Utilisation of Machine Learning Techniques in Testing and Training of Different Medical Datasets

Maad M. Mijwil[1], Israa Ezzat Salem[2] and Rana A. Abttan[3]

[1] Computer Techniques Engineering Department, Baghdad College of Economic Sciences University
Baghdad, Iraq
Email: mr.maad.alnaimiy [AT] baghdadcollege.edu.iq

[1] Computer Techniques Engineering Department, Baghdad College of Economic Sciences University
Baghdad, Iraq
Email: israa.ezzat [AT] baghdadcollege.edu.iq

[1] Computer Techniques Engineering Department, Baghdad College of Economic Sciences University
Baghdad, Iraq
Email: rana.ali.abttan [AT] baghdadcollege.edu.iq

ABSTRACT— ***On our planet, chemical waste increases day after day, the emergence of new types of it, as well as the high level of toxic pollution, the difficulty of daily life, the increase in the psychological state of humans, and other factors all have led to the emergence of many diseases that affect humans, including deadly once like COVID-19 disease. Symptoms may appear on a person, and sometimes they may not; some people may know their condition, and others may neglect their health status due to lack of knowledge that may lead to death, or the disease may be chronic for life. In this regard, the author executes machine learning techniques (Support Vector Machine, C5.0 Decision Tree, K-Nearest Neighbours, and Random Forest) due to their influence in medical sciences to identify the best technique that gives the highest level of accuracy in detecting diseases. Thus, this technique will help to recognise symptoms and diagnose them correctly. This article covers a dataset from the UCI machine learning repository, namely the Wisconsin Breast Cancer dataset, Chronic Kidney disease dataset, Immunotherapy dataset, Cryotherapy dataset, Hepatitis dataset and COVID-19 dataset. In the results section, a comparison is made between the execution of each technique to find out which one is the best and which one is the worst in the performance of analysis related to the dataset of each disease.***

Keywords— Disease, Machine Learning Techniques, COVID-19, Symptoms, Medical Datasets.

1. INTRODUCTION

The doctor or specialist makes analyses of the patient in order to ascertain his condition if there is a disorder in the physical or psychological function that affects the well-being and execution of the patient. The disease is usually associated with specific signs or symptoms that appear to him/her. For example, flu is usually associated with symptoms such as headache, runny nose, and fever. Frequently, some patients do not differentiate between disease and symptoms. Some diseases occur more frequently at certain times of the year. These diseases are also colloquially called seasonal diseases [1]. The most common seasonal illness is bronchus and influenza [2].

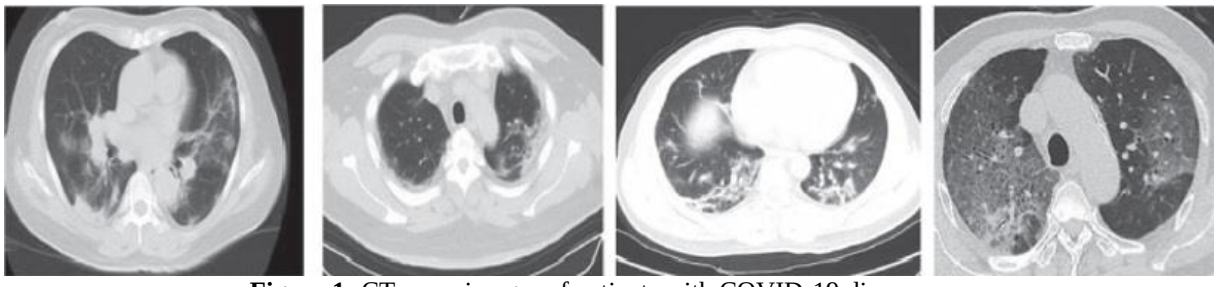

Figure 1: CT- scan images of patients with COVID-19 disease.

Presently, the volume of data is growing dramatically, and its complexity increases day by day. The task of analysing it and finding useful statistics in a traditional way by humans is challenging, and this is why there is an attempt to find suitable techniques to solve such a problem by the computer. In addition, medical data is one of these problems because

this data becomes more complicated with the large spread of diseases worldwide [3]. It has become difficult to control it, especially with the spread of COVID-19 disease and the increase in infections among humans and the increase in deaths [4]. This matter forced doctors and specialists to find techniques that help them in a significant way in diagnosing the injured and determining their condition quickly and accurately. From these techniques are machine learning techniques. Machine learning [5] is evolving and growing in the world of healthcare. Furthermore, healthcare [6] is always one of the most vital areas that witness a remarkable advancement in machine learning techniques. Recently, machine learning has been adopted to predict and analyse medical datasets due to its speed, accuracy, and low cost [7]. For example, it has been widely applied in analysing chest images of patients with the COVID-19 disease [8-11]. These techniques can be trained to look at these images to analyse them, locate the abnormalities, and point at areas where the virus is spread in the human lung, and to give us a high analysis [12]. With these types of advanced technologies, clinicians can be better informed in analysing patient information [13]. As well as it has the ability to predict early diseases such as stroke, breast cancer and many other diseases, which made these techniques of great value to doctors. Figure 1 shows a set of CT- scan images of people with COVID-19 disease [14].

The main contribution of this article is the exhibition of an investigation on the execution of machine learning techniques (Support Vector Machine, C5.0 Decision Tree, K-Nearest Neighbours, and Random Forest) to perform an analysis on a set of binary data that has been chosen from the University of California at Irwin machine learning repository to obtain the best technique with high results in analysing data for each disease so that this technique is supportive for doctors and specialists. This work is conducted by using Python. It is a high-level programming language that Guido Van Rossum invented while working at the Centrum Wiskunde & Informatica Research Centre in 1986. This language is widely used in artificial intelligence.

The following parts of this article are organised as follows: Section two reviews a set of recent studies that apply machine learning techniques to analyse medical datasets earned from UCI machine learning repository. Section three discusses the techniques and materials used in this research. Section four covers the results obtained through experiments as well as the comparison between these techniques. At the end of this article conclusion and future works are advised in Section five.

2. LITERATURE SURVEY

In this section, several previous works of literature that adopt the same views of the current paper and which has an impact on the author on its reading are presented. In addition, the researchers have not found find a similar published study to count the medical datasets chosen from the UCI repository website, and no study that applied the same techniques used in this paper, which make this paper unique.

The start is from a 2016 study conducted by Aswal et al. from India [15], they recommend implementing machine learning techniques (Support Vector Machine, C5.0 decision tree, k- Nearest Neighbour) on a medical dataset from the UCI machine learning repository, namely (Indian Liver Patient Dataset, Hepatitis Dataset, Thyroid Disease Dataset, Lung Cancer Dataset, and Pima Indians Diabetes Dataset). Their research explains that the best execution is the Support Vector Machine. In another paper issued at IEEE Xplore by Islam et al. in 2017 [16], they propose machine learning techniques (K-Nearest Neighbours and Support Vector Machine) to diagnose the breast cancer termed as Wisconsin breast cancer. This study has achieved an accuracy of more than 98% of support vector machine and earned more than 97% accuracy of K-Nearest neighbours. In another article conducted by Cahyani and Muslim [17], they make an improvement in the C4.5 Algorithm for Chronic Kidney Disease Diagnosis by adding two factors which are Discretization and Correlation-based Feature Selection. Their idea achieved success in analysing disease data, as they obtain an accuracy of more than 97%. This study is very impressive. In another study, Eedi and Kolla [18], they propose employing machine learning techniques (K-Nearest Neighbour, Random Forest, Naïve Bayes, Logistic Regression, and Decision Tree,) to detect Breast Cancer Wisconsin Diagnostic. Their research covers Breast Cancer Wisconsin dataset from the UCI machine learning repository. This research discovers the best execution for the random forest technique, with more than 93% accuracy. As for the previous study that will be covered in this section, it is an article conducted by Kumar et al. [19], on the application of one of the machine learning techniques, namely Support Vector machine with Genetic programming, on a dataset from the UCI repository, namely BUPA liver disorder, chronic kidney disease (CKD), fertility, and Wisconsin diagnostic breast cancer (WDBC). In this article, the authors obtain excellent accuracy for BUPA, Fertility, WDBC, and CKD as 75.36%, 85.0%, 99.12%, and 100%, respectively.

3. MATERIALS AND TECHNIQUES

This section is divided into two parts; the first part is about the repository from which the data is taken, and the second part is directed towards techniques that have been utilised in this article. The UCI Machine Learning Repository [20] is a website affiliated with the University of California that includes nearly 600 free datasets to serve researchers and authors in the machine learning community. Meanwhile, these datasets can be used easily with one condition, which is to make a citation for the reference of this data and this repository. The table below presents a concise description of all the datasets utilised in this comparison with their number of attributes and instance.

Table 1: Dataset's description

Datasets	Attributes	Instances
Wisconsin Breast Cancer [21]	32	569
Chronic Kidney disease [22]	25	400
Immunotherapy [23]	8	90
Cryotherapy [24]	7	90
Hepatitis [25]	19	155
COVID-19 [26]	7	14

In the second part, the importance of each technique utilised in this article is concisely discussed, where a set of machine learning techniques are utilised, which are outlined below.

Support Vector Machine (SVM)
SVM [27] is one of the most widespread supervised machine learning techniques invented in 1992 by three scientists: Bernhard Boser, Isabelle Guyon, and Vladimir Vapnik. This classifier is applied in classification and regression and performs operations using linear equations. The classifier has the ability to predict with high accuracy while avoiding overfitting of automatic data. We can summarize them as systems that employ a hypothesis for linear tasks in a high dimensional space and are trained from optimization theory that applies a learning bias derived from statistical learning theory. This technique employs hyperplanes to classify various classes in the dataset and practices various kernels like Poly, Sigmoid, Radial Basis Function, and Linear

C5.0 Decision Tree (C5.0 DT)
C5.0 [28] is an updated and revised version of the C4.5 decision tree. This tree intentionally creates branches in the process of using the Information gain measure. When creating a tree model, the attribute splitting is based on the maximum amount of information gained. The data acquisition mechanism is the process of multiplying the probability of multiplying the class by the probability register of that class. The attribute impurity measure is performed by entropy. Large quantities of information are generated based on calculating the entropy values of either the main tree or sub-tree features. This process continues until a decision is reached that no further division within the tree is required. The most significant characteristic of this version of the decision tree is the ability to create a large group of branches to receive the largest number of data and is also characterized by less memory consumption and faster implementation and support. Unfortunately, this technique does not work with small data.

K-Nearest Neighbours (K-NN)
K-NN [29] is one of the easiest arsenals of machine learning techniques to execute. This technique is based on the classification process, where this process is done by identifying the closest neighbours, for example, querying and using these neighbours to determine the query class. At the beginning of implementation, it is required to specify the value of K, which is set by default 5. Moreover, the group of examples is categorized based on the class of K's closest neighbours. Often it is necessary to take more than one neighbour into account, as these examples are required at run- time, meaning they must be stored in memory, so sometimes this technique is called Memory-Based Classification. A disadvantage of this technique is that it is a lazy learning method because the induction is delayed by the runtime. Besides, this technique uses measurement equations to calculate the distance between two points of the most famous of these equations is Euclidean Distance.

Random Forest (RF)
In 2000, Leo Breiman introduced a scheme that he called a random forest [30] whose goal is to build a set of predictions with other schemas that grow in subspaces that are randomly selected from the data. We can define this technique as a set of tree predictors so that each tree in the scheme depends on the values of a random vector, and samples are collected independently and with the same distribution for all trees in the forest. In addition, this technique has a generalization error that indicates the strength of individual trees in the forest and the continuous relationship between them. Also, the advantage of using a random group is to split each node in the tree into error rates that compare favourably with Adaptive Boosting and also lead to increased noise in it. This technique involves computing internal estimates that give strength, correlation, and error and is employed to prove the response to increasing the number of features used in segmentation. This technique can be used in the regression. This algorithm gives the best accuracy with less processing time for each dataset.

4. EXPERIMENTAL RESULTS

In this section, the results of the analysis of each technique are presented and its execution is evaluated based on various factors like Testing Accuracy, Training Accuracy, Testing Time, Training Time. Figure 2 shows the mechanism of this article in terms of input, processing and output of all medical data. Tables 2 to 5 display the execution evaluation effects for each technique in analysing the medical dataset. The computer specifications in which this work is applied consist of the following: Intel® Core™ i5-1130G7 Processor (4-Core), Hard disk:512GB SSD, 16GB RAM, Python v.3.7 with Spyder IDE v.4.2.1 and running on Windows 10.0 Home build 1904164-bit (last update on February 2021).

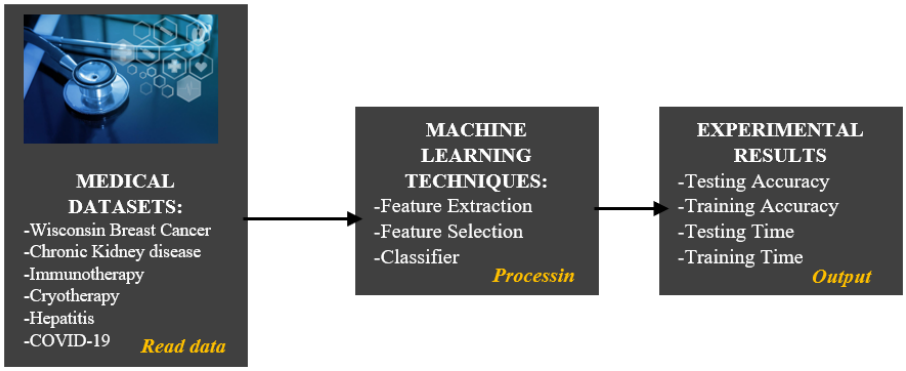

Figure 2: The stages of this work

Table 2: Execution evaluation of SVM

Medical Datasets	Testing Accuracy	Training Accuracy	Testing Time	Training Time
Wisconsin Breast Cancer	0.96114795214	0.9536719818323	0.04125	0.04125
Chronic Kidney disease	0.97578491347	0.9347588136591	0.05125	0.05125
Immunotherapy	0.76261839462	0.7162193529	0.013425	0.013425
Cryotherapy	0.92333333333	0.82131145451	0.013425	0.013425
Hepatitis	0.8361859564	0.7861292128	0.013425	0.013425
COVID-19	0.913968222222	0.891968222222	0.04125	0.04125

Table 3: Execution evaluation of C5.0

Medical Datasets	Testing Accuracy	Training Accuracy	Testing Time	Training Time
Wisconsin Breast Cancer	0.9381429581	0.887820375481	0.034825	0.034825
Chronic Kidney disease	0.9741222222	0.9537388134491	0.05125	0.05125
Immunotherapy	0.9332432782613	0.881323282321	0.015625	0.015625
Cryotherapy	0.976210000	0.890301386712	0.05125	0.05125
Hepatitis	0.88264867336	0.8119202925	0.05125	0.05125
COVID-19	0.71622412555	0.66731424444	0.066125	0.066125

Table 4: Execution evaluation of K-NN

Medical Datasets	Testing Accuracy	Training Accuracy	Testing Time	Training Time
Wisconsin Breast Cancer	0.94102895133891	0.93119309670342	0.066125	0.066125
Chronic Kidney disease	0.985222222222	0.925223452898	0.013425	0.013425
Immunotherapy	0.8132435886611	0.7955811238125	0.013425	0.013425
Cryotherapy	0.98888888888	0.97848589843	0.04125	0.04125
Hepatitis	0.826086956	0.7611940298	0.04125	0.04125
COVID-19	0.66666666666	0.63218467925	0.07125	0.07125

Table 5: Execution evaluation of RF

Medical Datasets	Testing Accuracy	Training Accuracy	Testing Time	Training Time
Wisconsin Breast Cancer	0.8891288722	0.9021282721	0.066125	0.066125
Chronic Kidney disease	0.875444444	0.915243742	0.074875	0.074875
Immunotherapy	0.9846421835	0.9280745516	0.066125	0.066125
Cryotherapy	0.733333333	0.7321862671	0.066125	0.066125
Hepatitis	0.928808192	0.928908288	0.074875	0.074875
COVID-19	0.9888281133	0.9828222111	0.074875	0.074875

5. CONCLUSIONS AND FUTURE DIRECTIONS

In fact, health is an invaluable blessing, and there is a wonderful saying by Anne Wilson Schaef (an American clinical psychologist), who says, “Good health is not something we can buy. However, it can be an extremely valuable savings account”. In this article, machine learning techniques are utilised to analyse medical datasets that have been chosen from the UCI repository. This article purposes to study the effect of each technique in analysing these data, as each group of these data has attributes and instances that differ from the other. Table 6 exhibits the effect of the execution of each technique, as the index included four points, which are excellent execution, good execution, Fair execution, and inadequate execution. In the future, other techniques can be applied in analysing other data or the same data collected in order to see the strength of their implementation in analysing medical data.

Table 6: The effect of executing all techniques

Medical Datasets	Excellent Execution	Good Execution	Fair Execution	Inadequate Execution
Wisconsin Breast Cancer	SVM	K-NN	C5.0	RF
Chronic Kidney disease	C5.0	K-NN	SVM	RF
Immunotherapy	RF	C5.0	K-NN	SVM
Cryotherapy	K-NN	C5.0	SVM	RF
Hepatitis	RF	C5.0	SVM	K-NN
COVID-19	RF	SVM	C5.0	K-NN

6. REFERENCES

[1] Grassly N. C. and Fraser C., “Seasonal infectious disease epidemiology,” *Proceedings. Biological sciences*, vol.273, no.1600, pp: 2541–2550, July 2006. https://doi.org/10.1098/rspb.2006.3604

[2] Tate M. D., Deng Y., Jones J. E., Anderson G. P., Brooks A. G., and Reading P. C., “Neutrophils Ameliorate Lung Injury and the Development of Severe Disease during Influenza Infection,” *The Journal of Immunology*, vol. 183, pp:7441-7450, November 2009. https://doi.org/10.4049/jimmunol.0902497

[3] Pandey S. C., “Data Mining Techniques for Medical Data: A Review,” *In Proceedings of International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES*), pp:1-12, Paralakhemundi, India, 3-5 October 2016. https://doi.org/10.1109/SCOPES.2016.7955586

[4] Jia Q., Guo Y., Wang G., and Barnes S. J., “Big Data Analytics in the Fight against Major Public Health Incidents (Including COVID-19): A Conceptual Framework,” *International Journal of Environmental Research and Public Health*, vol.17, no.6161, pp:1-20, August 2020. https://doi.org/10.3390/ijerph17176161

[5] Jones L. D., Golan D., Hanna S. A., and Ramachandran M., “Artificial intelligence, machine learning and the evolution of healthcare: A bright future or cause for concern?,” *Bone & Joint Research*, vol.7, no.33,pp:223-225, March 2018. https://doi.org/10.1302/2046-3758.73.BJR-2017-0147.R1

[6] Schmidt J., Marques M. R. G., Botti S., and Marques M. A. L., “Recent Advances and Applications of Machine Learning in Solid-State Materials Science,” *NPJ Computational Materials*, vol.5, no.83, pp:1-11, August 2019. https://doi.org/10.1038/s41524-019-0221-0

[7] Battineni G., Sagaro G. G., Chinatalapudi N., Amenta F., “Applications of Machine Learning Predictive Models in the Chronic Disease Diagnosis,” *Journal of Personalized Medicine*, vol.10, no.21, pp:1-11, March 2020. https://doi.org/10.3390/jpm10020021

[8] Pham T. D., “Classification of COVID-19 chest X-rays with deep learning: new models or fine tuning?” *Health Information Science and Systems*, vol.9, no. 2, November 2020. https://doi.org/10.1007/s13755-020-00135-3

[9] Mijwil, M. M., “Implementation of Machine Learning Techniques for the Classification of Lung X-Ray Images Used to Detect COVID-19 in Humans,” *Iraqi Journal of Science*, vol.62, no.6., pp: 2099-2109, 2 July 2021. https://doi.org/10.24996/ijs.2021.62.6.35.

[10] Mijwil, M. M. and Al-Zubaidi, E. A., "Medical Image Classification for Coronavirus Disease (COVID-19) Using Convolutional Neural Networks," *Iraqi Journal of Science*, vol.62, no.8, pp: 2740-2747, 31 August 2021. https://doi.org/10.24996/ijs.2021.62.8.27.

[11] Mijwil, M. M., Alsaadi, A. S, and Aggarwal K., "Differences and Similarities Between Coronaviruses: A Comparative Review," *Asian Journal of Pharmacy, Nursing and Medical Sciences*, vol.9, no.4, pp:49-61. 10 September 2021. https://doi.org/10.24203/ajpnms.v9i4.6696

[12] Borkowski A. A., Viswanadhan N. A., Thomas L. B., Guzman R. D., Deland L. A., and Mastorides S. M., "Using Artificial Intelligence for COVID-19 Chest X-ray Diagnosis," Federal practitioner: for the health care professionals of the VA, DoD, and PHS, vol.37, no.9, pp: 398–404, September 2020. https://doi.org/10.12788/fp.0045

[13] Sidey-Gibbons J. A. M. and Sidey-Gibbons C. J., "Machine Learning in Medicine: A Practical Introduction," *BMC Medical Research Methodology*, vol.19, no.64, pp:1-18, March 2019. https://doi.org/10.1186/s12874-019-0681-4

[14] Mishra A. K., Das S. K., Roy P., and Bandyopadhyay S., "Identifying COVID19 from Chest CT Images: A Deep Convolutional Neural Networks Based Approach," *Journal of Healthcare Engineering*, vol.2020, ID. 8843664, pp:1-7, August 2020. https://doi.org/10.1155/2020/8843664

[15] Aswal S., Ahuja N. J., and Ritika, "Experimental analysis of traditional classification algorithms on bio medical dtatasets," *In Proceedings of International Conference on Next Generation Computing Technologies (NGCT)*, pp:1-6, Dehradun, India, 14-16 October 2016. https://doi.org/10.1109/NGCT.2016.7877478

[16] Islam M., Iqbal I., Haque R., and Hasan K., "Prediction of breast cancer using support vector machine and K-Nearest neighbors," *In Proceedings of International Conference on Region 10 Humanitarian Technology (R10-HTC)*, pp:1-6, Dhaka, Bangladesh,21-23 December 2017. https://doi.org/10.1109/R10-HTC.2017.8288944

[17] Cahyani N., and Muslim M. A., "Increasing Accuracy of C4.5 Algorithm by Applying Discretization and Correlation-based Feature Selection for Chronic Kidney Disease Diagnosis," *Journal of Telecommunication, Electronic and Computer Engineering*, Vol.12 No.1, pp:25-32, March 2020.

[18] Eedi H. and Kolla M. "Machine Learning Approaches for Healthcare Data Analysis," *Journal of Critical Reviews*, vol.7, no.4, pp:806-81,1 February 2020. http://dx.doi.org/10.31838/jcr.07.04.149

[19] Kumar A., Sinha N., and Bhardwaj A., "A Novel Fitness Function in Genetic Programming for Medical Data Classification," *Journal of Biomedical Informatics*, vol. 112, December 2020. https://doi.org/10.1016/j.jbi.2020.103623

[20] Dua D. and Graff C., UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science, 2019.

[21] Street W. N., Wolberg W. H., and Mangasarian O. L., "Nuclear feature extraction for breast tumor diagnosis," *IS&T/SPIE 1993 International Symposium on Electronic Imaging: Science and Technology*, vol.1905, pp:861-870, California, United States, 1993.

[22] Thomas R., Kanso A., and Sedor J. R., "Chronic Kidney Disease and Its Complications," *Primary Care: Clinics in Office Practice*, vol.35, pp: 329–344, 2008. https://doi.org/10.1016/j.pop.2008.01.008

[23] Khozeimeh F., Azad F. J., Oskouei Y. M., M. Jafari, S. Tehranian S., Alizadehsani R., and Layegh P., "Intralesional Immunotherapy Compared to Cryotherapy in The Treatment of Warts," *International Journal of Dermatology*, vol.56, pp:474–478. 2017, https://doi.org/10.1111/ijd.13535

[24] Khozeimeh F., Alizadehsani R., Roshanzamir M., Khosravi A., Layegh P., and Nahavandi S., "An expert system for selecting wart treatment method," *Computers in Biology and Medicine*, vol. 81, pp:167-175, February 2017. https://doi.org/10.1016/j.compbiomed.2017.01.001

[25] Cestnik G., Konenenko I., and Bratko I., Assistant-86: A Knowledge-Elicitation Tool for Sophisticated Users. In: Bratko, I. and Lavrac, N., Eds., Progress in Machine Learning, Sigma Press, Wilmslow, pp: 31-45, 1987.

[26] Wua Y., Chena C., and Chan Y., "The outbreak of COVID-19: An overview," *Journal of the Chinese Medical Association*, vol.83, no.3, pp:217-220. March 2020. https://doi.org/10.1097/JCMA.0000000000000270

[27] Yu W., Liu T., Valdez R., Gwinn M., and Khoury M. J., "Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes," *BMC Medical Informatics and Decision Making*, vol.10, no.16, pp:1-7, March 2010. https://doi.org/10.1186/1472-6947-10-16

[28] Rajeswari S., and Suthendran K., "C5.0: Advanced Decision Tree (ADT) classification model for agricultural data analysis on cloud," *Computers and Electronics in Agriculture*, vol.156, pp:530-539, December 2018. https://doi.org/10.1016/j.compag.2018.12.013

[29] Zhang Z., "Introduction to machine learning: k-nearest neighbors," *Annals of Translational Medicine*, vol.4, no.11, pp:1-7, June 2016. https://doi.org/10.21037/atm.2016.03.37

[30] Biau G., and Editor: Yu B., "Analysis of a Random Forests Model," *Journal of Machine Learning Research*, vol.13, pp:1063-1095, April 2012.