ISSN 2278-3091
Volume 9, No.5, September - October 2020
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse15952020.pdf
https://doi.org/10.30534/ijatcse/2020/15952020

Machine Learning Models for the Prediction the Necessity of Resorting to ICU of COVID-19 Patients

Ramy Said Agieb
Department of Communications and Computer Engineering, Faculty of Engineering MTI University,
ramyagieb@gmail.com-drragieb@eng.mti.edu.eg

ABSTRACT

The world is currently facing many unrests and challenges due to the emergence of the COVID-19 Epidemic. The management of medical resources is considered one of the most important challenges posed by the emergence of this epidemic. The intensive care unit (ICU) plays an important role in saving the life of a COVID-19 patient, and therefore work has been done in this research to find models to predict the patient's need to enter ICU or not. The prediction models depend on Machine Learning (ML). Three models will be built to predict the state at which the patient needs to enter the ICU or not. The proposed predictor models based on three types of Supervised machine learning Naive Bayes, K -nearest neighbor (K-NN), and Support Vector Machine (SVM) according to the scarce datasets. Predictor model trained based on Extracted features from patients' X-ray images.

Key words: About four COVID-19-ICU, Machine learning, Supervised Learning, Naive Bayes, K -nearest neighbor, Support Vector Machine.

1. INTRODUCTION

Coronavirus illness (COVID-19) is a destined infection brought about by a newfound coronavirus [1]. A great many people tainted with the COVID-19 infection will encounter mellow to direct respiratory ailment and recoup without requiring exceptional treatment [2]. Older human beings and these with underlying medical problems like cardiovascular disease, diabetes, persistent respiratory disease, and most cancers are greater probable to enhance serious illness [3]. Currently, there's no remedy for the COVID-19 virus. Common remedies that have verified to be extremely positive in govt the signs and symptoms encompass taking over-the-counter medication, consuming a lot of water, getting unobjectionable rest, lamister overexertion, now not smoking, staying yonder from smoky areas, and the use of wipe mist vaporizers or humidifiers. Taking medicines like acetaminophen, ibuprofen, and naproxen can assist to decrease the ache and fever-related with the illness. The WHO and the CDC are no longer recommending any precise fitness measures at this time for dealing with the COVID-19 outbreak. They say human beings who have respiratory signs and symptoms they assume are associated with the coronavirus have to are searching for clinical help and advice. They say make the healthcare facility enlightened of your issues and do no longer return to work, school, or different crew things to do till you are unrepeatable the unprepossessing and flu-like signs and symptoms are no longer COVID-19 related [4].

The most dangerous problem when the number of patients increased which leads to Medical resources are collapsing because of the high demand for them. The most important part of the medical system is ICU, and management the necessity of resorting to ICU [4].

Deal with emergency situations requires more understanding of Data. Data gives us more facts more than it looks, so we use the new computerized techniques to extract the facts we search about it and discover what can't see [5]. The most important three techniques which deal with data to predict results are artificial intelligence (AI), machine learning (ML), deep learning (DL) [6]. In fact, ML is a subset of AI, and DL is a subset of ML [7]. The three techniques are trained using a training data set to create a model. When data is introduced to the proposed model, it makes a prediction for the results which are evaluated for accuracy and if the accuracy is acceptable, the model is deployed. If the accuracy is not appropriate, the model is trained again and again with an augmented training data set to reach the desired accuracy [8].

AI is a way that permits machines to mimic human behavior. ML is a technique that uses statistical methods to allow machines to enhance with experience. DL is a method that makes the calculation of a multi-layer neural system attainable. ML satisfies good result with small data sets but when used DL it needs large data sets, so in the current research, use ML according to scarcely of data sets, it makes a prediction on the basis of the model [9].

This paper is organized a follow. Section 2 describes the ML and its types. Section 3 describes the process of building the

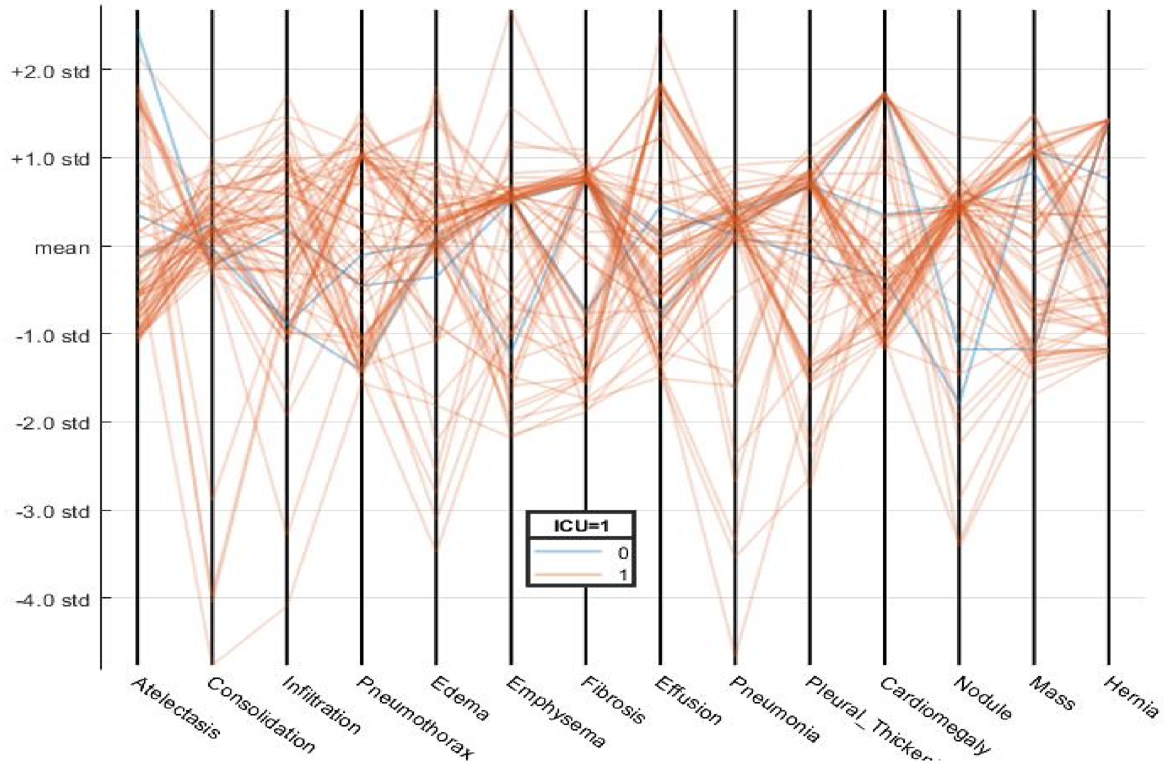

Figure 1: Parallel Representation of Dataset

three ML models used to predict the patient's enter ICU or not and the processes used to validate and testing the models. Section 4 presents the conclusion of the research.

2. MACHINE LEARNING

ML applied in many and wide tasks. It affects actually every enterprise from its safety malware seeks, to clinical applications, to climate forecasting, to stockbrokers seeking out reasonably-priced trades. The ML is created by supervised learning, Unsupervised Learning, and Reinforcement Learning [10]. Supervised learning is a method by which used inputs and their corresponding outputs to train the model. Unsupervised Learning is utilizing data that is neither arranged nor named and permitting the calculation to follow up on that data without direction [11]. Reinforcement learning is the model to make a sequence of decisions. The agent figures out how to accomplish an objective in a dubious, conceivably complex condition [12-13].

ML divided in many steps collect data, pre-process data, develop predictive model, validation and test predictive models [14-15].

3. BUILD THE THREE MODELS

3.1. COLLECT DATA

In this stage, I care about collect data from authorized and trusted sources. The main target of the research is to decide if the patient of COVID-19 pandemic needs to enter the ICU or not. So, I collect the data from the source which got it from the world health organization (who) [16].

3.2. PRE-PROCESS DATA

In this stage, I prepare the data to be suitable for working in the prediction models. The data divided into two parts, the first is an image folder which contains x-ray of different patients, and the second is a table contains a whole lot of inconsistencies like missing values, clean columns, abrupt values, and wrong statistics format which need to be cleaned. Determine the rows that have total information especially the confirmation about the patient enter ICU or not. Learning tools for chest X-ray diagnostics are used to extract the information from the patient's x-ray images. From the previous two steps, I have a table contains independent variables or features or Predictors {Atelectasis, Consolidation, Infiltration, Pneumothorax, Edema, Emphysema, Fibrosis, Effusion, Pneumonia, Pleural Thickening, Cardiomegaly, Nodule, Mass, and Hernia} and label or dependent variable [17].

3.3. DEVELOPPING MODELS

In this stage, three models are initialized based on training data. The three models will be used in the prediction process to gain an answer and defined the new undefined data. According to the dataset characteristics which features have many peaks, as shown in figure 1, i use three types of machine learning Naive Bayes, K-NN, and SVM.

3.3.1. NAIVE BAYES

Naive Bayes is a classification machine learning based on Bayes' Theorem used for binary (two-class) and multiclass classification problems. The key point in Naive Bayes is finding the contributions of each predictor of dataset in the decision. As in this research, the patient enters the ICU unit when he arrived at a certain level of Fatigue from the disease as reflected in the X-ray image. Predictors based on X-ray images are treated as a single player in determining the decision. Regardless of, whether these features rely upon one another or the presence of different highlights between each other. As in the current state, the prediction model treats the role of Atelectasis in the produced final result as the role and contribution of Consolidation in the process, and equal treatment for the other features. In other words, each feature is given the same importance or weight. The procedure as follow:

- Separate by Class as in this state, dataset separated int two categories enter ICU (1) and didn't enter ICU (0).
- From a given dataset calculate the mean and the standard deviation for each predictor depending on the class.
- Estimate the probabilities of each predictor given class. From figure 1, the predictors have a continuous distribution and multiple peaks or modes, so use Kernel Distribution to estimate the probabilities for the values given each class.

3.3.2. K-NN

K-NN is a supervised machine learning technique which describes as one of the simplest ways to predict belonging a particular element to a specific group. It can deal with both twofold or binary and multiclass information and makes no presumptions about the parametric type of the chosen limit. K-NN technique supposes the similarity between the new case/data and available groups and puts the new case into the category that is most similar to the available categories. K-NN doesn't have a preparation stage and is best portrayed through the basic procedure used to group new items. By using training data, which are classified into groups based on its Special characteristics. If the points are visualized, we will be able to remark some groups. Now, given an unknown point, we can assign it to its group based on its nearest neighbor belong to (the shortest distance). This implies a point near a group treated as a member of this group. In the case of 1-NN means find the nearest Euclidean distance between the unknown point and the other points which belong to defined groups. In other cases when, the value of K equal to two or more, the distribution of unknown data to the group which contains the highest number on neighbor nodes. K-NN by this method caused the high percentage error so; i used a modified version of it called weighted K-NN.

In weighted K-NN, the closest k cases are given a weight utilizing a function called the kernel function. The goal is to give more weight to the cases which are close by and less weight to the focuses which are farther away. Select any function compatible with training data and control the weighted.

In this research, the K value is optimized at five. The distance metric is Euclidean with squared inverse as distance weighted.

3.3.3.SVM

SVM used when label has two values only or two states. SVM makes its classifications based on finding the best hyperplane. The hyperplane separates between the two label's groups or states. The efforts are to find the best hyperplane which leading or meaning largest margin between two groups. Edge implies the maximal width of the section corresponding to the hyperplane that has no inside information focuses.

SVM kernel functions are the key to produce the best hyperplane that used to classify data and predict the state of un unknown data. The three kernel functions are Linear, Gaussian, and Radial Basis functions. The kernel must appropriate with the dataset. From analyzing the dataset, i decide to use gaussian as the kernel function. The gaussian kernel gives the ability to deal with non-liner predictors with linear SVM and use linear functions in the prediction model.

3.4. VALIDATION AND TEST PREDICTIVE MODELS

The next stage after built the predictive model is starting with validation the model. One of the most errors that happen by the designer of the model is the resubstitution error. This error accrues when used all the dataset to train the model and test by newcomer data or cases. To prevent this error, the data are split into two groups of data; one is used to train the data and the other to test the predictive model. On the other hand, improving the performance of previous models is done through applying Cross-Validation. Cross-Validation depends on dividing the data into many subsets or segments and trains the predictive model as follow:

- Reserve some segments of the test informational index.
- Using the rest informational index train the model.
- Test the model utilizing the reversed segment.
- The performance of the model is assessed.
- The average test error overall folds are calculated over all segments.

The predictive models were built and tested by using the Matlab tool. The performance in machine learning is measured by using the confusion matrix or the error matrix and receiver operating characteristic (ROC) curve. The three models are trained and tested used 5- Cross Validation. The three models have accuracy equal 100%.

Accuracy measures by applied the test data on the predictive models and compared the results with the real values of applied data on the models. Cross-validation helped us to switch between the partitions of data. In the confusion matrix, the true values of the label were represented in rows and the predicted values of the label were represented in the column. The main diagonal of the matrix display and indicate to the accuracy, the three models achieved maximum accuracy as shown in figure 2 and figure 3. The results can assessment by finding true positive rate (TPR) also known as sensitivity that represents the amount of agreement between the true and predicted positive samples. False prediction values (FPV) and false discovery rate (FDR) do the same role as TRP for false results. As the percentage of TRP increased means the better model, our models give 100% as shown in figure 2 and figure 3. Figure 4 shown ROC curve, the values in the curve between zero and one. For true positive rate as near to one and false positive rate near to zero meaning more accurate model, the three models satisfy maximum accuracy 100%. Previous work as in [18] used the Random Forest classifier that satisfied accuracy equal to 77.7 %. Another experiment is made by study the effect of features extracted from an x-ray on the prediction process; the result showed that the most important controller feature in the prediction process is the Infiltration feature. In feature more research must be directed to study the methods that control the infiltration feature.

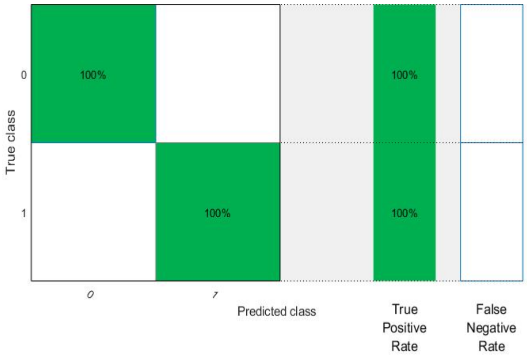

Figure 2: Confusion Matrix “TPR and FNR”

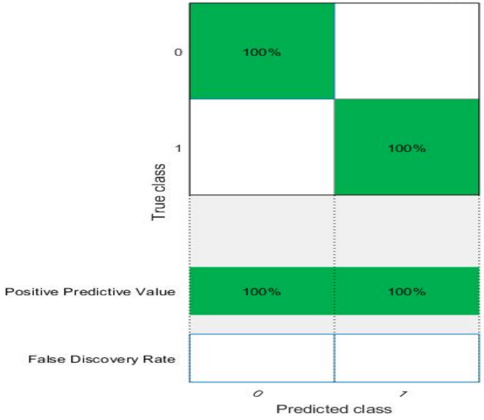

Figure 3: Confusion Matrix” FPV and FDR”

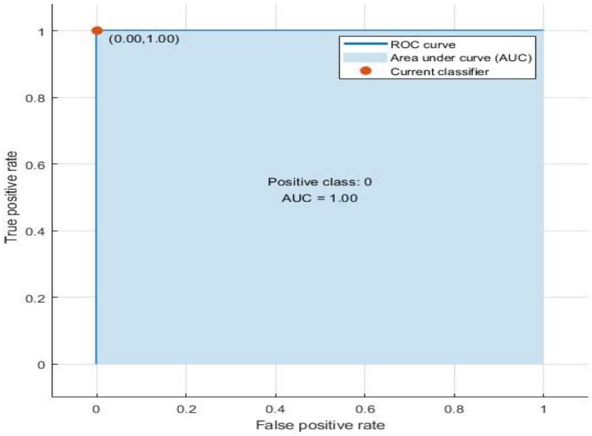

Figure 4: ROC Curve

The comparison between the different types of ML is as shown in table 1; the accuracy, training time, and prediction speed are the fields of comparison. The tree models achieved the highest accuracy compared to other types of ML, but the model which based on SVM (Fine Gaussian) is the optimum because it achieved the lower training time equal 12.184 sec and the maximum prediction speed equal 3400 obs/sec as shown in table 1.

Table 1: Comparison between different types of machine learning techniques applied to predict the decision to enter ICU or not enter ICU for COVId-19 patient.

Machine Learning techniques		**accuracy %**	**training time sec.**	**prediction speed obs/sec**
Tree	Fine	97.4	10.414	500
	Medium	97.4	9.322	550
	Coarse	97.4	8.950	520
Linear Discriminant		90.8	11.57	340
Logistic Regression		97.4	16.607	730
Naive Bayes	Gaussian	89.5	9.545	1300
	Kernel	100	12.2	340
SVM	Linear	92.1	11.706	710
	Quadratic	93.4	11.528	460
	Cubic	97.4	11.354	1900
	Fine Gaussian	100	11.056	3400
	Medium Gaussian	92.1	10.74	3800
	Coarse Gausian	92.1	10.482	610
K-NN	Fine	98.7	12.324	400
	Meduim	92.1	12.12	680
	Coarse	92.1	12.573	410
	Cosine	92.1	11.948	1500
	Cubic	92.1	12.321	910
	Weighted	100	12.184	1200
Ensemble	Boosted Tree	92.1	17.167	950
	Bagged Tree	93.4	17.940	200
	Subspace Discrimnant	92.1	17.737	1300
	Subspace K-NN	98.7	19.218	170
	RUSBoosted Tree	69.7	18.149	260

4. CONCLUSION

A conclusion This research worked and focused on using supervised machine learning to produce three models to predict the decision about entering the COVID-19 patient to ICU or not. Three models were built and tested based on the dataset and Matlab tool. The features extracted from the x-ray images of COVID-19 patients which used to train and tested the models. The three models made by applying Naive Bayes, K-NN, and SVM techniques. The three models achieved an accuracy of 100 % by comparing the results from the model and the real value by the cross-validation method. Confusion matrix used to detect other accuracy metrics like TPR, and sensitivity which proved that the optimal results from the models. ROC curve was used to measure the area under the curve, also given optimum results that equal one. The infiltration feature is the controller player and, In the future, the focus should be on finding solutions to control the factors that control this feature. The biggest and most important problem we faced and disturbed us is the scarcity of data. I hope in the future to find more data to give us the ability to apply more models in the production process.

REFERENCES

1. World Health Organization. **Coronavirus Disease (COVID2019) Situation Reports, World Health Organization, Geneva, Switzerland**, https://www.who.int/emergencies/diseases/noveloronavirus-situation-eports *,2020*
2. Z. Wu and J. M. McGoogan. **Characteristics of and important lessons from the coronavirus disease 2019 (COVID19) outbreak in China**, *JAMA*, vol. 323, 2020.
3. W. Guan, Z. Ni, Y. Hu et al. **Clinical characteristics of 2019 novel coronavirus infection in China**, *New England Journal of Medicine, vol. 395, pp. 1708–1720, 2020.*
4. N. Chen, M. Zhou, X. Dong et al. **Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan**, China: a descriptive study, *4e Lancet, vol. 395, no. 10223, pp. 507–513, 2020.*
5. Z Pratap Dangeti. ***Statistics for Machine Learning***, JAMA, Packt Publishing,2017.
6. W. Guan, Z. Ni, Y. Hu et al. **Modern Applications of Machine Learning,** New England Journal of Medicine**,** vol. 395, pp. 1708–1720, 2020.
7. Yen-Wei Chen, Lakhmi C. Jain. **Deep Learning in Healthcare**, 1st ed. Springer Nature Switzerland AG: Springer International Publishing,2020.
8. Simon Rogers, Mark Girolami. **A First Course in Machine Learning**, 2nd ed. New York: CRC press, 2017.
9. Aurélien Géron**. Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow**, 1st ed. California: O'Reilly Media,2019.
10. Murphy, K.P. **Machine Learning: A Probabilistic Perspective**, 1st ed. Cambridge, The MIT Press, 2012.
11. Abdulhamit Subasi**. Practical Guide for Biomedical Signals Analysis Using Machine Learning Techniques**, Elsevier,2019.
12. Christopher M. Bishop. **Pattern Recognition and Machine Learning**, 1st ed. Springer Nature Switzerland AG: Springer International Publishing,2006.
13. Petra Perner, Maria Petrou. **Machine Learning and Data Mining in Pattern Recognition**, 1st ed. Springer Nature Switzerland AG: Springer International Publishing,1998.
14. Sang Boem Lim. **Classification and Big Data Usages for Industrial Applications**, *International Journal of Advanced Trends in Computer Science and Engineering (IJATCSE)*, Volume 8, No.4, July- August 2019.
15. Sujeet More, Jimmy Singla. **Machine Learning Techniques with IoT in Agriculture**, *International Journal of Advanced Trends in Computer Science and Engineering (IJATCSE Volume 8, No.3, May - June 2019*.
16. Joseph Paul Cohen, Paul Bertin, Vincent Frappier. **Chester: A Web Delivered Locally Computed Chest X-ray Disease Prediction System**, MIDL, 2020.
17. X-ray system open source," https://mlmed.org/tools/xray".
18. Fu-Yuan Cheng, Himanshu Joshi. **Using Machine Learning to Predict ICU Transfer in Hospitalized COVID-19 Patients**, Journal of Clinical Medicine, June 2020.