INTERNATIONAL JOURNAL OF SYSTEMS APPLICATIONS, ENGINEERING & DEVELOPMENT
DOI: 10.46300/91015.2020.14.25 Volume 14, 2020

Death/Recovery Prediction for Covid-19 Patients using Machine Learning

Omar Mohamed Atef
Department of Computer Engineering
University of Sharjah
U16104886@sharjah.ac.ae

Ali Bou Nassif
Department of Computer Engineering
University of Sharjah
anassif@sharjah.ac.ae

Manar AbuTalib
Department of Computer science
University of Sharjah
mtalib@sharjah.ac

Qassim Nassir
Department of Electrical engineering
University of Sharjah
nasir@sharjah.ac.ae

***Abstract*—Covid19 is a newly discovered corona virus that has been officially announced as a pandemic by the World Health Organization in March 2020. It is a new virus in the medical field that has no specific treatment and no vaccines until this moment. Covid19 is spreading very fast as the medical systems over the world are not able to hospitalize all the patients which lead into a significant increase in the number of the virus death. This work uses machine learning models to predict which patient has a higher probability of death. Three different algorithms such as multilayer perceptron, support vector machine and K nearest neighbor were used in this work. The accuracies achieved were between 92% to 100% with MLP, SVM and KNN. SVM achieved the highest accuracy. The models were evaluated through precision, accuracy, recall and F measure.**

***Keywords*—Corona virus, Covid19, Multilayer perceptron, support vector machine, K nearest neighbor.**

I. Introduction

According to World Health organization, Covid-19 deaths now are more than 500K people and more than 16M case around the world[1]. The main reason of that is because physicians are not able to hospitalize all the patients, so they mostly choose the patients that have higher probability of surviving. Many people think that only old people have higher probability of death, but this is not always the case. The higher probability of risk depends on several factors such as the symptoms, age, dates of symptoms, date of hospitalization and date of confirmation of covid-19 and if the patient has any other disease.

Recently, Artificial intelligence models have achieved a great success in the medical imaging due to its high capability of feature extraction [1], [2]. In addition, Artificial intelligence is also used to predict diseases in order to early detect the disease so that the treatment will be much easier and much safer [3].

In this work, we are utilizing Machine learning to predict which patients have higher priority for hospitalization as they have higher probability of death. We are building some classification algorithms to predict that, and we will evaluate the performance through some evaluation metrics.

The "Novel Corona Virus 2019 Dataset" available on Kaggle has over 10k of data from different patients in different countries with different attributes[2]. The main goal of this work is to predict the probability of death, recovery, stable or severe by using three machine learning algorithms.

II. Related Work

Recently, death/recovery prediction was not done before. Despite the huge amount of work done in Covid19 in a short time.

In [1], the authors developed an automatic framework to detect Covid19 from the CT scan of the patient's chest. They used deep learning model. The authors used a dataset contains 4352 chest CT scans from 3322 patients. Their model achieved an accuracy higher than 90%.

While in [4], the authors present a novel methodology using machine learning to reliably forecast Covid19 activity in Chinese provinces. They did not focus on diagnosing or treatment; the main objective was to forecast the behaviour of the virus in Chine.

In [5], the authors aim to investigate about Covid19 patients and how the virus affected their digestive system. They analysed 204 patients with full laboratory data and imaging data. Then they found that 50.5% reported digestive symptoms.

III. Technical Background

A. Corona Virus Covid-19

Covid-19 is a new corona virus. It was firstly discovered in China in December 2019 and then it spread all over the world. The cause of covid-19 is still not clear until now. But its cause is the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) [6].

[1] World Health Organization, Coronavirus disease Pandemic available: https://www.who.int/emergencies/diseases/novel-coronavirus-2019

[2] https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset

Coivd19 has no specific treatment or vaccines until this moment but in hospitals, they are trying to treat only the symptoms. The virus infects the respiratory system and in the critical cases the virus can damage the lungs and the patient dies. In 11 March 2020 World Health Organization "WHO" has characterized Covid19 as pandemic.

Covid19 symptoms are not yet all known but some common symptoms are like normal flu symptoms e.g. fever, headache, fatigue, etc. Recent researches focuses on monitoring smell and taste feelings and how they are related to Covid19 patients [7].

B. *Machine Learning*

Machine learning is a field of Artificial Intelligence that tries to make the machine think like a human, it mimics the behaviour of the Human brain [8].

Machine Learning is working by training the mathematical models on the training set and as a result it gains experience and can predict and take decisions with being explicitly programmed.

Machine Learning can be divided into supervised and non-supervised learning. Supervised learning is when the model is trained on a specific dataset giving the input and the expected outputs. Through some mathematical process it will be able to adjust some parameters to be able to predict and take decisions in the testing phase.

Non-supervised learning works in opposite way. The models are not given the expected output, so they are used in other applications such as clustering.

C. *Artificial Neural Network*

Artificial Neural Network "ANN" is a supervised machine learning algorithm that used mainly for classification problems as a classifier[9]–[11].

The simplest form is composed of input layer and output layer. This model is called single perceptron and it is used for easy classification problems.

Neural Networks mostly used architecture is composed of 3 parts and it is called Multilayer perceptron "MLP". Input, hidden and output layers. It operates by some process that mimics the way of how human brain works. The input layer contains the inputs of the system. These inputs will be multiplied by the weights which are randomly initially and then they will be adjusted in the training stage. The hidden layer takes the results which are the inputs multiplied by the weights. In the hidden layer, there are several neurons. Each neuron contains a function e.g. sigmoid, tanh, etc. then the output of the hidden layers will be multiplied by some weights then they will be the input to the output layer. The output layer maps the result to the closest class according to the activation function in it.

D. *Support vector machine*

Support vector machine "SVM" is another supervised machine learning algorithm that mainly used in classifications and can also be used in regression through some modifications [12].

The mechanism of SVMs is by separating the classes from each other using hyperplanes. The classes are represented as data points that are 'n' dimensional feature vector and the hyperplane has a geometric shape that occupies 'n-1' dimensions.

The simplest form of the SVMs is where the data is linearly separable. Where two parallel hyperplanes are used to separate the classes such that the distance between the hyperplanes is maximum to minimize the error of classification. The observation that lies on the hyperplanes is known as support vectors.

E. *K-Nearest Neighbours*

K nearest neighbour "KNN" is a supervised machine learning algorithm that can be used in classification and regression [13], [14]. KNN classifies new cases based on the similarity measure e.g. distance functions.

Choosing the optimal value for K is best done by first inspecting the data. In general, a large K value is more precise as it reduces the overall noise but there is no guarantee. Cross-validation is another way to retrospectively determine a good K value by using an independent dataset to validate the K value. Historically, the optimal K for most datasets has been between 3-10. That produces much better results.

Data pre-processing and filtering is the most consuming time stage in any Machine learning/Deep learning project.

The Novel Corona virus 2019 dataset has over 10k of entries, but it contains many missing values.

To begin with, we removed ID of each patient, the city were the patient lives, longitude and latitude of the city, all travel history data, death or discharge and some other references columns. We left the most important attributes for our system which are age, gender, symptoms, date on set symptoms, date hospitalization, date confirmation and outcome. The outcome is the dependent variable "outputs" that contains 4 classes which are "Discharge", "Death", "Stable", "Severe" and the rest of the attributes are independent variables "inputs".

Secondly, all data were converted to numeric data to be easy to handle missing data using different techniques for instance: mean, median and mode. We used to change the gender into 1 for female and 2 for male. Then using the imputer function in Sklearn in Python3, we used to fill the missing gender data with the mean of the available data and take the ceil of the results to make them either 1 or 2. Also, another imputer used the mean to fill the missing age data. Then for the missing dates, mean or median method would not work so all missing dates were removed. Unfortunately, the dataset has been reduced a lot to be 256 rows.

Finally, symptoms column contains for each patient several symptoms or no symptoms. Each symptom has been extracted to be in 1 column individually and the attribute will be 1 is true symptom or 0 false symptom e.g. column fever for patient 2 is 1 indicating that he has fever. Symptoms used in this work are 5 symptoms which are fever, malaise, chills, cough, fatigue.

F. *Features*

1. Age: age of the patient.
2. Sex:

Identify applicable funding agency here. If none, delete this text box.

- 1: Female.
- 2: Male

3. Fever:
    - 1: Positive
    - 0: Negative
4. Malaise:
    - 1: Positive
    - 0: Negative
5. Chills:
    - 1: Positive
    - 0: Negative
6. Cough:
    - 1: Positive
    - 0: Negative
7. Fatigue:
    - 1: Positive
    - 0: Negative
8. Syms hosp: days between symptoms appearance and hospital admission.
9. Syms conf: days between symptoms appearance and confirmation of covid19.
10. Outcome: dependent attribute consists of 4 classes:
    - Discharge: Patient is recovered.
    - Death: Patient Died
    - Stable: Patient is in stable condition.
    - Severe: patient in critical condition.

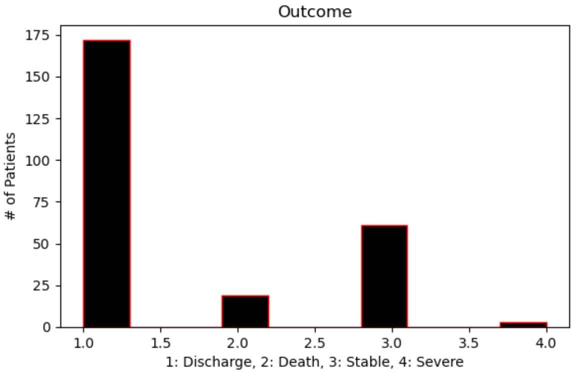

Figure 1: Outcome data

The output data are distributed as shown in Fig. 1 the classes are not equally distributed most of the data are discharge and stable while death and severe are small.

IV. Methodolgy

A. Feature selection

In this paper, feature selection was done to get the correlation among the variables. Table 1 is showing the correlation between the variables.

Table 1: Correlation between variables

	age	sex	outcome	fever	malaise	headache	chills	cough	fatigue	syms_hosp	syms_conf
age	1	-0.0063	0.273058	0.009971	-0.02842	0.023482	0.009038	-0.0106	0.118847	0.109634	0.199019
sex	-0.0063	1	0.007838	-0.03812	0.05553	-0.1127	-0.09312	0.050502	-0.23992	0.05666	-0.06067
outcome	0.273058	0.007838	1	-0.19939	-0.04149	-0.11901	0.050789	-0.09719	-0.07531	-0.21728	0.038009
fever	0.009971	-0.03812	-0.19939	1	-0.03784	0.298444	0.278605	0.506934	0.221155	0.049354	0.040768
malaise	-0.02842	0.05553	-0.04149	-0.03784	1	-0.01129	-0.01054	0.152304	-0.01332	0.041197	0.018155
headache	0.023482	-0.1127	-0.11901	0.298444	-0.01129	1	-0.03024	0.053603	0.293971	-0.16276	-0.02029
chills	0.009038	-0.09312	0.050789	0.278605	-0.01054	-0.03024	1	0.407803	0.200596	0.104757	-0.00929
cough	-0.0106	0.050502	-0.09719	0.506934	0.152304	0.053603	0.407803	1	0.022137	0.046479	0.067718
fatigue	0.118847	-0.23992	-0.07531	0.221155	-0.01332	0.293971	0.200596	0.022137	1	-0.06146	0.162333
syms_hosp	0.109634	0.05666	-0.21728	0.049354	0.041197	-0.16276	0.104757	0.046479	-0.06146	1	0.789457
syms_conf	0.199019	-0.06067	0.038009	0.040768	0.018155	-0.02029	-0.00929	0.067718	0.162333	0.789457	1

There is high correlation between the fever and cough. In addition, highly correlation between symptoms to hospital and symptoms confirmation.

B. Model Design

In this work, Sklearn in Python3 was used to build the three machine learning models. The data was split randomly into 70% training and 30% testing.

Table 2 shows the three models with the best parameters for each of them after simulating these parameters in Python using PyCharm.

Table 2: Tested parameters

	Parameters tested	Best	Best Testing Acc
MLP	1-tanh-sgd-5 Acc = 74% 2-tanh-sgd-10, Acc = 65% 3-relu-sgd-10, Acc = 64% 4-relu-lbfgs-10, Acc= 97% Alpha in all of them = 1e-5	4	98.7%
SVM	1-Kernel: Linear Acc = 100% 2-Kernel: Polynomial Acc = 71.5% 3-Kernel: Sigmoid Acc = 68%	1	100%
KNN	1-K=1, Acc = 95% 2-K=2, Acc = 79% 3-K=3, Acc = 76%	1	95%

As shown in the table, very high testing accuracy achieved between 95% to 100%.

C. *Performance*

The parameters used to evaluate each model was through accuracy, recall, F-measure, precision. These metrics extracted from the confusion matrix generated from each model.

The equation of each metric is represented as:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

$$Recall = \frac{TP}{TP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

Table 3: Accuracy for each model

Model \ Metric	Accuracy
Multilayer perceptron	99%
Support Vector Machine	100%
K Nearest Neighbor	95%

$$Fmeasure = \frac{2 * Precision * Recall}{Precision + Recall}$$

Where TP are the true positives which are the correctly classified patients. While TN are the true negatives which are patients that does not belong to a specific class and classified correctly.

The FP are the false positives which are data wrongly classified as positives of a specific class. While FN are false negatives which are patients belongs to a specific class and classified as does not belong to that class.

V. Results and Discussions

a) *Multilayer precptron:*

The MLP performance was very high as shown in confusion matrix and the performance metrics. The precision recall and F measure of class 1,2 is 1 supporting 54,4 samples from classes 1,2 respectively which is excellent. In addition, class 3 achieved 0.95 precision, 1 recall and 0.97 for F1 measure supporting 18 samples. The 4th class achieved 0 supporting only 1 class. This is expected because in the whole dataset class 4 counts are available only 4 times. The overall accuracy was very high as shown here it reached 0.99.

b) *Support Vector Machine:*

SVM performance was the highest among the three models. As mentioned before using linear kernel, the model reached testing

The classification accuracy is 100%. The precision, recall and F1-measure are 1 for all classes and accuracy is 100%. The reason of that is that the data after filtering and removing missing data was less than 300 row which is very low amount of data.

c) *K nearest neighbour:*

The algorithm achieved its highest accuracy at K=1. It was the lowest among the 3 models. The precision is 0.98 for class 1 and the recall was 0.94 while the f1measure was 0.96. This is considered good among the 54 counts. However, in the 2nd class, it achieved 0.67 precision and this is very low due to the low amount of the 2nd class among the testing data while the recall was 1 and the f-measure was 0.80. in the 4th class it achieved 1 in the 3 metrics. The 3rd class it achieved 0.94 in the 3 metrics since it classified 18 counts.

d) *Overall Metrics evaluation:*

Table 3 below shows the accuracy of our three models. The overall metrics are represented in Fig. 2 using the average precision, recall and F measure of each class. Based on the results, SVM was the best among the three models.

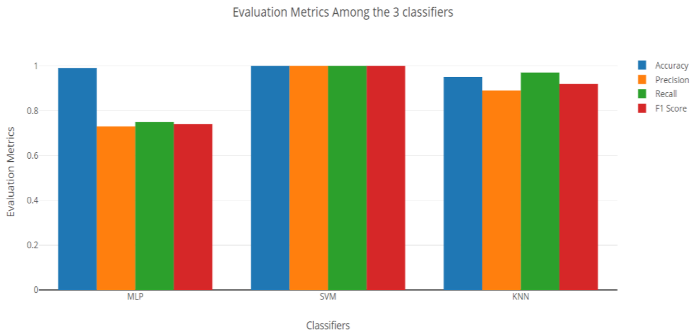

Figure 2: Evaluation Metrics

VI. Conclusion

To sum up, this work compares three different classification algorithms which are MLP, SVM and KNN to classify 4 classes of Covid19 patients. The Performance was very high, and this is due to the small amount of data used. The data was very large but with a lot of missing data and after filtering the data it becomes about 300 rows. This work was able to predict the patients that has higher risk of death or critical condition according to several patient's data from different countries.

For future work we are planning to search for another datasets and to apply different machine learning models.

References

[1] L. Li *et al.*, "Artificial Intelligence Distinguishes COVID-19 from Community Acquired Pneumonia on Chest CT," *Radiology*, p. 200905, Mar. 2020.

[2] R. Hamoudi, M. Bettayeb, A. Alsaafin, M. Hachim, Q. Nassir, and A. B. Nassif, "Identifying Patterns of Breast Cancer Genetic Signatures using Unsupervised Machine Learning," in *2019 IEEE International Conference on Imaging Systems and Techniques (IST)*, 2019, pp. 1–6.

[3] X. Zhou *et al.*, "A Comprehensive Review for Breast Histopathology Image Analysis Using Classical and Deep Neural Networks," *IEEE Access*, vol. 8, pp. 90931–90956, 2020.

[4] D. Liu *et al.*, "A machine learning methodology for real-time forecasting of the 2019-2020 COVID-19 outbreak using Internet searches, news alerts, and estimates from mechanistic models," *ArXiv*, Apr. 2020.

[5] L. Pan *et al.*, "Clinical characteristics of COVID-19 patients with digestive symptoms in Hubei, China: A descriptive, cross-sectional, multicenter study," *Am. J. Gastroenterol.*, vol. 115, no. 5, pp. 766–773, May 2020.

[6] K. J. Clerkin *et al.*, "COVID-19 and Cardiovascular Disease," *Circulation*, vol. 141, no. 20. Lippincott Williams and Wilkins, pp. 1648–1655, May-2020.

[7] C. H. Yan, F. Faraji, D. P. Prajapati, C. E. Boone, and A. S. DeConde, "Association of chemosensory dysfunction and COVID-19 in patients presenting with influenza-like symptoms," *Int. Forum Allergy Rhinol.*, vol. 10, no. 7, pp. 806–813, Jul. 2020.

[8] O. T. Ali, A. B. Nassif, and L. F. Capretz, "Business intelligence solutions in healthcare a case study: Transforming OLTP system to BI solution," in *2013 3rd International Conference on Communications and Information Technology, ICCIT 2013*, 2013, pp. 209–214.

[9] A. B. Nassif, L. F. Capretz, and D. Ho, "Estimating software effort using an ANN model based on use case points," in *Proceedings - 2012 11th International Conference on Machine Learning and Applications, ICMLA 2012*, 2012, vol. 2, pp. 42–47.

[10] A. B. Nassif, D. Ho, and L. F. Capretz, "Towards an early software estimation using log-linear regression and a multilayer perceptron model," *J. Syst. Softw.*, vol. 86, no. 1, pp. 144–160, 2013.

[11] A. B. Nassif, "Software Size and Effort Estimation from Use Case Diagrams Using Regression and Soft Computing Models," University of Western Ontario, 2012.

[12] M. Lataifeh, A. Elnagar, I. Shahin, and A. B. Nassif, "Arabic audio clips : Identification and discrimination of authentic Cantillations from imitations," *Neurocomputing*, vol. 418, pp. 162–177, 2020.

[13] A. B. Nassif, O. Mahdi, Q. Nasir, M. A. Talib, and M. Azzeh, "Machine Learning Classifications of Coronary Artery Disease," in *2018 International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP)*, 2018, pp. 1–6.

[14] C. López-Martín, Y. Villuendas-Rey, M. Azzeh, A. Bou Nassif, and S. Banitaan, "Transformed k-nearest neighborhood output distance minimization for predicting the defect density of software projects," *J. Syst. Softw.*, vol. 167, p. 110592, Sep. 2020.