Proceedings of the 24th International Conference on Automation & Computing, Newcastle University, Newcastle upon Tyne, UK, 6-7 September 2018

Prediction of Diabetes Using Machine Learning Algorithms in Healthcare

[1]Muhammad Azeem Sarwar, [2]Nasir Kamal, [3]Wajeeha Hamid,[4]Munam Ali Shah
[1,4]Department of Computer Science, COMSATS Institute of Information technology, Islamabad, Pakistan
[2]Electrical Engineering, School of Electrical Engineering and Computer Science, NUST. Pakistan.
[3]Center for Advanced Studies in Engineering, Islamabad, Pakistan
azeem261@gmail.com,mshah@comsats.edu.pk,
14mseenkamal@seecs.edu.pk,wajeeha094@gmail.com,wajeeha.hamid@xflowresearch.com, mshah@comsats.edu.pk

***Abstract*—There are several machine learning techniques that are used to perform predictive analytics over big data in various fields. Predictive analytics in healthcare is a challenging task but ultimately can help practitioners make big data-informed timely decisions about patient's health and treatment. This paper discusses the predictive analytics in healthcare, six different machine learning algorithms are used in this research work. For experiment purpose, a dataset of patient's medical record is obtained and six different machine learning algorithms are applied on the dataset. Performance and accuracy of the applied algorithms is discussed and compared. Comparison of the different machine learning techniques used in this study reveals which algorithm is best suited for prediction of diabetes. This paper aims to help doctors and practitioners in early prediction of diabetes using machine learning techniques.**

Keywords-Big data analytics; Predictive Analytics; Machine Learning; Healthcare.

I. INTRODUCTION

As the technology is advancing, devices are generating large amount of data every. There is a global outburst in the availability of data for researchers. The complexity, huge size and heterogeneity of data require one to search, discover and adopt new software tools and mechanisms in order to successfully manage, analyse, and visualize the data [1]. In [2], author have obtained results from Google Scholar for the term "Big data" from year 2008-2015. These results shows how this field is evolved through years and the increasing rate of publications in the field of big data. This exponential growth in the field of big data started from 2012 and still this area of research is attracting more and more researchers.

A report by McKinsey states that 50% of Americans are the victim of one or more chronic diseases, and they spend around 80% of American medical care fee on treatment of these chronic diseases [4]. Around 2.7 trillion USD are being spent on the treatment of those chronic diseases annually. This amount of 2.7 trillion USD contains 18% of the annual Gross Domestic Product (GDP) of the United States. Many other countries are also suffering from these chronic diseases. According to a Chinese report published in 2015 86.6% of deaths are caused by these chronic diseases in China [5]. Considering the annual growth of data generation, by 2020 data we generate annually will reach 44 trillion gigabytes which is ten times the size of the data generated in 2013 [6].

Big data in healthcare industry refers to electronic health datasets so large and complex for traditional software tools to process. Healthcare analytics refers to the systematic use of these healthcare datasets for business insights, decision making, planning, learning, early prediction and detection of diseases by using different statistical, predictive and quantitative models and techniques. Figure 1, shows the fast increase in the number of publications referring to "predictive analytics in healthcare" from year 2005 to 2017.

Healthcare analytics needs a technology that helps to perform a real time analysis on the massive dataset. In healthcare industry the application of predictive analytics are significantly high. Predictions can be made about patients, which patients, areas or geographic will be affected by some disease. Due to these applications in healthcare industry predictive analytics have received a huge amount of interest from researchers in past few years.

Recent developments in machine learning has enhanced radically the capability of computers to identify and label images, identify and translate speech, play games which involves skills and higher IQ, prediction of diseases and improved decision making over data. In these applications of machine learning, the objective is usually to train a computer to do as humans or better than a human [7]. Traditionally supervised learning algorithms are used for training the model with labelled data and then testing data is used for evaluation using testing data [8, 9].

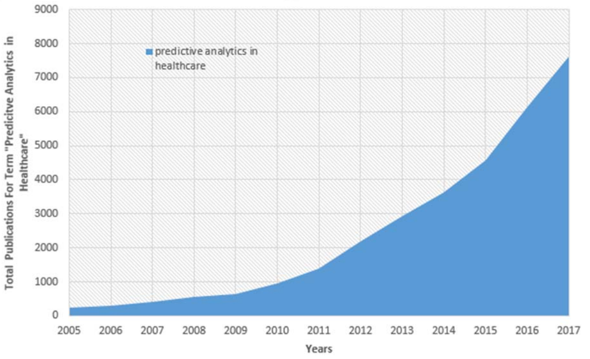

Figure 1. Publications in the area of predictive analytics in healthcare

As different machine learning algorithms are suitable for different size and kind of data and has limitations. This paper discusses the predictive analytics in healthcare. For experiment purpose a large dataset of healthcare is obtained and different machine learning algorithms are applied on the dataset. Performance and accuracy of the applied algorithms is discussed according to the nature of dataset. The objective of the study is to provide enough understanding to reader about how healthcare industry can utilize big data analytics for better decision making or disease prediction. Secondly, performance evaluation of machine learning algorithms in predictive analytics for diabetes disease.

The remaining article is structred as follows. We describe the Related work in Section II. Dataset and atributes of the data are described in Section III. Methodology and machine learning algorthims used are described in Section IV. Results are discussed in the Section V. Finally, Section VI concludes this research work

II. Related work

In recent years, many researchers around the globe worked in big data analytics and predictive analytics in healthcare and other domains, to predict or forecast about the future challenges and opportunities. Taxonomy of big data and analytics is presented in Figure 2. This taxonomy is adopted from [3] and extended in this work. There are different big data sources from where data is coming, then different components and big analytics technologies are given. In this paper we will focus on machine learning for predictive analytics. Research work by different authors is studied as a basis for our research and understanding. In this regard few research papers are discussed below.

In [10], Pisapia et al. used image analysis and machine learning for the prediction of Hydrocephalus. They used the cerebral ventriculomegaly and extracted 77 imaging features. Machine learning algorithm support vector machines was applied on the ventricular features of 25 children. The question was who needed shunts and who did not. Results were obtained and compared. Results shows that every 3 out of 4 children need shunts with 75% sensitivity and 95% specificity. A new fuzzy rule-based classifier is proposed in [11]. Algorithms are designed on the basis of expectation-maximization and fuzzy-rule base classifier for applying analytics and cluster formation. Proposed scheme is compared with existing schemes and results were analysed on the basis of accuracy, response time, false positive rate and computation cost. Results show that proposed technique performs better than Bayes network, multi-layer and decision tables.

In [12], authors predicated the diabetes types, complications and type of treatment which can be given to patients. Predictive analysis algorithm and Hadoop map reduce was used for the prediction and the treatment types. Large data set gathered from different laboratories, clinics, EHR and PHR processed in Hadoop, final results then distributed over different servers according to the geographical locations. Jiang Zheng and Aldo Dagnino in [13] presented a comprehensive survey of literature over big data and analytics. The focus of the authors were to apply machine learning algorithms on industrial power systems and applications for the prediction of faults and power load. In [14], a healthcare prediction system based on Naïve Bayes algorithm is presented. Proposed system discovers and extracts hidden data related to different diseases from disease database. This system allows users to share their health related problems and then using Naive Bayes predict the correct illness. For better prediction of heart diseases in frequent chronic disease outbreak communities, authors streamline the machine learning algorithms in [15]. Authors proposed a new convolutional neural network based multimodal disease risk prediction algorithm. For evaluation of proposed algorithm real life hospital data was collected from central China for the period of 2013-2015. Experiment was done on chronic disease of cerebral infraction. Experiments results shows that for structured data Naive Bayes performs better and for structured and text when combined proposed algorithm, performs better. A proof of concept study is presented in [16]. Because of the clinical importance of sepsis, authors used sepsis mortality as the prediction use case. Data was acquired from four emergency departments for a period of 12 months. Processing and clustering of data was done using K-mean clustering, random forest technique was used for prediction. Logistic regression model and CART were used as traditional model of prediction in emergency care. Results shows that random forest predict more accurate results as compared to other models.

Das *et al.* studied the cases of dengue and malaria in Delhi, India and performed predictive analysis on the data in [17]. Simi *et al.* explored the importance of early detection of female infertility in [18]. Authors in their research work used 26 variables and 8 classes of female infertility, Results identified that Random Forest technique outperformed other techniques and provide 88% accuracy. Lafta *et al.* proposed an intelligent recommender system that assists the patients and practitioners about the short term risk assessment of heart failures in [19]. Authors proposed a heart disease prediction model, according to the results the system also provides recommendations to the patients that about the need of taking some test and visiting a doctor. The main component of recommendation system is based on time Series data analysis algorithm. For evolution of the purposed system real life data was used. Authors conducted a pilot study on group of heart failure patients and gathered data using daily medical readings. There were 7147 records of patient data conducted for the period of May to November 2012. According to the results recommendation accuracy of the proposed system ranges between 75% and 100%. The dataset used contains only few patients and the data and readings taken were only numerical values. Author focused on Diabetes mellitus, a diseases in which body cannot retain level of maltose in blood [20]. System performs

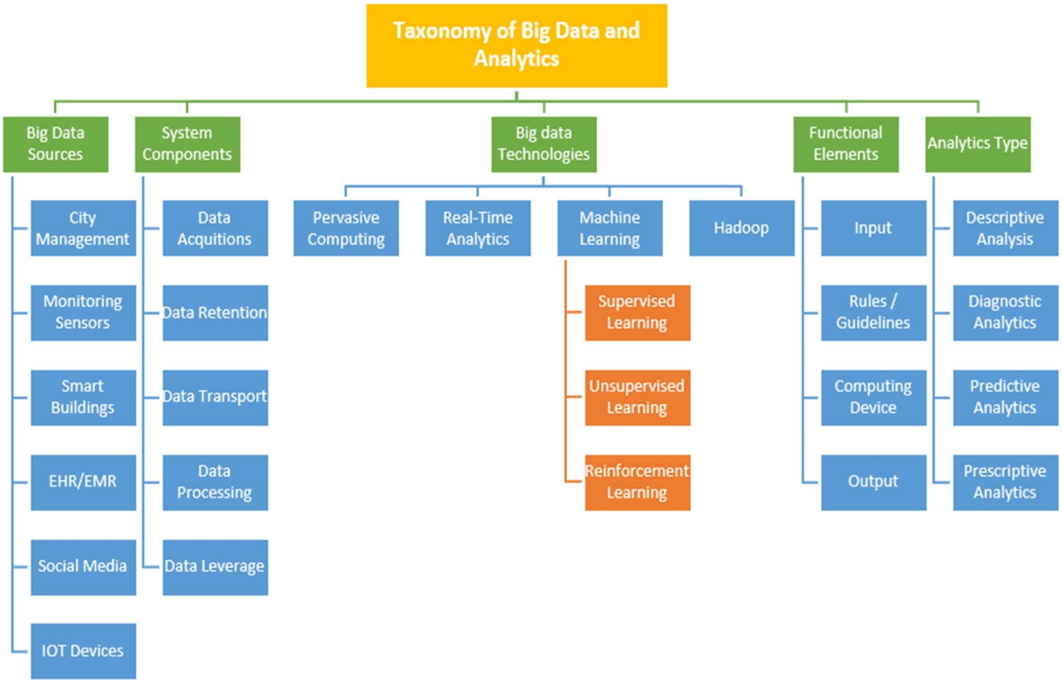

Figure 2. Taxonomy of Big data and analytics

prediction in Hadoop/Map reduce environment using different algorithms. Proposed system predicts the type of diabetic mellitus a person may have. For prediction of patient's status as non CKD or CKD, machine learning classification algorithms were used in [21]. A dataset consisting 400 data records, 25 attributes was taken from UCI repository. Authors used the reduced dataset consisting 14 attributes related to CKD. Using Microsoft Azure Machine Learning Studio, different machine learning algorithms were applied on dataset. Results verify that Multiclass Decision Forest Algorithm performs better and provides 99.1% accuracy. In [22], authors focused on the predicting the survivability of patients with breast cancer. Dataset consisting data of more than 683 cases was obtained from the UCI machine learning repository. There were 26 variables related to the disease in each dataset record. Authors used five machine learning algorithms on the dataset. According to the results support vector machine performed better than other algorithms and provided 97% accuracy. Tele-monitoring data for prediction of asthma exacerbations before their occurrence using machine learning algorithms in [23]. 7001 records submitted by asthma patients was used for training and testing of algorithm. Adaptive Bayesian network, naive Bayesian classifier and support vector machines, three algorithms were used. The study showed that machine learning techniques have significant potential in predicting asthma exacerbations over tele-monitoring data. The next section will describe dataset and attributes we used for disease predictive analytics.

III. Dataset and Attributes

This research paper uses openly available dataset [24] which is downloaded from the UCI machine learning repository. Selected data set is part of a larger data set held by the National Institutes of Diabetes and Digestive and Kidney Diseases. For predictive analytics, many researchers used this dataset in their research work [25-28]. This dataset contains 768 patient records of Pima Indian women with 9 attributes. Table 1 describes the used attributes of the dataset whereas Table 2 describes the basic data statistics. In particular, all the patients in this dataset are females with at least 21 years of age who belongs to Pima Indian heritage. The objective is to predict if a person has diabetes or not based on the diagnostic measurements of the patient.

Table 1. Description of Attributes

Attribute	Description
num_preg	Number of pregnancies (Numeric).
glucose_conc	Plasma glucose concentration (Numeric).
diastolic_bp	Diastolic blood pressure (mm Hg) (Numeric).
thickness	Triceps skin fold thickness (mm) (Numeric).
insulin	2-Hour serum insulin (mu U/ml) (Numeric).
Bmi	Body mass index (Numeric).
diab_pred	Diabetes pedigree function (Numeric)
age	Age (Numeric).
Diabetes	Diabetes or not diabetes (True/False).

Table 2. Data Statistics

Attributes	Count	Mean	STD	Min	Max
num_preg	768	3.84	3.36	0.00	17.00
glucose_conc	768	120.89	31.97	0.00	199.00
diastolic_b	768	69.10	19.35	0.00	122.0

p					0
thickness	768	20.53	15.95	0.00	99.00
insulin	768	79.79	115.2	0.00	846.00
bmi	768	31.99	7.88	0.00	67.10
diab_pred	768	0.47	0.33	0.07	2.42
age	768	33.24	11.76	21.0	81.00
skin	768	0.80	0.62	0.00	3.90

IV. Methodology

Data mining is one of the major and important technology that is currently being used in the industry for performing data analysis and gaining insight into the data. Data mining uses different data mining techniques such as artificial intelligence, machine learning and statistical analysis. In this study, machine learning technique is used for disease prediction. Machine learning provides a pool of tools and techniques, using these tools and techniques raw data can be converted into some actionable, meaningful information by computers. There are four types of machine learning algorithms that are currently being used. Figure 3, shows these four types of machine learning algorithms. Supervised learning involves classification and regression problems. It is used mostly for predictive analytics as it builds a model from data, this data also includes the outcomes or responses. Model is trained using labelled data. Unsupervised learning is used when outcome or responses are unknown, model is trained using unlabelled data. This type of learning is mostly used for pattern detection and descriptive modelling. Unsupervised learning involves clustering problems. Semi supervised learning is combination of both supervised and unsupervised learning. Lastly reinforcement learning aims at using observations gathered by interacting with the environment to take actions which increases the reward or decreases the risks. As this research work evaluates the performance of machine learning algorithms for predictive analytics in healthcare, supervised learning is used in this research work. Figure 4, describes the basic process of supervised learning algorithms. In supervised learning, algorithm or often called model is fed with data for training purpose. This data given includes the input values, often referred as predictor values and correct output values. With the help of this data, model learns the dependencies, patterns and relationships between given features and targeted output value. Once the models learns these patterns, we can use it for predicting the responses against new data.

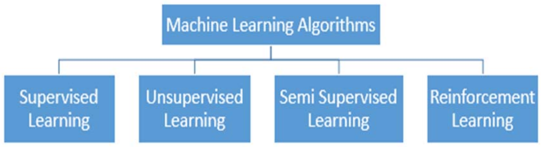

Figure 3. Types of machine learning algorithms

In this paper, six machine learning algorithms are used to predict diabetes disease. These six algorithms are K-Nearest Neighbours (KNN), Naive Bayes (NB), Support Vector Machine (SVM), Decision Tree (DT), Logistic Regression (LR) and Random Forest (RF).

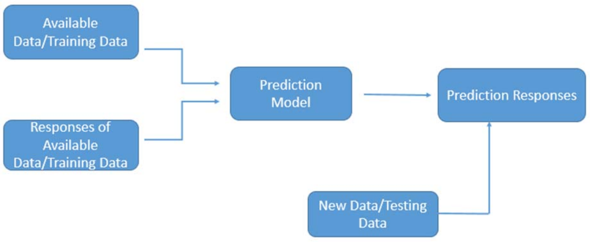
Available Data/Training Data
Responses of Available Data/Training Data
Prediction Model
Prediction Responses
New Data/Testing Data

Figure 4. Supervised learning process

The process used of developing and evaluating the predictive models is shown in the Figure 5. Coding was done in python programing language using tool Enthought Canopy. Enthought Canopy offers a verified scientific and analytical Python package distribution with key important integrated tools for application development, iterative data analysis and data visualization [29].

After obtaining the dataset from UCI machine learning repository. In first step, data pre-processing was performed on the data. For efficient performance and analysis data has to be in structured form. Data was

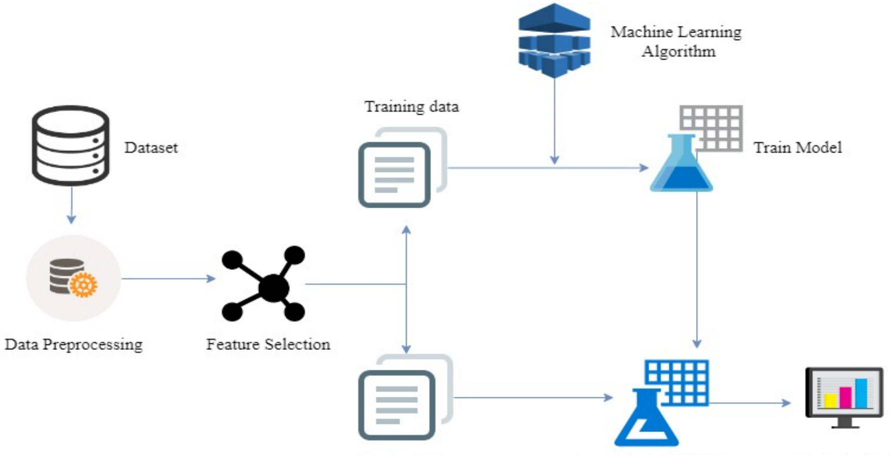

Figure 5. Process followed for evaluation of algorithms

checked for missing values and instances of diabetes are transformed into numerical values e.g. 1 or 0. Through the data analysis it was noticed that number of instances having zero value was quite high. Data imputing was performed on the dataset to overcome missing or zero values. After that feature selection was performed, out of 9 features 8 features were selected. In next step, data was divided into two sets which are training data and testing data. Then machine learning model was trained on that training data for predictions. Once the model trains itself using training data, testing data was used for predicting the responses and checking the accuracy, and lastly the model was evaluated. This process was followed for all 6 machine learning algorithms used in this paper. Experiments were performed and results were obtained. The next section describes the results and our findings in details

V. RESULTS AND DISSCUSSION

In this experimental study, six machine learning algorithims were used. These algorthims are NB, KNN, SVM, LR, DT and RF. All these algorithms were applied on PIMA Indian dataset. Data was divided into two portions, training data and testing data, both these portions consisting 70% and 30% data respectively. All these six algorithms were applied on same dataset using Enthought Canaopy and results were obtained. Predicting accuracy is the main evaluation parameter that we used in this work. Accuracy can be defied using equation 1. Accuracy is the overall success rate of the algorithm.

$$\textbf{Accuracy} = \textbf{(TP+TN) / (P + N)} \qquad \textbf{(1)}$$

All predicted true positive and true negative divided by all positive and negative. True Positive (TP), True Negative (TN), False Negative (FN) and False Positive (FP) predicted by all algorithms are presented in table 3. In our case TP means actual diabetes and predicted diabetes. FN, actual diabetes but predicted to not diabetes. FP, predicted diabetes but actually not diabetes. TN, actual not diabetes and predicted not diabetes.

Table 3. TP, TN, FP and FN predicated by algorithms

Algorithm	TP	FN	FP	TN
DT	61	48	19	103
LR	44	23	36	128
RF	43	30	37	121
NB	52	33	28	118
KNN	41	20	23	137
SVM	37	16	37	141

Figure 6, shows the importance of the features. It can be seen that Plasma glucose concentration have highest importance among other features. Body mass index and age are second and third important features respectively. It can be inferred that these important features plays an important role in prediction and are indicative of if patient will have diabetes or not.

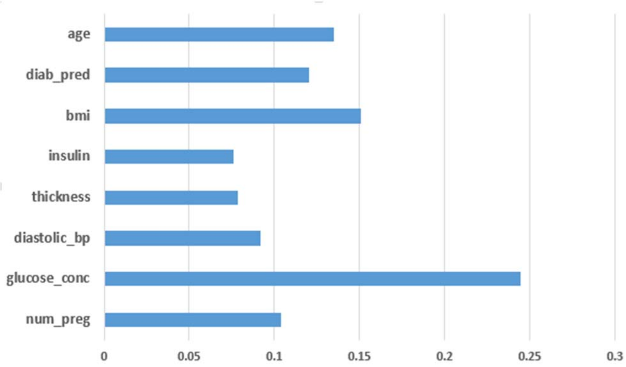

Figure 6. Feature importance

Accuracy of algorithms were measured and presented in Figure 7. LR gives 74% accuracy, SVM gives 77% accuracy, 74% accuracy was achieved by using NB, DT and RF achieved 71% accuracy and KNN achieved 77%. So SVM and KNN achieved highest accuracy which is 77%. From the experimental results obtained, it can be concluded that the Support vector machine and K-nearest neighbor algorithm is appropriated for predicting the diabetes status of patients.

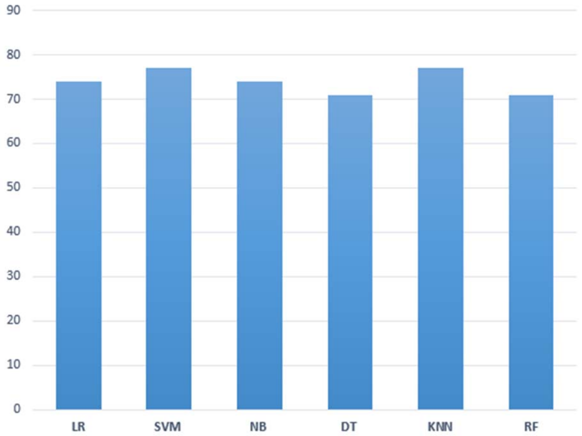

Figure 7. Accuracy of algorithms

VI. CONCLUSION

Predictive analytics in healthcare can change the way how medical researchers and practitioners gain insights from medical data and take decisions. In this paper, we used six popular machine learning algorithms for predictive analytics. These algorithms include SVM, KNN, LR, DT, RF and NB. Predictions were made about diabetes on PIMA Indian dataset consisting 768 records. 8 attributes were selected for training and testing the predictive model. From the experimental results obtained, it can be seen that SVM and KNN gives highest accuracy for predicting diabetes. Both these algorithms provide 77% accuracy which is highest as compared to other four algorithms used in this paper. Therefore, it can be concluded that SVM and KNN is appropriated for predicting the diabetes disease.

Some limitations of this study are the size of dataset and missing attribute values. To build a prediction model for diabetes with 99.99% accuracy, we will need thousands of records with zero missing values. 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 &amp; Company." [Online]. Available: https://www.mckinsey.com/industries/healthcare-systems-and-services/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.

[16] R. A. Taylor *et al.*, "Prediction of In-hospital Mortality in Emergency Department Patients With Sepsis: A Local Big Data-Driven, Machine Learning Approach," *Acad. Emerg. Med.*, vol. 23, no. 3, pp. 269–278, Mar. 2016

[17] S. Das and A. Thakral, "Predictive analysis of dengue and malaria," in *2016 International Conference on Computing, Communication and Automation (ICCCA)*, 2016, pp. 172–176.

[18] M. S. Simi, K. S. Nayaki, M. Parameswaran, and S. Sivadasan, "Exploring female infertility using predictive analytic," in *2017 IEEE Global Humanitarian Technology Conference (GHTC)*, 2017, pp. 1–6.

[19] R. Lafta, J. Zhang, X. Tao, Y. Li, and V. S. Tseng, "An Intelligent Recommender System Based on Short-Term Risk Prediction for Heart Disease Patients," in *2015 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, 2015, pp. 102–105.

[20] S. T. Prasad, S. Sangavi, A. Deepa, F. Sairabanu, and R. Ragasudha, "Diabetic data analysis in big data with predictive method," in *2017 International Conference on Algorithms, Methodology, Models and Applications in Emerging Technologies (ICAMMAET)*, 2017, pp. 1–4.

[21] W. H. S. . Gunarathne, K. D. . Perera, and K. A. D. C. . Kahandawaarachchi, "Performance Evaluation on Machine Learning Classification Techniques for Disease Classification and Forecasting through Data Analytics for Chronic Kidney Disease (CKD)," in *2017 IEEE 17th International Conference on Bioinformatics and Bioengineering (BIBE)*, 2017, pp. 291–296.

[22] S. Jhajharia, S. Verma, and R. Kumar, "Predictive Analytics for Breast Cancer Survivability," in *Proceedings of the Second International Conference on Information and Communication Technology for Competitive Strategies - ICTCS '16*, 2016, pp. 1–5.

[23] J. Finkelstein and I. cheol Jeong, "Machine learning approaches to personalize early prediction of asthma exacerbations," *Ann. N. Y. Acad. Sci.*, vol. 1387, no. 1, pp. 153–165, Jan. 2017.

[24] J. W. Smith, J. E. Everhart, W. C. Dickson, W. C. Knowler, and R. S. Johannes, "Using the ADAP Learning Algorithm to Forecast the Onset of Diabetes Mellitus," *Proc. Annu. Symp. Comput. Appl. Med. Care*, pp. 261–265, 1988.

[25] B. M. K. Prasad, K. K. Singh, N. Ruhil, K. Singh, and R. O'Kennedy, *Communication and Computing Systems : Proceedings of the International Conference on Communication and Computing Systems (ICCCS 2016), Gurgaon, India, 9-11 September, 2016.* CRC Press, 2017

[26] H. Wu, S. Yang, Z. Huang, J. He, and X. Wang, "Type 2 diabetes mellitus prediction model based on data mining," *Informatics Med. Unlocked*, vol. 10, pp. 100–107, Jan. 2018.

[27] D. M. Renuka and J. M. Shyla, "Analysis of Various Data Mining Techniques to Predict Diabetes Mellitus," *Int. J. Appl. Eng. Res. ISSN*, vol. 11, no. 1, pp. 973–4562, 2016.

[28] K. Kayaer and T. Yildirim, "Medical Diagnosis on Pima Indian Diabetes Using General Regression Neural Networks," *Iternational Conf. Artif. Neural Networks Neural Inf. Process.*, pp. 181–184, 2003.

[29] "Canopy | Scientific Python Packages &amp; Analysis Environment | Enthought." [Online]. Available: https://www.enthought.com/product/canopy. [Accessed: 12-May-2018].