Real-time Covid-19 Detection based on Symptoms using Machine Learning Models

Youssef Mellah, Zakaria Kaddari, Mohammed Ghaouth Belkasmi, Noureddine Rahmoun and Toumi Bouchentouf
Mohammed First University, LARSA/SmartICT Laboratory, ENSAO, Oujda, Morocco

Keywords: COVID-19, SARS-Cov-2, PCR, X-Ray, Machine-Learning, World Health Organization.

Abstract: The rapid spread of coronavirus has led to the 2019 global coronavirus disease pandemic (COVID-19). In this sense, people suspected of having this virus must know quickly if they are infected so that they can receive appropriate treatment, isolate themselves and inform people with whom they have been in close contact. This situation has attracted worldwide attention, developing exact methods for the identification and isolation of patients infected with SARS-CoV-2. However, those methods (as PCR test, chest X-ray and CT scan images... etc.) take time to get the exact result of the patient. In this paper, we propose different classifier methods using Machine Learning for real-time COVID-19 detection based only on symptoms. These methods can achieve an accuracy of up to 98.0%, evaluated on a DataSet based on significant COVID-19 symptoms according to the World Health Organization (WHO).

1 INTRODUCTION

Coronavirus is a specific type of virus, which depending on the inherited property, grows and replicates. The severity of this virus is due to its nature of spreading very quickly. Common signs of disease are fever, cough, respiratory symptoms, shortness of breath, and difficulty breathing. The disease, in more severe cases, can cause kidney failure, severe acute respiratory syndrome, pneumonia, and even death.

Standard recommendations for limiting the spread of this virus include covering the mouth and nose when coughing and sneezing, washing hands regularly, and taking a bib.

In this sense, and to limit the spread of the pandemic, research is popping out to give guidance in this domain. Authors in Reference (Shan, 2020) used the deep learning-based segmentation method to identify regions of interest in CT Images of lungs for quantifying COVID-19.

Authors (Gozes, 2020) used 2D and 3D deep learning techniques and obtained 99.6% accuracy in the context of COVID-19. The work (Wang, 2020) used a bidirectional Recurrent Neural Network with and attentional mechanisms in the context of COVID-19. Authors in the paper (Xu, 2020) come out with 86.7% accuracy on benchmark datasets using deep learning in the context of coronavirus. In addition, other researches use X-Ray images to detect COVID-19. The literature is growing and many efforts are underway in the scientific community to limit the adverse effects of this virus.

From our side, in this paper, we propose to incorporate Machine Learning models for real-time COVID-19 detection based only on symptoms. We applied these models to a DataSet built base on World Health Organization (WHO) 1 reports. This dataset was published on Kaggle2, containing 5,435 possible combinations, each with seven significant symptoms of COVID-19 according to the same organization. By doing that, we achieve 98.0% as our best accuracy reached by Artificial Neural Network.

2 SETTING UP MACHINE LEARNING CLASSIFIERS

Classification is the way to predict the given data class. Each method attempt to recognize a pattern that best matches the relationship between data.

The objective of training is to have a predictive model that predicts the correct class, especially on data not seen during the training phase. The

Mellah, Y., Kaddari, Z., Belkasmi, M., Rahmoun, N. and Toumi, B.
Real-time Covid-19 Detection based on Symptoms using Machine Learning Models.
DOI: 10.5220/0010731100003101
In *Proceedings of the 2nd International Conference on Big Data, Modelling and Machine Learning (BML 2021)*, pages 206-210
ISBN: 978-989-758-559-3

classification is a method of grouping a pattern from the data in input. There are diverse algorithms for solving a classification problem such as K-neighbours classifier, decision tree classifier, and gradient boosting classifier, without forgetting neural networks, which have shown great success.

2.1 Random Forest Classifier

The Random forest (Culter, 2012) is a supervised learning algorithm. The idea is to build a “forest” like a set of decision trees, generally trained with combining several models with the “bagging” method, which helps in increasing the general precision.

This algorithm can be used for both regression and classification problems, which make it useful and powerful.

2.2 Logistic Regression Classifier

Logistic regression (Wright, 1995) is a predictive technique. It aims to build a model making it possible to predict / explain the values taken by a qualitative target variable (most often binary, we then speak of binary logistic regression; if it has more than 2 modalities, we speak of polychromous logistic regression) from a set of quantitative or qualitative explanatory variables (coding is necessary in this case).

2.3 Decision Tree Classifier

This decision support or data mining tool allows you to represent a set of choices in the graphic form of a tree. It is one of the most popular supervised learning methods for data classification problems. Concretely, a decision tree models a hierarchy of tests to predict a result.

The possible decisions are located at the ends of the branches (the "leaves" of the tree) and are reached based on decisions made at each stage.

2.4 Naïve Bayes Classifier

The Naive Bayesian classification (Leung, 2007) method is a supervised machine learning algorithm that classifies a set of observations according to rules determined by the algorithm itself. There is a theory called Bayes, where the name of the algorithm comes from.

This classification tool must first be trained on a training dataset, which shows the expected class according to the inputs during the learning phase, the algorithm develops its classification rules on this data set, in order to apply them secondly to the classification of a prediction data set.

2.5 Support Vector Machine Classifier

The main idea behind the Support Vector Classifier (Hearst, 1998) is to find a decision boundary with maximum width that can classify both classes. Maximum margin classifiers are extremely sensitive to outliers in training data, which makes them quite lame. Choosing a threshold that allows classification errors is an example of the Bias-Variance tradeoff that affects all machine learning algorithms.

2.6 K-Nearest Neighbours Classifier

In artificial intelligence, more precisely in machine learning, the k nearest neighbours (Peterson, 2009) method is a supervised learning method. In abbreviated form k-NN or KNN, from English k-nearest neighbors.

In this context, we have a training database made up of N "input-output" pairs. To estimate the output associated with a new input x, the k nearest neighbors method consists of taking into account (identically) the k training samples whose according to a defined distance, input is closest to the new input x.

For example, in a classification problem, we will retain the most represented class among the k outputs associated with the k inputs closest to the new input x.

2.7 Gradient Boosting Classifier

Gradient Boosting (Friedman, 2002) classifiers are a category of machine learning algorithms that combine multiple learning models to create a stronger one. Decision trees are generally used when increasing gradients. Gradient enhancement models are popular due to their efficiency in classifying complex data sets and have recently been used to win many Kaggle Data Science competitions.

2.8 Artificial Neural Network Classifier

First, the neural network is a concept. It's not physical. The concept of Artificial Neural Networks (ANN) (Wang, 2003) was inspired by biological neurons. In a biological neural network, several neurons work together, receive input signals, process information, and trigger an output signal. The artificial intelligence neural network is based on the same model as the biological neural network.

Neural networks are trained with a multitude of input data coupled with their respective output data. They then calculate the output data, they compare it to the known actual output data and constantly update themselves to improve the results (if necessary).

There are several types of ANN, such as recurrent neural network used in natural language processing, convolutional neural network used in image processing, and hybrid ones. Thy have been very successful and outperform most other algorithms, especially with new advancements like the attention mechanism and the architecture of Transformers.

3 IMPLEMENTATION

3.1 Environment

To make all Machine Leaning Classifiers, we use sklearn, Numpy, and Pandas. All packages are managed through pip. Weused Jupyter Notebook as the code editor.

3.2 Steps for Setting up the Various Machine Learning Classifiers

3.2.1 Dataset Preprocessing

The DataSet is already cleaned and there are no missing values. Using StandardScaler, we scale various feature value ranges in standard format.

3.2.2 Dataset Splitting

We split the DataSet into:

- Training Set: 80% from the whole DataSet
- Test Set: 20% from the whole DataSet

3.2.3 Make and Train the Various Models

We have set up all the models mentioned below, then we trained and evaluated them in order to compare them and see which classifier gives the best results. The training is made using a model.fit() method.

3.2.4 Evaluate the Various Models

We evaluate our models on the Test Set to see the score and the generalizability of them, by comparing predicted and expected outputs. Using sklearn, we get the predictions as follow:

```
Begin
 Prediction=
 model.predict(inputs_features)
END
```

Also, the accuracy, precision, and other scores are calculated between the predicted and the expected output of all the various MC classifier models by employing classification reports from sklearn.metrics. The evaluation is made by the following snippet code:

```
Begin
 print(classification_report(expected,
 predicted);
END
```

4 RESULTS AND ANALYSIS

4.1 Evaluation and Results

The different statistical metric used for evaluating our models are described as follow (see Table 1 for obtained results):

- Accuracy
- Recall
- Precision
- F1 score
- AUC

Table 1: Results of the evaluation of the various Models.

Models	Accuracy	F1	Precision	Recall	AUC
Artificial Neural Network	**0.980**	**0.987**	0.989	0.986	**0.971**
K-Neighbors Classifier	0.977	0.986	0.989	0.982	0.969
Gradient Boosting Classifier	0.977	0.986	0.984	**0.987**	0.961
Support Vector Machines	0.977	0.986	0.984	**0.987**	0.961
Random Forrest Classifier	0.974	0.984	0.985	0.982	0.961
Logistic Regression	0.962	0.977	0.968	0.986	0.925
Decision Tree Classifier	0.948	0.967	0.972	0.963	0.925
Naïve Bayes	0.759	0.824	**1.000**	0.701	0.850

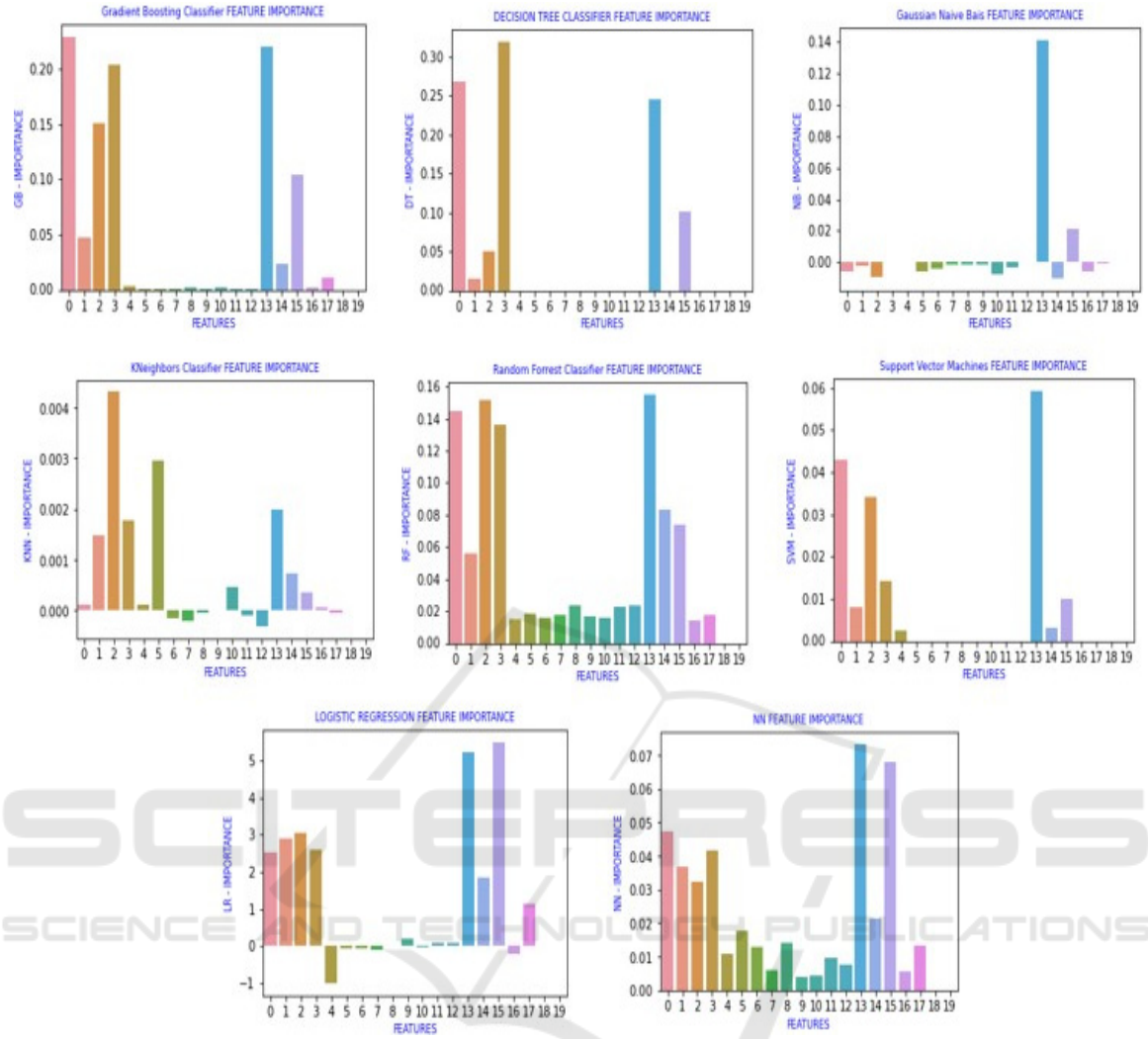

Figure 1: Features Importance of the various Machine Learning Classifiers

Also, we are interested in features impotence, which plays a vital role in a predictive modelling project, Figure 1 shows the importance of the features given by the various classifiers.

4.2 Analysis and Discussion

Artificial Neural Network (ANN) gave the highest score of 0.980 in terms of precision metrics, compared to all other classifiers.

The same thing for F1 and AUC, ANN provides the highest score compared to other ML classifiers. In terms of Precision, the Naïve Bayes Classifier reaches the top one achieving a score of one.

While Gradient Boosting Classifier and the Support

Vector Machines are given top scores on Recall with a minimal difference (0.001) compared to ANN.

As shown in Figure 1, the majority of ML classifiers do not take into account all features, except Random Forest and Neural Network ones so that we can consider this last as the more powerful one compared to all models, in terms of accuracy as well as features importance.

5 CONCLUSIONS

The exponential rate of spread of the COVID-19 pandemic requires the implementation of a solid and effective strategy for the detection of patients affected by the virus, especially in terms of detection time. In this sense, we have implemented machine learning models for real-time detection for more efficiency and time savings. We have found that the model

based on artificial neural networks gives more precision than other classical classifiers, reaching 98.0% on a DataSet based on symptoms. In addition, it takes into consideration all the features (symptoms) which make it the most optimal.

REFERENCES

Shan, F., Gao, Y., Wang, J., Shi, W., Shi, N., Han, M., ... & Shi, Y. (2020). Lung infection quantification of COVID-19 in CT images with deep learning. arXiv preprint arXiv:2003.04655.

Gozes, O., Frid-Adar, M., Greenspan, H., Browning, P. D., Zhang, H., Ji, W., ... & Siegel, E. (2020). Rapid ai development cycle for the coronavirus (covid-19) pandemic: Initial results for automated detection & patient monitoring using deep learning ct image analysis. arXiv preprint arXiv:2003.05037.

Wang, Y., Hu, M., Li, Q., Zhang, X. P., Zhai, G., & Yao,
N. (2020). Abnormal respiratory patterns classifier may contribute to large-scale screening of people infected with COVID-19 in an accurate and unobtrusive manner. arXiv preprint arXiv:2002.05534

Xu, X., Jiang, X., Ma, C., Du, P., Li, X., Lv, S., ... & Li, L. (2020). A deep learning system to screen novel coronavirus disease 2019 pneumonia. Engineering, 6(10), 1122-1129

Cutler, A., Cutler, D. R., & Stevens, J. R. (2012). Random forests. In Ensemble machine learning (pp. 157-175). Springer, Boston, MA.

Wright, R. E. (1995). Logistic regression.

Leung, K. M. (2007). Naive bayesian classifier. Polytechnic University Department of Computer Science/Finance and Risk Engineering, 2007, 123-156.

Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J., & Scholkopf, B. (1998). Support vector machines. IEEE Intelligent Systems and their applications, 13(4), 18-28.

Peterson, L. E. (2009). K-nearest neighbor. Scholarpedia, 4(2), 1883.

Friedman, J. H. (2002). Stochastic gradient boosting. Computational statistics & data analysis, 38(4), 367-378.

Wang, S. C. (2003). Artificial neural network. In Interdisciplinary computing in java programming (pp. 81-100). Springer, Boston, MA.