decision Tree (DT):

A decision tree is a decision support tool that uses a tree-like model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm that only contains conditional control statements. Decision trees use multiple algorithms to decide to split a node into two or more sub-nodes. The creation of sub-nodes increases the homogeneity of resultant sub-nodes. In other words, we can say that the purity of the node increases with respect to the target variable.

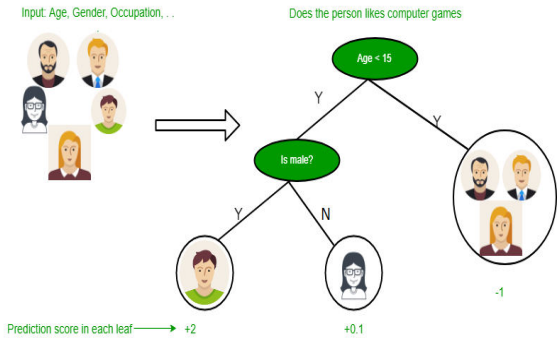

Fig.5: Decision tree model

Support Vector Machine (SVM):

Support Vector Machine(SVM) is a supervised machine learning algorithm used for both classification and regression. Though we say regression problems as well its best suited for classification. The objective of SVM algorithm is to find a hyperplane in an N-dimensional space that distinctly classifies the data points. Support vector machines (SVMs) are a set of supervised learning methods used for classification, regression and outliers detection. The advantages of support vector machines are: Effective in high dimensional spaces. Still effective in cases where number of dimensions is greater than the number of samples.

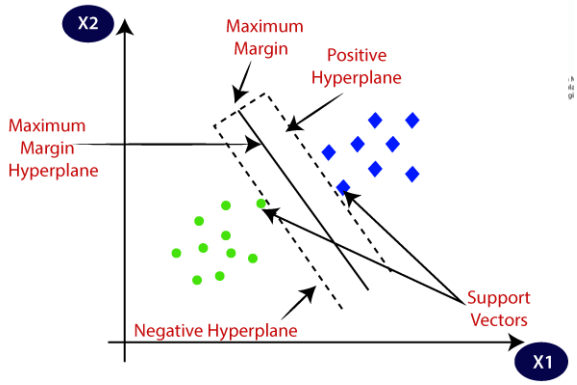

Fig.6: SVM model

Gradient Boosting (GB):

Gradient boosting is a method standing out for its prediction speed and accuracy, particularly with large and complex datasets. From Kaggle competitions to machine learning solutions for business, this algorithm has produced the best results. Gradient boosting is a machine learning technique used in regression and classification tasks, among others. It gives a prediction model in the form of an ensemble of weak prediction models, which are typically decision trees.

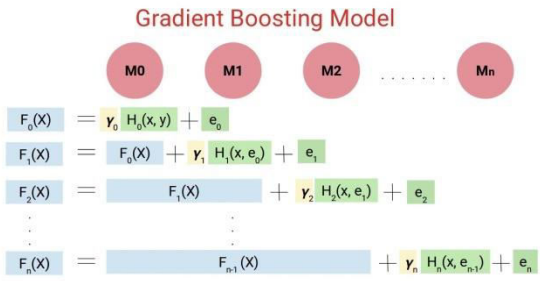

Fig.7: Gradient boosting model

Random Forest (RF):

Random forest is a Supervised Machine Learning Algorithm that is used widely in Classification and Regression problems. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression. Random Forest is a supervised machine learning algorithm made up of decision trees. Random Forest is used for both classification and regression—for example, classifying whether an email is “spam” or “not spam”.

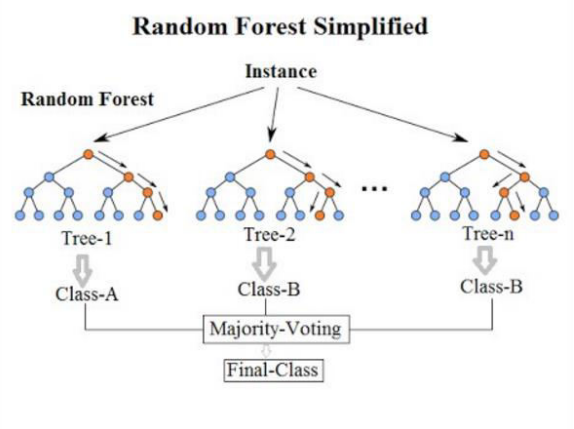

Fig.8: Random forest model

5. EXPERIMENTAL RESULTS

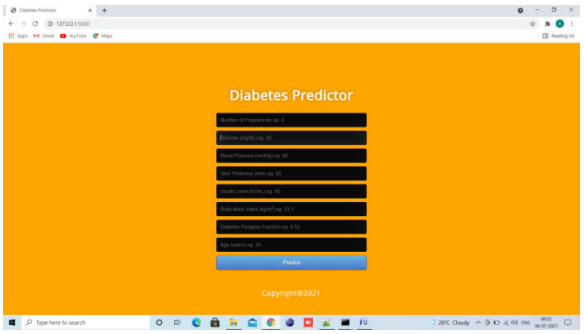

Fig.5: Output screen

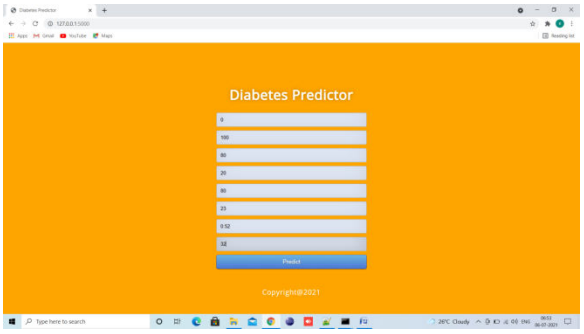

Fig.6: Output screen

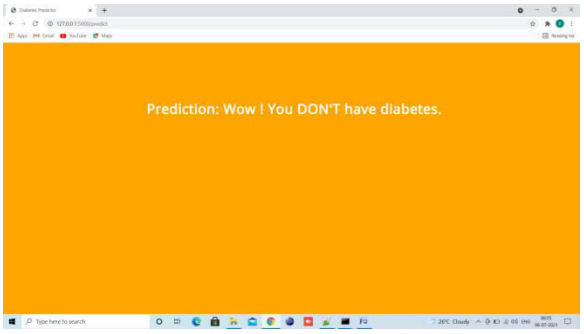

Fig.7: Output screen

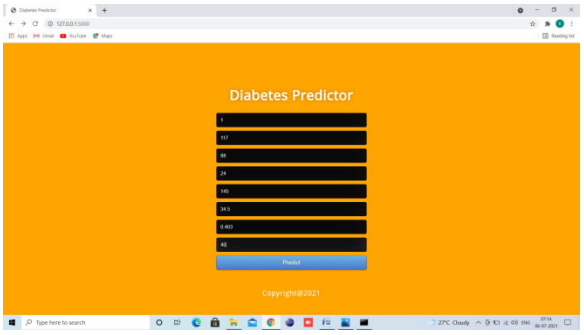

Fig.8: Output screen

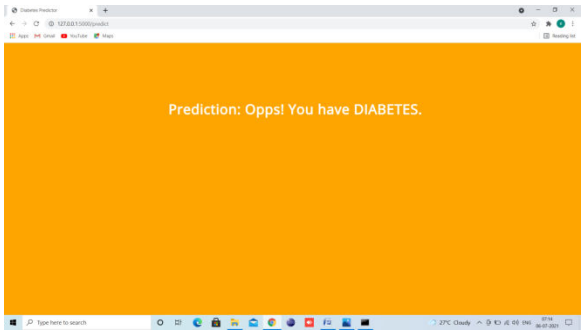

Fig.9: Output screen

6. CONCLUSION

In this study, various machine learning algorithms are applied on the dataset and the classification has been done using various algorithms of which Logistic Regression gives highest accuracy of 96%. Application of pipeline gave AdaBoost classifier as best model with accuracy of 98.8%. We have seen comparison of machine learning algorithm accuracies with two different datasets. It is clear that the model improves accuracy and precision of diabetes prediction with this dataset compared to existing dataset. Further this work can be extended to find how likely nondiabetic people can have diabetes in next few years.

7. FUTURE SCOPE

Our future work will focus on integration of other methods into the used model for tuning the parameters of models for better accuracy. Then testing these models with large dataset having minimum or no missing attribute values will reveal more insights and better prediction accuracy.

REFERENCES

[1] A. Belle, R. Thiagarajan, S. M. R. Soroushmehr, F. Navidi, D. A. Beard, and K. Najarian, “Big Data Analytics in Healthcare,” Hindawi Publ. Corp., vol. 2015, pp. 1–16, 2015.

[2] J. Andreu-Perez, C. C. Y. Poon, R. D. Merrifield, S. T. C. Wong, and G.-Z. Yang, “Big Data for Health,” IEEE J. Biomed. Heal. Informatics, vol. 19, no. 4, pp. 1193–1208, 2015

[3] E. Ahmed et al., “The role of big data analytics in Internet of Things,” Comput. Networks, vol. 129, no. December, pp. 459–471, 2017

[4] “The big-data revolution in US health care: Accelerating value and innovation | McKinsey & Company.” [Online]. Available: https://www.mckinsey.com/industries/healthcare-systems-andservices/our-insights/the-big-data-revolution-in-us-health-care. [Accessed: 12-May-2018].

[5] M. Chen, Y. Hao, K. Hwang, L. Wang, and L. Wang, “Disease Prediction by Machine Learning over Big Data from Healthcare Communities,” IEEE Access, vol. 5, no. c, pp. 8869–8879, 2017.

[6] L. Zhou, S. Pan, J. Wang, and A. V. Vasilakos, “Machine learning on big data: Opportunities and challenges,” Neurocomputing, vol. 237, pp. 350–361, May 2017.

[7] J. B. Heaton, N. G. Polson, and J. H. Witte, “Deep learning for finance: deep portfolios,” Appl. Stoch. Model. Bus. Ind., vol. 33, no. 1, pp. 3–12, Jan. 2017.

[8] K. Lin, M. Chen, J. Deng, M. M. Hassan, and G. Fortino, “Enhanced Fingerprinting and Trajectory Prediction for IoT Localization in Smart Buildings,” IEEE Trans. Autom. Sci. Eng., vol. 13, no. 3, pp. 1294–1307, Jul. 2016.

[9] 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. Informatics, vol. 13, no. 4, pp. 1932– 1940, Aug. 2017.

[10] P. A. Chiarelli, J. S. Hauptman, and S. R. Browd, "Machine Learning and the Prediction of Hydrocephalus," JAMA Pediatr., vol. 172, no. 2, p. 116, Feb. 2018.

[11] A. Jindal, A. Dua, N. Kumar, A. K. Das, A. V. Vasilakos, and J. J. P. C. Rodrigues, "Providing Healthcare-as-a-Service Using Fuzzy Rule-Based Big Data Analytics in Cloud Computing," IEEE J. Biomed. Heal. Informatics, pp. 1–1, 2018.

[12] N. M. S. kumar, T. Eswari, P. Sampath, and S. Lavanya, "Predictive Methodology for Diabetic Data Analysis in Big Data," Procedia Comput. Sci., vol. 50, pp. 203–208, Jan. 2015.

[13] J. Zheng and A. Dagnino, "An initial study of predictive machine learning analytics on large volumes of historical data for power system applications," in 2014 IEEE International Conference on Big Data (Big Data), 2014, pp. 952–959.

[14] International Journal of Advanced Computer and Mathematical Sciences. Bi Publication-BioIT Journals, 2010.

[15] M. Chen, Y. Hao, K. Hwang, L. Wang, and L. Wang, "Disease Prediction by Machine Learning Over Big Data From Healthcare Communities," IEEE Access, vol. 5, pp. 8869–8879, 2017.