Machine Learning and Statistical Modelling for Prediction of Novel COVID-19 Patients Case Study: Jordan

Ebaa Fayyoumi[1]*, Sahar Idwan[2], Heba AboShindi[3]
Department of Computer Science and Applications
Faculty of Prince Al Hussein Bin Abdullah II for Information Technology
The Hashemite University, P.O. Box 330127, Zarqa 13133, Jordan[1, 2]
Abt Associates,P.O.Box 851275, Sweifieh 11185, Amman, Jordan[3]

***Abstract*—As of December 2019, the world's view on life has been changed due to ongoing COVID-19 pandemic. This requires the use of all kinds of technology to help identify coronavirus patients and control the spread of this disease. In this paper, an online questionnaire was developed as a tool to collect data. This data was used as an input for various prediction models based on statistical model (Logistic Regression, LR) and machine learning model (Support Vector Machine, SVM, and Multi-Layer Perceptron, MLP). These models were utilized to predict potential patients of COVID-19 based on their signs and symptoms. The MLP has shown the best accuracy (91.62%) compared to the other models. Meanwhile, the SVM has shown the best precision 91.67%.**

***Keywords*—Novel COVID-19; machine learning; logistic regression; support vector machine; multi-layer perceptron**

I. Introduction

Coronavirus disease (COVID-19) is an infectious disease caused by newly discovered coronavirus [1]. It is considered as zoonosis which is caused by microbes that are transferred between animals and people [2]. Coronavirus is mysterious since it had other previously reported versions such as SARS-CoV [3] which was transferred from cats to humans and MERS-CoV which is transported from camels to the humans [4].

COVID-19 disease was identified in December 2019 at Wuhan, the capital of China's Hubei province, and spread to the countries of the globe developing an ongoing COVID 2019-2020 pandemic [5]. Based on the World Health Organization (WHO) global COVID-19 outbreak pandemic report, there are 2,858,635 confirmed cases, and 196,295 deaths all over the world till April, 27th 2020. While, 165,379 confirmed cases in Eastern Mediterranean. In Jordan, 449 confirmed cases and 7 deaths were reported [1]. The majority people who get COVID-19 displayed mild to moderate symptoms and recovered without special treatment [1]. This virus causes a few noticeable symptoms for its patients such as coughing, high fever and pneumonia which can be utilized to detect the disease in possible patients [6].

Jordan highlighted their surveillance to prospectively have early diagnosis for new COVID-19 cases. Hence, Jordan conducted nationwide unprecedented actions on March 2020 to contain the spread of disease such as large-scale quarantine, extensive controls on travels, social distancing, continuous monitoring of suspected COVID-19 cases, and blocking areas in order to decrease the number of infected cases. Fig. 1 shows the number of COVID-19 confirmed patients in Jordan between March 3rd, 2020 and April 27th, 2020. Yet, it is uncertain whether these subsequent policies have had an impact on the containment of epidemic and what is coming in the future? Accordingly, it is crucial to examine the epidemic progression in the globe by predicting the new COVID-19 cases from the most common symptoms which could effectively control the spread of disease.

Recently, many researchers tackled COVID-19 in their research to prevent the spread of it. Naudé, W. [7] discussed several fields where Artificial Intelligence (AI) can be utilized to influence the fights against COVID-19 such as early warning and alerts, tracking and prediction, data dashboards, diagnoses and prognosis, treatments and cures and social control. He concluded that data shortage or having too much information is considered as an obstacle for using AI against COVID-19.

To prevent the spreading of the COVID-19 many predication models have been utilized by officers and leaders in different countries to generate the appropriate decisions and regulations which help to overcome this pandemic. Authors of [8] explored using of the machine learning for shaping the exponential growth of COVID-19. They concluded that Multi-Layer Perceptron and adaptive network-based fuzzy inference system can be used as a useful tool to handle the epidemic. Other researchers [9] discussed the integration of an improved mathematical modelling in machine learning with the cloud computing. This envisages the expansion of the COVID-19 which follows the exponential dispersion. Machine learning can be developed to predict the extent of COVID-19 by using the high-speed computations in the cloud computing.

Having COVID-19 pandemic in a short period of time raises the need to study the behavior of this virus and the infected patients. The researchers of [10] selected the Prophet Algorithm as the most suitable predictive algorithm. This is done based on what is the analytical question that they need to address in their study and how the predictive algorithms can be approved to get the best results [10].

*Corresponding Author

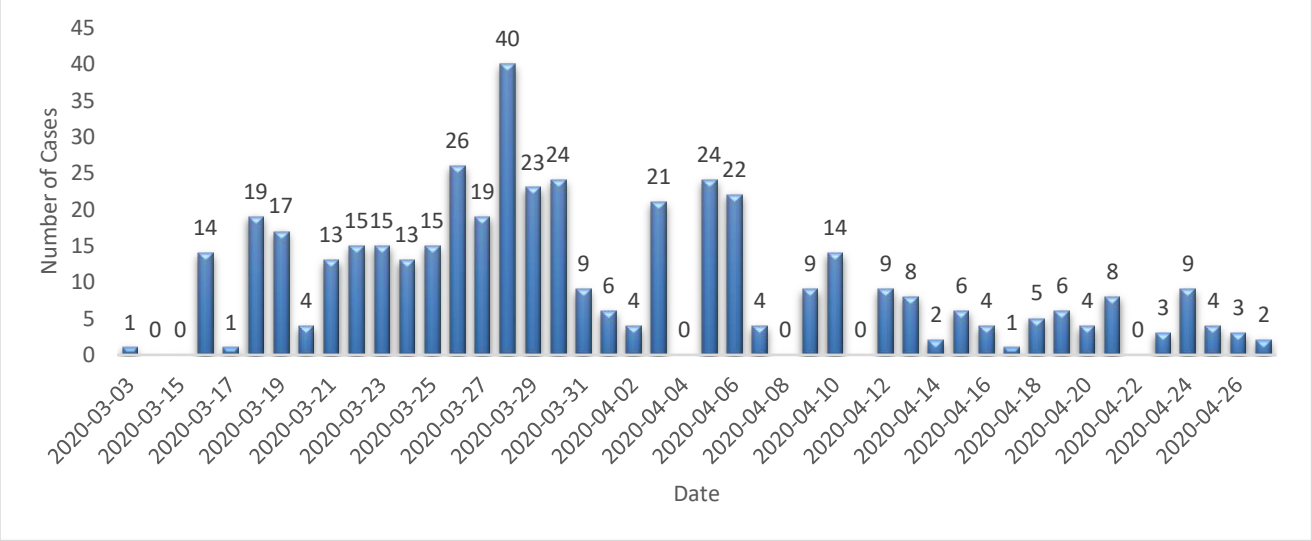

Fig. 1. New COVID-19 Cases between 3-3-2020 to 27-4-2020 in Jordan.

Yadav, D. et al. [11] investigated the foresee of the COVID-19 in different countries including United States of America. The foresee achieved by invoking the machine learning data-driven Prophet time series that analyzed the infected, active and cured cases to outburst the predication.

Large amount of the information is available in the internet and presented numerically and graphically. Thus, utilizing statistical analysis and machine learning are essential to provide better understanding of the results and generate informed decision to meet community, national and international challenges in many fields like medical, business, economics, web search engine, Facebook, spam filters and commerce [12].

To the best of our knowledge, the mathematical model and machine learning model have not been used to predict the infected cases of novel COVID-19 based on the signs and symptoms. There is an urgent need globally and in Jordan particularly to screen patients quickly due to limited availability of the Polymerise Chain Reaction tests. The main objective of this paper is to establish a reliable trusted model to predict the potential patients of COVID-19 by using either statistical or machine learning models. The following procedure had been adopted to achieve the outlined objective:

- Generating a questionnaire to collect data from individual citizens according to their health states during the last two weeks.
- Employing the statistical modelling and machine learning methods individually to assess the obtained data from the questioner in order to predict health condition for people in different cities in Jordan.
- Evaluating the performance of each model to choose the most appropriate one to our domain problem.

This paper is organized as follows: The methodology is presented in Section II, while the experimental results and discussion are reported in Section III. Finally, Section IV draws the conclusion and future works.

II. Methodology

A cross-sectional quantitative study was conducted to build up a reliable trusted model to forecast COVID-19 diagnosis from the signs and symptoms that participants had. The signs and symptoms of novel COVID-19 used in this study were obtained from Jordan[1]. Our strategy includes four processing stages, namely data collection, data preprocessing, classification, and performance evaluation. The classification stage can be accomplished either by building statistical model or by invoking machine learning model. In the statistical model we used Logistic Regression (LR), while in machine learning model we invoked Support Vector Machine (SVM), and Mulit-Layer Perceptron (MLP). Finally the performance of each classifier is quantified. A block diagram of the proposed work is illustrated in Fig. 2.

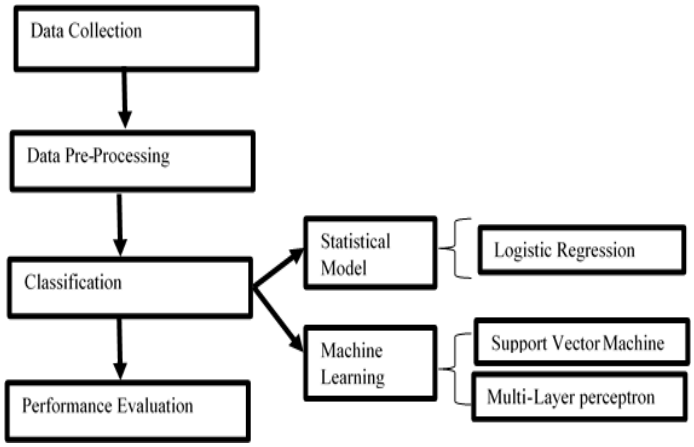

Fig. 2. Block Diagram of Novel COVID-19 Predictor Model.

[1]Jordan is divided geographically into three regions: North, Middle and South. The middle region of Jordan includes four cities: Amman, Zarqa, Madaba and Al-balqa. According to the Jordanian Department of Statistics, population of the middle region of Jordan was estimated as 63.5 % of all population [20]. The questionnaire has been an online survey questionnaire and all eligible participants from all regions can fulfill the survey.

A. Data Collection

Prior to starting data collection, the ethical approval was obtained from the Institutional Review Board at the Hashemite University (HU-IRB: 2020/2019/7/1). Moreover, the eligible participants' approval to participate in this study was obtained using online consent form before starting the survey. Consent form included information about the purposes of the study, significance, benefits and risks. All participants were aware that participating in this study was voluntary and they could withdraw from it at any time they want without any physical or emotional harm. Also, the consent form has a clear statement that the participation was totally voluntary without any risk of participation or withdrawal from the study. Beside the decision to participate or not in this study would not affect their treatment plan.

Furthermore, participants were informed about the privacy and confidentiality of this study. This is achieved by understanding that data will be used only for the research purposes and no one other than the researcher can access them. Besides understanding that the questionnaire contains code number not their names.

Data were collected using online survey questionnaire on April 2020. The participants who have the willingness to participate in this study was asked if they met the criteria of eligibility [2], then they signed the consent form before answering the survey. The contact information of the primary researcher was available in the online survey in order to answer or clarify any misunderstanding of questions. As well as, the survey needs as an average 3 minutes to be fulfilled.

B. Data Pre-Processing

The target population in this study is the potential patient for novel COVID-19. The size of the collected sample was 120. It is worthy to highlight that there were 15 rejected samples due to incomplete or inconsistent (ex. Age: adult) answered survey. The reliability of the accepted sample 105 was 87.50%. The purpose of the questionnaire was to utilize a machine learning and statistical models to predict novel COVID-19 potential patient based on the signs and symptoms they have.

Our real novel COVID-19 data set present imbalanced classification problem, the majority class is referred to as the negative outcome (Negative PCR Test) with 64 out of 105 (60.95%), and the minority class is referred to as the positive outcome (Positive PCR Test) with 41out of 105 (39.05%).

The collected dataset consists of thirteen attributes. One class attribute, A, and twelve test feature attributes that present the signs and symptoms of the candidate patient for novel COVID-19: age, smoker, positive chest x-ray, fever, sore throat, aches and pain, dry cough, nasal congestion, absence of smell, diarrhea or vomiting, and breathing difficulty. It is worth to mention that all the attributes are binary categorical type except the age attribute which is integer; therefore the age attribute is normalized to be aligned with other attributes. The descriptive statistics for the collected samples are shown in Table I, while the description of symptoms for novel COVID-19 (N=41) and non-novel COVID-19 (N= 64) are shown in Table II.

C. Classification

Classification is a process related to categorization, the process in which negative and positive outcomes are recognized and understood. This is usually achieved by either utilizing several statistical models or invoking various machine learning models.

Machine learning is widely used in different applications due to its powerful prediction and high accuracy while statistical analysis show cases emphasis in models that can be interpreted easily with uncertainty and precision [13].

In this paper, we used the following models:

- Logistic Regression (LR): is a predictive analysis which computes the probability of one dependent variable based on the observations of one or more independent variables. It is the most widely used algorithm for solving problems in different scales. It works properly with the minor instances of multicollinearity and in high dimensional datasets [14]. LR provides the direction of the relationship as well as the degree of the significance of the predicator [15]. It is very easy to implement and explain the results using this model.
- Support Vector Machine (SVM): is a simple and effective neural network. It is utilized for prediction and classification in order to increase predictive correctness by excluding over-fit to the data. It generates various classes by establishing the best hyperplane in multidimensional space in order to reduce the error [16]. It is more applicable in large dimensional spaces where the border of the partition between classes are defined clearly.
- Multi-Layer Perceptron (MLP): is a standard type of neural network which is used for prediction and classification problems. This is achieved by building relationships between inputs and outputs, and computing the required patterns. It consists of a set of neurons in different layers with a set of adaptive weights [17]. The number of hidden layers determines whether the machine learning model is deep or shallow.

D. Performance Evaluation

Evaluation metric plays predominant role in quantifying the performance of the various models. Generally speaking, metrics include comparing the expected class label to the predicted class label. There are many standard metrics that are used for evaluating the predictive models such as accuracy, sensitivity, specificity, and precision. Those metrics are easily calculated from the confusion matrix, since our classification problem is a binary imbalance type as follows [18]:

[2]The eligible criteria for participants are (a) their age should be 18 years or above (b) they can read and understand Arabic Language (c) they did a test of Polymerise Chain Reaction (PRC) before two weeks or more, and finally (d) they have no critical illnesses at the time of data collection (hemodynamically stable).

TABLE I. THE DESCRIPTIVE STATISTICS FOR THE COLLECTED SAMPLES

	Min	Max	Median	Mean	Variance
Positive PCR	0	1	0	0.390	0.240
Age	19	75	40	40.629	128.120
Gender	1	2	1	1.362	0.233
Smoker	0	1	1	0.581	0.246
Positive X-ray	0	1	0	0.095	0.087
Fever	0	1	0	0.286	0.206
Sore Throat	0	1	1	0.819	0.150
Aches and Pain	0	1	1	0.771	0.178
Dry Cough	0	1	1	0.571	0.247
Nasal Congestion	0	1	0	0.238	0.183
Absence of Smell	0	1	0	0.143	0.124
Diarrhea or Vomiting	0	1	0	0.305	0.214
Breathing	0	1	0	0.343	0.227

TABLE II. DESCRIPTION OF SYMPTOMS FOR COVID-19 (N=41) AND NON COVID-19 (N= 64)

Symptoms	**COVID-19 n (%)**	**Non COVID-19 n (%)**
Aches and pains	41 (100 %)	40 (62.5 %)
Sore throat	39 (95.1 %)	47 (73.4 %)
Dry cough	36 (87.8 %)	24 (37.5 %)
Difficulty of breathing	27 (65.9 %)	9 (14.1%)
Diarrhea or vomiting	27 (65.9 %)	5 (7.8%)
Fever	24 (58.5 %)	6 (9.4%)
Nasal Congestion	14 (34.1 %)	11 (17.2 %)
Absence smell	15 (36.6 %)	0 (0 %)
Abnormal chest x-ray	7 (17.1 %)	3 (4.7 %)

Accuracy: The percentage of test set tuples that are correctly classified.

$$Accuracy = \frac{\text{Correct Prediction}}{\text{Total Prediction.}}$$

Error Rate: The percentage of test set tuples that are incorrectly classified.

$$Error = \frac{\text{Incorrect Prediction}}{\text{Total Prediction.}}$$

It is also well-known as the complement of classification accuracy as: Error = 1-Accuracy.

Sensitivity: A metric measures the ability to correctly detect patient who do have the disease (The portion of actual positives that are correctly identified). It is well-known as a type I error.

$$Sensitivity = \frac{\text{True positive}}{\text{True Positive} + \text{False Negative}}$$

Specificity: A metric measures the ability to reject healthy patient without a condition. (The portion of actual negatives that are correctly identified). It is well-known as type II error.

$$\text{Specificity} = \frac{\text{True Negative}}{\text{False Positive} + \text{True Negative}}.$$

The domain of our data set is related to diagnostic novel COVID-19 test, so it is important to have a highly sensitive[3] and highly specific[4] test. Therefore; these two metrics can be mutually joint into single score that balances both concerns so called geometric mean (G_Mean) as follows:

$$G\_Mean = \sqrt{\text{Sensitivity} * \text{Specificity}}$$

Precision: A metric measures the number of positive class predictions that actually belong to the positive class.

$$Precision = \frac{\text{True positive}}{\text{True Positive} + \text{False Positive}}$$

III. EXPERIMENTAL RESULTS AND DISCUSSION

Python programming language has been used to build the various models in this study. It is an interpreted, high-level, general-programming language. Python is described as a "batteries included" language due to its comprehensive standard library [19].

All models were built using 10-fold cross validation and a tolerance value was set to 0.001. All test attributes are used in building each model. In the LR a logit model was built with a confidence interval 95% and the cut point was set to 0.5. In the SVM the Sequential Minimum Optimization algorithm, (SMO), parameters were set as follows: c = 1.0, Epsilon = 1×10^{-12}, and the kernel type was chosen to be linear. Finally, the MLP is considered as a shallow deep learning model. It was built with one hidden layer of 12 neurons and the activation function was set to sigmoid one because it exists between zero and one.

The performance evaluations for the statistical model LR and the machine learning models SVM and MLP are illustrated in Table III. It is clearly shown that the machine learning model competes the statistical model with respect to the accuracy. The percentage accuracy of the LR model was as high as 85.00%, while it was equal to 90.00% and 91.62% by using the SVM and MLP, respectively. The reason behind this is referred to the fact that statistical model used to characterize the relationship between the test attributes and the class attribute (outcome variable) to assess the model's legitimacy. Therefore; the accuracy of prediction by using this inference model is not that robust comparable to the machine learning models. The MLP has the best accuracy compared to the other techniques because of its ability to capture very complex features in the hidden layers and the usage of the nonlinear stimulation functions. In general, the machine learning models sacrifice interpretability for the prediction power.

As we know, the metrics of the sensitivity and specificity reflect completely different aspects of the prediction model. As mentioned earlier, the domain of our data set is related to diagnostic novel COVID-19 test, so it is important to have a

[3]Highly sensitive test infrequently overlooks an actual positive PCR.
[4]Highly specific test infrequently registers a positive classification for anything that is not PCR of the test.

highly sensitive and highly specific test. Since sensitivity and specificity are two conflict metrics, researchers have concentrated on either of these metrics. While this is an accepted practice, we feel that a more fair index would be one which considers both of them simultaneously which is known as geometric mean (G_Mean). It is clearly noted that the shallow deep learning MLP model scored the highest value of geometric mean which was equal to 90.73 while the statistical model LR scored the minimum value of the geometric mean which was equal to 88.53. The SVM model scored 89.53 as geometric mean value. Table III shows the sensitivity, specificity, and geometric mean for each model.

Precision in any prediction model refers to how closely the observed value to the model's prediction. As shown in Table III, the SVM model is the most precise model, it reached up to 91.67%, due to its high correlation shape which can be detected by our data set On the other hand, the most imprecise model was the LR model, and it reached up to 66.67%. This obviously reducing usefulness of the prediction and making the mistake very costly.

Based on these results, we could conclude that MLP represents a useful model of novel COVID-19 detection based on the sign and symptoms. On the contrary, LR is not an appropriate technique to predict potential COVID-19 patients due to relative error obtained and low precision.

As a summary a machine learning model competes the statistical model in predicting the infected and non- infected cases of novel COVID-19 based on the signs and symptoms. Utilizing the machine learning model will help in screening patients quickly due to limited availability of the Polymerise Chain Reaction tests in Jordan.

TABLE III. EVALUATION THE PERFORMANCE METRICS FOR VARIOUS CLASSIFICATION TECHNIQUES

Metrics	**Statistical Method**	**Machine Learning**	
	Logistic Regression (LR)	*Support Vector Machine (SVM)*	*Multi-Layer perceptron (MLP)*
Accuracy (%)	85.00	90.00	91.62
Sensitivity (%)	100.00	91.67	87.80
Specificity (%)	78.57	87.50	93.75
G_Mean	88.64	89.53	90.73
Precision (%)	66.67	91.67	90.00

IV. CONCLUSION AND FUTURE WORKS

Several measures are utilized to combat the extent of the novel COVID-19 in the world and in Jordan specifically. The obtained data were analyzed to aid the government in predicting potential patients of COVID-19 in order to save lives or to provide best health care. Several predication models were invoked based on either statistical or machine learning approaches to envisage the citizens' health status.

We foresee two avenues for future work. The first avenue is to implement these models in hospitals in Jordan with aim of identifying COVID-19 patients in a quick, safe method leading to a decrease in the rapid spread of the virus. The second involves applying the studied models on bigger size dataset and study its corresponding parameters.

ACKNOWLEDGMENT

Authors would like to thank Institutional Review Board at the Hashemite University. We are also extremely grateful to the anonymous Referees of this paper, for their valuable and constructive comments.

REFERENCES

[1] W. H. Organization, "Coronavirus disease 2019 (COVID-19): situation report," 2020.

[2] "CDC Works 24/7," 2020. [Online]. Available: https://www.cdc.gov/. [Accessed April 25,2020].

[3] "SARS CORONA Virus," [Online]. Available: https://www.sinobiological.com/research/virus/sars-coronavirus-overview. [Accessed April 25,2020].

[4] I. Al-Turaiki, M. Alshahrani and T. Almutair, "Building predictive models for MERS-CoV infections using data mining technique," Journal of Infection and Public Health (2016), vol. 9, p. 744—748, 2016.

[5] A. Du Toit, "Outbreak of a novel coronavirus," Nature Reviews Microbiology, vol. 18, no. 3, pp. 123-123, 2020.

[6] W. Wang, J. Tang and F. Wei, "Updated understanding of the outbreak of 2019 novel coronavirus (2019 - nCoV) in Wuhan, China," Journal of medical virology, vol. 92, no. 4, pp. 441-447, 2020.

[7] W. Naudé, "Artificial Intelligence against COVID-19: An Early Review," IZA Discussion Papers 13110, 2020.

[8] S. F. Ardabili, A. Mosavi, P. Ghamisi, F. Ferdinand, A. R. Varkonyi-Koczy, U. Reuter, T. Rabczuk and P. M. Atkinson, "COVID-19 Outbreak Prediction with Machine Learning, 2020040311," Preprints 2020, doi: 10.20944/preprints202004.0311.v1.

[9] S. Tuli, S. Tuli, R. Tuli and S. S. Gill, "Predicting the Growth and Trend of COVID-19 Pandemic using Machine Learning and Cloud Computing," Internet of Things, p. 100222, 2020.

[10] P. N. Mahalle, N. P. Sable, N. P. Mahalle and G. R. Shinde, "Predictive Analytics of COVID-19 Using Information, Communication and Technologies," Preprints 2020, 2020040257 (doi: 10.20944/preprints202004.0257.v1).

[11] D. Yadav, H. Maheshwari and U. Chandra, "Outbreak prediction of covid-19 in most susceptible countries," Global Journal of Environmental Science and Management, vol. 6, no. 4, p. 2020.

[12] S. Das, A. Dey, A. Pal and N. Roy, "Applications of Artificial Intelligence in Machine Learning: Review and Prospect," International Journal of Computer Applications, vol. 115, pp. 31-41, 2015.

[13] A. Palmer, J. Rafael and E. Gervilla, Data Mining: Machine Learning and Statistical Technique, 2011.

[14] C. Molnar, Interpretable Machine Learning: A Guide for Making Black Box Models Explainable, Lulu, 1st edition, 2019.

[15] J. F. and J. Fang, "Why Logistic Regression Analyses Are More Reliable Than Multiple Regression Analyses," Journal of Business and Economics, vol. 4, no. 7, pp. 620-633, 2013.

[16] V. Jakkula, "Tutorial on support vector machine (SVM)," School EECS,, Washington State Univ, Washington, DC, USA, 2011.

[17] L. Dencelin and R. Thenmoley, "Analysis of multilayer perceptron machine learning approach in classifying protein secondary structures," Biomedical Research (2016) Computational Life Sciences and Smarter Technological Advancement, pp. 166-S173, 2016.

[18] J. Han and M. Kamber, "Data mining: concepts and techniques," San Francisco, Morgan Kaufmann Publishers, 2001.

[19] E. MATTHES, Python crash course: a hands-on, project-based introduction to programming, 2016.

[20] DOS (2015), "Jordanian Department of Statistics. Retrieved 10 4, 2016, from," [Online]. Available: http://dos.gov.jo/dos_home_e/main/linked-html/jordan_no.htm. [Accessed April 2020].