PREDICTING THE SEVERITY OF NEW SARS-COV-2 VARIANTS IN VACCINATED PATIENTS USING MACHINE LEARNING

MEROUANE ERTEL[1],AZEDDINE SADQUI[2], SAID AMALI[3], INTISSAR MAHMOUDI[4], YOUNES BOUFERMA[5],NOUR-EDDINE EL FADDOULI[6]

[1,2]Informatics and Applications Laboratory (IA), Faculty of Sciences, Moulay Ismail University, Morocco

[3]Informatics and Applications Laboratory (IA), FSJES, Moulay Ismail University, Morocco

[4]Faculty of Medicine and Pharmacy, Sidi Mohamed Ben Abdellah University, Fes, Morocco
[5]Faculty of Legal, Economic and Social Sciences, Moulay Ismail University, Meknes, Morocco
[6]RIME Team, MASI Laboratory, E3S Research center EMI, Mohammed V University, Morocco

Email:[1]m.ertel@edu.umi.ac.ma,[2]a.sadqui@umi.ac.ma,[3]s_amali@yahoo.com,[4]dr.mahmoudi.intissar@gmail.com, [5]bahaebik@gmail.com, [6]noureddine.elfaddouli@um5.ac.ma

ABSTRACT

Given the increasing number of COVID-19 cases and the risk of new variants, early prediction of disease severity in critical care patients is essential to optimize treatment options. In this study, we set up an experiment on 236 patients infected with COVID-19 and hospitalized at the Sidi Said hospital in Meknes, Morocco.

This work proposes a new multivariate classification model to predict which patients admitted to hospital with COVID-19 will require special care (oxygen therapy, intensive care, resuscitation) or will die following an abrupt deterioration in their state of health. This model will help healthcare professionals (doctors) make decisions about recommending appropriate medical treatments to patients. A comparative study of different multivariate machine learning algorithms (Support Vector Machine (SVM), K-nearest neighbor (KNN), Decision Tree (DT) and Random Forest (RF)) is also presented in this article. The result obtained shows that the SVM classifier is a reliable, powerful and efficient algorithm to predict the level of risk of patients contaminated with COVID-19.

Keywords: *Covid-19; Clinical Decision Support; Machine Learning; Ordinal Classification, Multi-Class Classification; Personalized Medicine*

1. INTRODUCTION

The 2019 coronavirus (COVID-19) pandemic first emerged in China in December 2019 and quickly spread around the world. On March 11, 2020, the World Health Organization classified the COVID-19 outbreak as a global pandemic. In September 2020, the number of deaths worldwide exceeded one million [1].

From the beginning of the Covid-19 spread until these days of slow return to normal, the Moroccan government has taken all the necessary precautions to preserve the health of the population as part of its anti-pandemic program. Moreover, Morocco is one of the most advanced African countries in terms of vaccination, with 63.2% of Moroccans receiving two doses of AstraZeneca, Sinopharm , Johnson & Johnson and Pfizer vaccines since the campaign began on January 28, 2021 [2].

Despite the fact that the epidemiological crisis has been controlled in Morocco to a large extent, the waves of the pandemic of COVID-19 have experienced a variation of daily cases of coronavirus plus its severity varied, lately the variant Omicron, more contagious, causes an important influx of patients of severe or critical to hospitals. As a result, the 3rd dosage of the vaccine has been approved, in accordance with the recommendations of the national scientific commission and the specialized world authorities, with the aim of providing citizens with a collective immunity, so that the worst situations in terms of infection rates, hospitalizations in intensive care and deaths are avoided in Morocco.

In the reference hospital sidi Said of Meknes in Morocco, the hospital structure specialized in the management of patients with this disease [3], [4]. This wave has highlighted the incredible efforts of clinicians and all stakeholders within the hospital, in terms of the positive results of the treatment strategies of patients and the allocations of medical resources necessary for them. However, there is still a need for technical means to help predict the severity of cases, and early identification of patients at risk for complications, to ensure good decision making in the intensive care and resuscitation departments in Sidi Said Hospital.

In response to this problem, researchers are working to solidify the process by using machine learning-based predictive analytics techniques that allow clinicians to predict the likelihood of hospitalization of a covid-19 patient based on their comorbidities[5]–[8]. Other researchers have shown how data science and daily data streams from hospital intensive care units, can help learn much faster how to treat patients with COVID-19 based on their daily symptoms and needs[9]–[11]. In a recent study in China,Liang et al presented a risk prediction model to predict the occurrence of severe disease in patients hospitalized for COVID-19 [12].In Liang's model, the response variable had only two levels of severity: severe and non-severe. In reality, the mortalities of COVID-19 patients with different severity levels are variable.

Our objective is to propose an automatic classification model to predict which patients admitted to the hospital with COVID-19 will need special care (oxygen therapy, intensive care, resuscitation) or will die following a sudden deterioration of their health condition. This classification model allows to rationalize patient care (estimation of oxygen supply needs, estimation of resuscitation block saturations...).

In the second section of this paper we will present the predictor variables introduced in the model. The description of several coding techniques (collection, pre-processing, cleaning, transformation) of the dataset (Demography - Comorbidities - Clinical classification of cases - Vaccine) will be presented in the third section. The fourth section explains the proposed ordinal regression techniques. The analysis of the results used in the prediction of multi-class ordinal variables of disease severity (Covid-19), will be presented in the last section.

2. MATERIALS AND METHODS

2.1 Data Understanding

2.1.1 Data Source

We conducted a retrospective observational study, at the sidi-said hospital in Meknes, Morocco. Study participants were consecutive adult patients (≥ 18 years of age) with documented COVID 19 infection (i.e., tested by reverse polymerase chain reaction (RT-PCR) test for SARS-CoV-2), requiring NIV at the time of ICU admission or during ICU stay were prospectively enrolled between April 1, 2021, and December 31, 2021, and followed up until death or hospital discharge.

The dataset in our system contains 254 records and 16 variables. These variables provide demographic, clinical, and therapeutic information about the patient, including the target variable (Outcomes).The medical records of the enrolled patients were accessed by the respective providers and the data were extracted manually, allowing for detailed case ascertainment.

2.1.2 Variable of interest

In this study we used the following characteristics: gender, age, obesity, smoker, alcoholic, clinical classification of cases (moderate, severe, critical), type of screening test (chest X-ray as a screening test, followed by a CT scan in doubtful cases), as well as history of neurological, cardiovascular, respiratory and cancer disorders (see Table 1)

Table 1: Clinical features of patients infected with SARS-CoV-2

Characteristics	Description of features	Feaures attributes
Demographics	Gender (F=1, M=2)	Numerical variables
	Age	Numerical variables
	Obesity	Binary variables
	Smoking	Binary variables
	Alcohol addiction	Binary variables
	Pregnant women	Binary variables
Comorbidites	Diabete No (0) ,Yes (1)	Binary variables
	HTA	Binary variables
	Chronic kidney disease (CKD) > stage III	Binary variables
	Asthma	Binary variables
	Chronic obstructive pulmonary disease (COPD)	Binary variables

	Cardiac disease	Binary variables
	Cancer (active or < 5 years)	Binary variables
Type of screening test	RAT (1) - PCR (2) -CT-Scan (3)	Numerical variables
SpO2 < 92%	No (0) - Yes (1)	Binary variables
Clinical classification of cases	Benign (1) - moderate (2) - strict (3) - critical (4)	Numerical variables
NB of COVID-19 Vaccine Doses	Dose (0 - 1 – 2- 3)	Numerical variables
Outcomes	O2 therapy	Text/categorical
	Intensive care unit (ICU)	Text/categorical
	Resuscitation	Text/categorical
	Death	Text/categorical

2.2 Data Preparation

Data preparation is made up of several stages: Data cleaning, Data Transformation.

2.2.1 Data cleaning

The data collected from a computerized registry of the Sidi-Said Hospital, Meknes, Morocco, are structured in the form of a relational database. This database has undergone a cleaning process to eliminate and reduce noise:

- ✓ Attribute noise is caused by input errors, missing variable values and redundant data.
- ✓ Class noise which is due to errors introduced when assigning instances to classes.

After removing rows with substantial missing values, we checked for missing or null data points in the database using the Python pandas library (Table 2).

Table 2: Dataframe information

#	Column	Non-Null Count	Dtype
0	Gender	236 non-null	int64
1	Age	236 non-null	int64
2	Obesity	236 non-null	int64
3	Smoking	236 non-null	int64
4	Alcohol addiction	236 non-null	int64
5	Pregnantwomen	236 non-null	int64
6	Diabete	236 non-null	int64
7	HTA	236 non-null	int64
8	Chronic kidney disease (CKD)	236 non-null	int64
9	Asthma	236 non-null	int64
10	Chronic obstructive pulmonary disease (COPD)	236 non-null	int64
11	Cardiac disease	236 non-null	int64
12	Cancer	236 non-null	int64
13	RAT - PCR -CT-Scan	236 non-null	int64
14	SpO2 < 92%	236 non-null	int64
15	Clinical classification of cases	236 non-null	int64
16	NB of COVID-19 Vaccine Doses	236 non-null	int64
17	Outcomes	236 non-null	object

The number of records kept is 236 records, each showing a different case of a patient infected with covid-19. Each of these cases is represented by 15 independent predictors/variables, plus an ordinal categorical target variable that reflects the level of risk in covid-19 infected patients.

2.2.2 Data Transformation

2.2.2.1 *Logistic Regression Ordinale model*

Ordinal logistic regression, or proportional odds model, is an extension of the logistic regression model that can be used for ordered target variables. It was first created in the 1980s by Peter McCullagh[13].Ordinal regression problems are machine learning problems in which the goal is to categorize patterns using a categorical scale with labels in a natural order. This labeling structure is prevalent in many real-world applications, which has led to a growth in the number of approaches and algorithms created in this area in recent years [14], [15].In this study, we used this multi-class classification model to predict ordinal variables signifying the level of risk for each patient, in order to determine which patient has the highest risk of mortality from covid-19. The independent variables used in the prediction are divided into four ordinal levels (Table 3):

Table 3: Risk Level of patients infected with Covid-19

Outcomes	Description
Low	O2 therapy: administered when saturation was ≤ 92% at rest in room air; nasal cannula or Venturi mask required; NIV: noninvasive ventilation required.
Medium	Intensive care: these units are specialized in the management of a potentially serious isolated failure (neurological, cardiac...)
High	Resuscitation: these units treat patients with several simultaneous acute failures (circulatory, respiratory, etc.) that threaten their vital prognosis and require the use of heavy techniques
Highest	Deaths

2.2.2.2 *Transforming the Ordinal Classification Problem*

In this study we used the Label-Encoder method, which required the target column to be of the "Category" data type. The default type of a non-numeric target column is "object" (see Table 2). We converted this column to "Category" type using the scikit-learn package (Table 4).

Table 4: Dataframe information and target categorical variable

#	Column	Non-Null Count	Dtype
0	Gender	236 non-null	int64
1	Age	236 non-null	int64
2	Obesity	236 non-null	int64
3	Smoking	236 non-null	int64
4	Alcohol addiction	236 non-null	int64
5	Pregnantwomen	236 non-null	int64
6	Diabete	236 non-null	int64
7	HTA	236 non-null	int64
8	Chronic kidney disease (CKD)	236 non-null	int64
9	Asthma	236 non-null	int64
10	Chronic obstructive pulmonary disease (COPD)	236 non-null	int64
11	Cardiac disease	236 non-null	int64
12	Cancer	236 non-null	int64
13	RAT - PCR -CT-Scan	236 non-null	int64
14	SpO2 < 92%	236 non-null	int64
15	Clinical classification of cases	236 non-null	int64
16	NB of COVID-19 Vaccine Doses	236 non-null	int64
17	Outcomes	236 non-null	Category

Next, we used the label encoder to set a numeric value for each individual class within this categorical target variable, we did not use hot encoding as this would result in the loss of critical (ranking) information. Then, we defined the sensible order (Low< Medium < High <Highest) and mapped to the corresponding variable, to create a new column called (Risk Level) and delete the variable (Outcomes), as explained below (see Table 5 and 6).

Table 5: Convert categorical target variable to numeric variables ordinal

Outcomes	Risk level
Low	0
Medium	1
High	2
Highest	3

Table 6: Dataframe information with the new ordinalnumeric variables "Risk Level"

#	Column	Non-Null Count	Dtype
0	Gender	236 non-null	int64
1	Age	236 non-null	int64
2	Obesity	236 non-null	int64
3	Smoking	236 non-null	int64
4	Alcohol addiction	236 non-null	int64
5	Pregnantwomen	236 non-null	int64
6	Diabete	236 non-null	int64
7	HTA	236 non-null	int64
8	Chronic kidney disease (CKD)	236 non-null	int64
9	Asthma	236 non-null	int64
10	Chronic obstructive pulmonary disease (COPD)	236 non-null	int64
11	Cardiac disease	236 non-null	int64
12	Cancer	236 non-null	int64
13	RAT - PCR -CT-Scan	236 non-null	int64
14	SpO2 < 92%	236 non-null	int64
15	Clinical classification of cases	236 non-null	int64
16	NB of COVID-19 Vaccine Doses	236 non-null	int64
17	Risk level	236 non-null	int64

2.2.3 *Modeling*

2.2.3.1 *Development model*

The data preprocessing step is followed by a modeling process, which involves training the machine learning algorithms to predict classes from the features. In this study, we built our multi-variate logistic regression model, based on demographic data, level of disease severity (moderate, severe and critical), therapeutic and evolutionary, for ordinal prediction of the level of risk for aCOVID-19 patient(need for oxygen inhalation, admission to intensive care, admission to resuscitation, death) in accordance with the good clinical practice protocol. The Python package scikit-learn was used to create classification models using support vector machine (SVM), K-nearest neighbors (kNN), decision tree (DT) and random forest (RF). We used a 10-fold cross-validation test to evaluate the classifiers, which is a predictive model evaluation approach that divides the original set into a training sample for learning the model and a series of tests to measure its effectiveness and efficiency, in order to construct a prediction of the risk level for patients infected with COVID-19 “Fig. 1”.

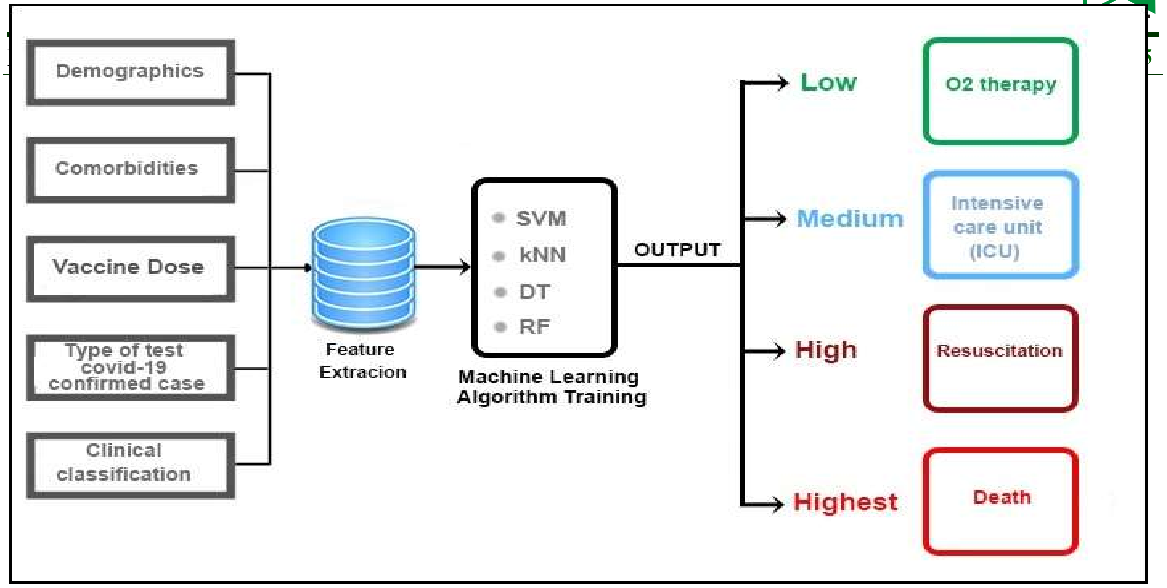


Fig.1 Prediction model used in this study

2.2.3.2 *Classification Methods*

In the present study, four machine learning approaches were used and compared to predict risk levels in covid-19 infected patients: K-Nearest Neighbors, Support-vector machines, decision tree and Random forest. The approaches are listed above, along with their results on the training and validation sets.

a)KNN: K Nearest Neighbors (KNN): is a supervised machine learning algorithm that is one of the most basic, It is a non-parametric classifier based on the location of data points, according to the number of neighbors, similar features are grouped together[16]. The data is not analyzed presumptively; the essential notion behind k-NN is that related samples are clustered in the feature space[17]. As a result, the class label of a test example can be easily identified as the most common class label among those assigned to its k nearest neighbors[18]. The mapping of the dataset onto a metric space can be used to explain this procedure. Euclidean and Manhattan distance metrics are two commonly used distance measures[19]. In this study we used KNN to identify the level of risk in patients infected with Covid-19.

b)Support-vector machines: Support vector machine (SVM) is a machine learning technique based on statistical learning theory, used for classification and regression. SVM provides a better classification that generates a more complex boundary between classes [20]. SVM was chosen as one of the learning techniques to test the performance of the model because it better captured the fundamental properties of the data despite its small size[21] .

c)Décision Tree: The data is split multiple times into decision tree models based on the feature cutoff parameters. As a result of the split, several subsets of the dataset are produced, with each instance belonging to one of them. The ultimate subsets are end nodes or leaves, while the intermediate subsets are inner nodes or splits. To predict the outcome in each leaf node, the average outcome of the training data in that node is used. Decision trees can be used for classification and regression [22].

d)Random forest : Breiman introduced Random Forests (RF) as a tree-based ensemble learning approach for classification and regression in 2001[15], [23]. It has been frequently used in the healthcare field due to its simple structure and superior performance compared to other machine learning approaches[24].

2.2.3.3 *Performance measures:*

Evaluating model performance is an essential part of developing a successful machine learning model[25]. In this study, we used confounding metrics to evaluate the performance of each predictive model, including accuracy, specificity, precision, sensitivity, recall curve, and area under the receiver operating characteristic curve (AUC). Metrics for classification problems are essentially comparing actual classes to classes predicted by the model. They can also be used to understand the probability of those classes that were predicted. The performance of all of these metrics was evaluated to determine the optimal model for predicting risk levels for COVID-19 infected patients.

a) Confusion matrix

A confusion matrix is commonly used to visualize the performance of a classification algorithm. Figure 2 shows the confusion matrix for a multi-class model with N classes [25]. Observations on correct and incorrect classifications are collected in the confusion matrix $\mathrm{C}(_{\mathrm{Cij}})$, where C_{ij} represents the frequency with which class i is identified as class j. In general, the confusion matrix provides four types of classification results with respect to a classification target k :

✓ True positive (TP) : correct prediction of the positive class ($c_{k,k}$)
✓ True negative (TN) : correct prediction of the negative class $\sum_{i,j\in N\setminus\{k\}} c_{ij}$
✓ False positive (FP) : incorrect prediction of the positive class $\sum_{i\in N\setminus\{k\}} c_{ik}$
✓ False negative (FN) : incorrect prediction of the negative class $\sum_{i\in N\setminus\{k\}} c_{ki}$

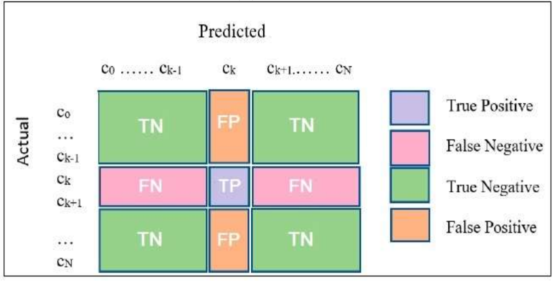


Fig.2 Confusion Matrix for Multi-Class Classification

b) Classification report: A classification report is a tool to evaluate the accuracy of the predictions of a classification algorithm. How many predictions are true and how many are false. As shown below, true positives, false positives, true negatives, and false negatives are used to predict the metrics of a classification report.

Accuracy is the proportion of the total number of correct predictions. It is defined as the total number of positive instances of the model divided by the total number of instances. The accuracy parameter provides the percentage of correctly classified instances. The accuracy of the model is defined as:

$$Overall\ Accuracy = \frac{\sum_{i=1}^{N} c_{i,i}}{\sum_{i=1}^{N}\sum_{j=1}^{N} c_{i,j}} \qquad (1)$$

Precision (2) is the ratio of true positives to all positives. For our problem statement, this parameter is used to determine the degree of the attribute to correctly classify the combination of effective adjuvant treatments, is defined as:

$$Precision_{class} = \frac{TP_{class}}{TP_{class} + FP_{class}} \qquad (2)$$

The true negative rate (Specificity) is defined by equation (3). The false positive rate is the proportion of negative data points that are correctly considered negative, out of all negative data points.

$$\boldsymbol{Specificity}_{\ \boldsymbol{class}} = \frac{\boldsymbol{TN_{class}}}{\boldsymbol{FP_{class} + TN_{class}}} \qquad \mathbf{(3)}$$

The recall (sensitivity) is the true positive rate defined by equation (4). This rate is the proportion of positive data points that are correctly considered as positive, on all positive data points.

$$Recall_{\ class} = \frac{TP_{class}}{TP_{class} + FN_{class}} \qquad (4)$$

Sensitivity and specificity are also called quality parameters and used to define the quality of the predicted class. To determine the quality of the medical diagnostic model, three parameters are basically used; these three parameters are accuracy, sensitivity and specificity.

F1-Score: This harmonic mean metric of accuracy and Recall. Although F1-Score is not as intuitive as Precision, it is useful for measuring the accuracy and robustness of the classifier[26].

$$\boldsymbol{F\,1 - Score} = \frac{2 * TP_{class}}{2 * TP_{class} + FN_{class} + FP_{class}} \qquad (5)$$

The Roc and AUC curve

A receiver operating characteristic (ROC) curve is a curve that plots the rate of true positives (sensitivity) against the rate of false positives (1 - specificity) as the decision threshold changes[22]. The area under the curve (AUC) is a measure of the probability that the model correctly classifies a positive random example versus a negative random example. Its values range from 0 to 1. By analogy, the higher the AUC, the better the model is at distinguishing between covid-19 infected patients at high risk of the disease and those who are not.

The comparison of the performance of the learning algorithms, discussed in the next section, is based on these indicators (Accuracy; Precision; Specificity; Recall; AUC).

3. RESULTS AND DISCUSSION

3.1 Analysis of Result

In this study, the quality of the ordinal classification model is assessed by the classification methods and the confusion matrix. We used the variables: gender, age, Obesity, Smoker, Alcoholic, clinical classification of cases (moderate, severe, critical), type of screening test (Chest X-ray as a screening test, followed by CT scan in doubtful cases), number of covid-19 vaccine, as well as history of neurological, cardiovascular, respiratory, and cancer disorders, to train our machine learning model to identify high-risk individuals with symptoms of COVID-19. This method allows for rapid identification of high-risk patients at four different clinical phases, ranging from the onset of COVID19 to the requirement for expert treatment, such as intubation, intensive care units, and resuscitation.

All the results below are 10-fold cross-validation results, each representing the optimal result of this method, The results of these classification algorithms are presented in Table 6 below, each row of the confusion matrix represents the instances of a real class and each column represents the instances of a predicted class, which it allows to get an overview of the correct predictions and false predictions.

Table 6: The Multi-Class Confusion Matrix Of The Classification Models Used

Classifier	Predicted					n = 236
	0	1	2	3		
SVM	**94**	7	0	1	**0**	**Current**
	14	**47**	3	2	**1**	
	1	5	**38**	2	**2**	
	5	4	2	**11**	**3**	
Random Forest	**87**	8	4	3	**0**	
	10	**45**	8	3	**1**	
	2	3	**40**	1	**2**	
	3	3	3	**13**	**3**	
DecisionTree	**85**	12	0	5	**0**	
	19	**40**	5	2	**1**	
	0	10	**35**	1	**2**	
	5	5	2	**10**	**3**	
KNN	**91**	8	3	0	**0**	
	35	**26**	5	0	**1**	
	22	6	**17**	1	**2**	
	14	2	4	**2**	**3**	

The results showed that the SVM classifier was more accurate in predicting the correct risk levels in covid-19 patients and that the highest false predictive number was 14 for value 1, with a total of 190 correctly classified instances versus 46 incorrectly classified instances. Followed successively by Random Forest, Decision Tree and KNN. On the other hand, we observe that the KNN classifier is the highest in terms of false predictions, with the highest false prediction number (35) for value 1, with a total of 102 misclassified instances.

According to Table6, we observe that the predicted value 1, that it represents the average risk level in covid-19 infected patients, was misclassified by the classifiers (KNN Decision Tree, SVM, Random Forest), followed successively by the predicted values 3,2, 0.

3.2 Performance Evaluation

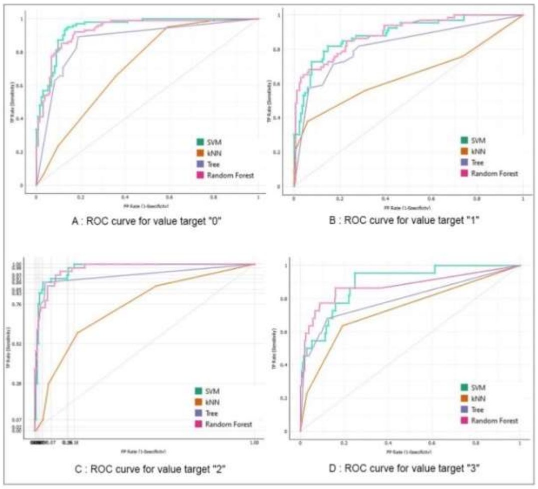
Classification measures were calculated to compare the performances of four algorithms. Table 7 shows that SVM achieves the highest Accuracy (80.5%), Sensitivity (80.5%), Precision (80.1%) and f1 measure (80%), followed successively by Random Forest and Decision Tree with an Accuracy score of 78.4% and 72%, respectively. Considering the kNN classifier, we can notice that at a poor result in terms of Accuracy (57.6%), sensitivity (57.6%), Precision (59, 2%) and the f1 measurement (53.6%), AUC (70.4%). This is why the accuracy of SVM is better than the other classification techniques used in our study, with a score of (80.1%) and a lower error.

Table 7: Evaluation Of The Different Machine Learning Algorithms

Model	AUC	CA	F1	Precision	Recall
SVM	0.930	0.805	0.800	0.801	0.805
Random Forest	0.923	0.784	0.782	0.784	0.784
DecisionTree	0.861	0.720	0.718	0.718	0.720
KNN	0.704	0.576	0.536	0.592	0.576

3.3 Roc and AUC curve

The SVM machine learning classifiers, Random Forest and Decision Tree, give an accuracy level greater than 86% for the classification of risk levels in patients with symptoms of COVID-19. This indicates that the performance of these classification techniques is excellent for prediction. Based on the ROC curves of the models (Figure 3), the SVM model outperformed other machine learning models, in terms of sensitivity and specificity.



A : ROC curve for value target "0"

B : ROC curve for value target "1"

C : ROC curve for value target "2"

D : ROC curve for value target "3"

Fig.3 ROC Curve For The Four Predicted That Signify The Level Of Risk For Patients Infected With The Covid-19 Virus

We observe in Figure 3 that the ROC and AUC curves obtained from the test dataset show the areas under the ROC curves with similar patterns in the upper left corner. This means that the classifiers correctly predict the value 1, followed successively by the values 3, 2, 0 presented in the ROC curves (D, C, B).

4. CONCLUSION

In this paper we have proposed a new multivariate classification model to predict which patients admitted to hospital with COVID-19 will need special care (oxygen therapy, intensive care, resuscitation) or will die following a sudden deterioration of their health. This model will help health professionals (doctors) in decision-making for the recommendation of adequate medical treatments to patients, and to optimize decision-making processes in the management of COVID-19 patients (allocation of medical resources, planning of hospital capacities, estimation of oxygen supply needs, estimation of the saturations of intensive care units, etc.). A comparative study of different multivariate machine learning algorithms (SVM, KNN, DT and RF) showed that the SVM classifier is a reliable, powerful and efficient algorithm in predicting the risk level of patients contaminated by COVID-19.

One key difference between this approach and previous studies is the use of machine learning algorithms to make predictions. Previous studies may have used traditional statistical models or relied on empirical observations. Machine learning algorithms have the ability to analyze complex datasets and identify patterns that may not be immediately apparent using traditional methods. Additionally, this approach focuses specifically on vaccinated patients, whereas previous studies may have looked at unvaccinated patients or the general population. This allows for a more targeted approach to predicting the severity of new variants in those who have already received the vaccine.

We envisage as perspectives of our work, to generalize our study by integrating other parameters such as the dose of vaccination (1st dose, 2nd dose, 3rd dose and 4th dose) and the type of vaccines (AstraZeneca, Sinopharm, Johnson & Johnson and Pfizer) as well as data from patients hospitalized in the various centers in Morocco.

REFERENCES:

[1] "WHO Coronavirus (COVID-19) Dashboard." Accessed: Sep. 09, 2021. [Online]. Available: https://covid19.who.int/

[2] "Coronavirus (COVID-19) Vaccinations." Accessed: Sep. 15, 2021. [Online]. Available: https://ourworldindata.org/covid-vaccinations?country=MAR

[3] T. C. Harvey-Dunstan, A. R. Jenkins, A. Gupta, I. P. Hall, and C. E. Bolton, "Patient-related outcomes in patients referred to a respiratory clinic with persisting symptoms following non-hospitalised COVID-19," *Chron Respir Dis*, vol. 19, p. 147997312110693, Jan. 2022, doi: 10.1177/14799731211069391.

[4] W. Khan, A. A. Khan, J. Khan, N. Khatoon, S. Arshad, and P. D. los Ríos Escalante, "Death caused by covid-19 in top ten countries in Asia affected by covid-19 pandemic with special reference to Pakistan," *Braz. J. Biol.*, vol. 83, p. e248281, 2023, doi: 10.1590/1519-6984.248281.

[5] J. Ebinger *et al.*, "A Machine Learning Algorithm Predicts Duration of hospitalization in COVID-19 patients," *Intelligence-Based Medicine*, vol. 5, p. 100035, 2021, doi: 10.1016/j.ibmed.2021.100035.

[6] W. Cai *et al.*, "CT Quantification and Machine-learning Models for Assessment of Disease Severity and Prognosis of COVID-19 Patients," *Academic Radiology*, vol. 27, no. 12, pp. 1665–1678, Dec. 2020, doi: 10.1016/j.acra.2020.09.004.

[7] M. Al-Emran, M. N. Al-Kabi, and G. Marques, "A Survey of Using Machine Learning Algorithms During the COVID-19 Pandemic," in *Emerging Technologies During the Era of COVID-19 Pandemic*, vol. 348, I. Arpaci, M. Al-Emran, M. A. Al-Sharafi, and G. Marques, Eds. Cham: Springer International Publishing, 2021, pp. 1–8. doi: 10.1007/978-3-030-67716-9_1.

[8] M. Martínez-Lacalzada *et al.*, "Predicting critical illness on initial diagnosis of COVID-19 based on easily obtained clinical variables: development and validation of the PRIORITY model," *Clinical Microbiology and Infection*, vol. 27, no. 12, pp. 1838–1844, Dec. 2021, doi: 10.1016/j.cmi.2021.07.006.

[9] A. Althnian, A. A. Elwafa, N. Aloboud, H. Alrasheed, and H. Kurdi, "Prediction of COVID-19 Individual Susceptibility using Demographic Data: A Case Study on Saudi Arabia," *Procedia Computer Science*, vol. 177, pp. 379–386, 2020, doi: 10.1016/j.procs.2020.10.051.

[10] N. Alballa and I. Al-Turaiki, "Machine learning approaches in COVID-19 diagnosis, mortality, and severity risk prediction: A review," *Informatics in Medicine Unlocked*, vol. 24, p. 100564, 2021, doi: 10.1016/j.imu.2021.100564.

[11] "An ensemble prediction model for COVID-19 mortality risk," p. 28.

[12] W. Liang et al., "Development and Validation of a Clinical Risk Score to Predict the Occurrence of Critical Illness in Hospitalized Patients With COVID-19," JAMA Intern Med, vol. 180, no. 8, p. 1081, Aug. 2020, doi: 10.1001/jamainternmed.2020.2033.

[13] P. McCullagh and P. Lang, "Stochastic Models for Rock Instability in Tunnels," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 46, no. 2, pp. 344–352, 1984, [Online]. Available: http://www.jstor.org/stable/2345520

[14] D. G. Kleinbaum and M. Klein, "Ordinal Logistic Regression," in *Logistic Regression*, New York, NY: Springer New York, 2010, pp. 463–488. doi: 10.1007/978-1-4419-1742-3_13.

[15] M. Ertel, S. Azeddine, A. Said, and E. F. Nour-eddine, "PREDICTION OF THE MOST EFFECTIVE ADJUVANT THERAPEUTIC COMBINATIONS FOR BREAST CANCER PATIENTS USING MULTINOMIAL CLASSIFICATION," *Journal of Theoretical and Applied Information Technology*, vol. 100, no. 23, Dec. 2022, [Online]. Available: http://www.jatit.org/volumes/onehundred23.php

[16] S. D. Bay, "Nearest neighbor classi®cation from multiple feature subsets," *Intelligent Data Analysis*, p. 19, 1999.

[17] S. E. Buttrey, "Nearest-neighbor classification with categorical variables," *Computational Statistics & Data Analysis*, vol. 28, no. 2, pp. 157–169, Aug. 1998, doi: 10.1016/S0167-9473(98)00032-2.

[18] T. Hastie and R. Tibshirani, "Discriminant adaptive nearest neighbor classification," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 18, no. 6, pp. 607–616, Jun. 1996, doi: 10.1109/34.506411.

[19] R. Todeschini, "k-nearest neighbour method: The influence of data transformations and metrics," *Chemometrics and Intelligent Laboratory Systems*, vol. 6, no. 3, pp. 213–220, Sep. 1989, doi: 10.1016/0169-7439(89)80086-3.

[20] "Performance of Support Vector Machine Kernels (SVM-K) on Breast Cancer (BC) Dataset," *ijrte*, vol. 8, no. 2S7, pp. 412–417, Sep. 2019, doi: 10.35940/ijrte.B1076.0782S719.

[21] M. Ertel and S. Amali, "'Classification by Logistic Regression for predicting metastasis in breast cancer patients'. (2021). 1st International Congress on Pure and Applied Sciences ICPAS 21′, June 23 - 25, Meknes, Morocco.," 2021. [Online]. Available: https://orcid.org/0000-0002-3510-9722

[22] A. Syarif, Y. Yun, and M. Gen, "Study on multi-stage logistic chain network: a spanning tree-based genetic algorithm approach," *Industrial Engineering*, p. 16, 2002.

[23] M. Ertel, S. Amali, and N. E. Faddouli, "Multinomial classification to predict the most effective adjuvant combination therapies for breast cancer patients," In Review, preprint, Apr. 2022. doi: 10.21203/rs.3.rs-1574021/v1.

[24] F. Krüger, "Activity, context, and plan recognition with computational causal behavior models," 2018, doi: 10.18453/ROSDOK_ID00002015.

[25] E. Merouane, A. Said, and E. F. Nour-eddine, "Prediction of Metastatic Relapse in Breast Cancer using Machine Learning Classifiers," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 2, p. 6, 2022.

[26] L. Sazonova, G. Osipov, and M. Godovnikov, "Intelligent system for fish stock prediction and allowable catch evaluation," *Environmental Modelling & Software*, vol. 14, no. 5, pp. 391–399, Mar. 1999, doi: 10.1016/S1364-8152(98)00100-5.