E-ISSN 2581 – 7957
P-ISSN 2277 – 3916

CVR Journal of Science and Technology, Volume 22, June 2022
DOI: 10.32377/cvrjst2204

Supervised Learning for COVID Mortality Span Prediction

A. Srinivasa Reddy
Assoc. Professor, CVR College of Engineering/CSIT Department, Hyderabad, India
Email: srinivas.asr@gmail.com

Abstract: **The covid-19 outbreak is causing concern among the world's population. Because there are no preventative precautions in place, current treatment approaches are limited to treating patients who tested positive for covid-19. In this case, determining the severity of the patient's disease is critical in lowering the covid-19-related death rate. The pathology findings are used by clinical professionals to scale the severity of the condition. Clinical specialists' diagnosis is strongly dependent on their level of competence. Unlike the other dimensions, sensitivity, or accuracy in disease-prone situations, is extremely important in clinical practice. This paper describes a supervised learning strategy for performing computer-assisted covid-19 mortality scope using the target patient's pathology records. The experimental investigation shows the utility of the proposed strategy in predicting death with the least amount of false alerting.**

Index Terms: **COVID-19, Computed Tomography, Feature Optimization, World Health Organization, Machine Learning.**

I. Introduction

There have been several pandemic diseases that have afflicted humans in the past. The WHO (World Health Organization) is collaborating with numerous national authorities as well as professionals to combat these pandemics. As soon as the first instance of COVID-19 sickness was identified in the Wuhan district of China in December 2019, it spread throughout the world until it was ultimately eradicated on January 30th, 2020.COVID-19 has been classified as an infectious disease caused by a new coronavirus and found in WUHAN, China, according to the research in [1]. SARS-CoV-2 (severe acute respiratory syndrome coronavirus-2) is a new form of virus that has never been seen before in humans. Furthermore, the virus is spread primarily through respiratory difficulties, droplets from sneezing, coughing, or when people meet one another. When these droplets are inhaled or land on surfaces that other people touch with their hands, they become infected with the virus. If their hand meets their nose, eyes, or mouth, they become infected with the virus.

The coronavirus, which has become an epidemic disease, may exist for a few days on a variety of surfaces such as plastic and stainless steel, and for a few hours on copper and cardboard. Nonetheless, the number of potential viruses may decrease over time and may not always be there to induce infection. Furthermore, in the case of people, the viral symptoms might appear anywhere from 1 to 14 days after infection. Later, it spreads quickly, leaving no time to prepare for a newly discovered renowned and contagious virus, prompting the WHO to declare COVID-19 a pandemic, as mentioned in [2], because to its rapid transmission among humans.

Various trials have been conducted in the medical labs in distinct phases to assess the efficacy of covid-19, but no findings have been released too far. Because it is a new virus, there is no vaccination available. Even though multiple pharmaceutical and research companies have begun to work on a vaccine, it may be months or even years before the vaccine is available to the public. Because there aren't enough ventilators, hospital beds, kits, or oxygen tanks, and there's no proper treatment or vaccination available, it's important to investigate the number of positive cases, the number of recovery cases, and other factors that could influence virus proliferation. Everyone should be aware of the virus and take the appropriate steps to avoid infection.

II. Related Work

For gaining insights and spreading the sources of disease, effective outbreak estimation methodologies must be used. As mentioned in [1,3] various legislative and political organizations rely on insights from estimating methods for adopting fresh policies and monitoring the efficiency of implemented rules. The current worldwide pandemic disease COVID-19 has been envisioned as intricate and non-linear, according to the work [4].

Furthermore, the outbreak differs from other recent epidemics, raising the question of whether established methodologies can produce reliable results, as noted in [3].In addition, as noted in [1,5], various unknown as well as recognized variables included in the spread, the complexities of large population behavior in several geopolitical areas, and variances in containment systems all exacerbated technique uncertainty. As a result, traditional epidemiological methodologies face new hurdles in producing more reliable results. As indicated in [3-5], multiple unique ways have arisen to overcome this difficulty. These methods add diverse assumptions into modeling.

Machine learning has been used to improve the screening and diagnosis of identified patients using a radio imaging scheme that is like clinical data from blood samples and CT scans (computed tomography). Radiology imaging such as CT scans and X-rays are routinely used by healthcare experts to supplement traditional screening and diagnosis. Inappropriately, the performance of such devices is moderate during the peak of the SARS-CoV-2 pandemic. In this regard, the work [6] demonstrates feasible ML devices by providing a unique technique that includes both valid and quick SARS-C detection.

As mentioned in [7], the current investigations propose an accessory device for improving COVID-19 diagnosis accuracy with a novel Automatic COVID-19 identification model based on a deep learning algorithm. The proposed

Received on 04-08-21 Revised on 04-11-2021, Accepted on 04-02-2022.

CVR Journal of Science and Technology, Volume 22, June 2022
DOI: 10.32377/cvrjst2204

method makes use of raw chest x-ray pictures from 127 affected patients. The binary class has a performance accuracy of 98.08 percent, and the multiclass has a performance accuracy of 87.02 percent. The forecast approach used a judgment rule to forecast fast and estimate the number of infected people at high risk; patients who have been declared infected should be considered for intensive care, and the shortness rate should be reduced. By employing a deep-learning algorithm over a large short-term network, a forecasting approach based on Canadian time-series has been built.

III. Supervised Learning

Supervised learning is a sort of machine learning in which machines were trained using well-labeled training data and then predict the output based on that data. The labelled data indicates that some of the input data has already been tagged with the appropriate output.

In supervised learning, the training data presented to the machines acts as a supervisor, instructing the machines on how to correctly predict the output. It uses the same notion as when a student learns under the guidance of a teacher. The process of supplying input data as well as proper output data to the machine learning model is known as supervised learning. A supervised learning algorithm's goal is to discover a mapping function that will map the input variable(x) to the output variable(y).

Supervised learning can be utilized in the real world for things like risk assessment, image categorization, fraud detection, spam filtering, and so forth.

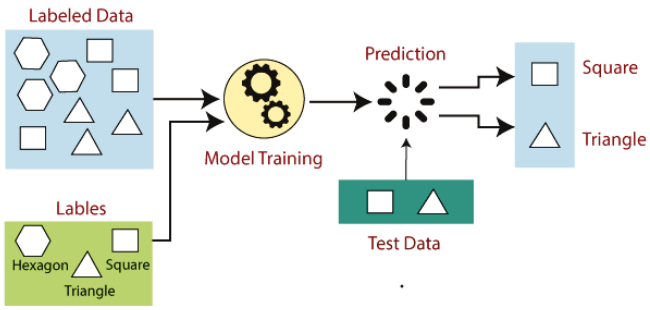

Figure 1. Model of Supervised Learning

The Process of Supervised Learning:

1. Determine the type of training dataset you'll be using.
2. Obtain the tagged training data by collecting/gathering it.
3. Divide the training dataset into three parts: training, test, and validation.
4. Determine the training dataset's input characteristics, which should contain enough information for the model to accurately predict the output.
5. Choose an appropriate algorithm for the model, such as a support vector machine or a decision tree.
6. Use the training dataset to run the algorithm. Validation sets, which are a subset of training datasets, are sometimes required as control parameters.
7. By giving the test set, you may assess the model's correctness. If the model correctly predicts the outcome, then our model is accurate.

There are two types of algorithms from supervised learning:

1. Classification
2. Regression

Regression:

If there is a relationship between the input and output variables, regression procedures are applied. It's used to predict continuous variables like weather forecasting, market trends, and so on. These algorithms include Regression trees, Linear regression, Bayesian linear regression and Polynomial regression.

Classification:

When the output variable is categorical, meaning there are two classes, such as Yes-No, Male-Female, True-False, and so on, classification methods are utilized. These algorithms include Logistic regression, Random Forest, Support Vector Machine and Decision trees.

IV. Proposed Method

This section delves into the specifics of the data, methodologies, and materials used in the suggested statistical evaluation method for predicting the mortality scope of covid-19-positive patients.

A. The Data

Age, gender, weight, and clinical reports connected to pathological reports related to blood tests are among the demographic parameters. The information gathered from clinical reports is a subset of the pathology reports for diabetes types I and II. Each record has 31 qualities, one of which is the mortality scope, which can be either positive or negative.

B. The Features

The training datasettD is bipartite as two sets i.e. tD_+ , $tD_-$with entries labeled positive and negative in the order they appear in the training dataset. Allow the set to include all attributes (excluding mortality scope) that indicate the anticipated values in each record of both sets. Additionally, all potential subsets of size 1 to $|aL|$ of the setaL size are discovered. These subsets are referred to as n-gram feature labels, and the projected values are referred to as n-gram feature labels.

The frequency of each n-gram feature in datasets tD_+ , tD_- indicates their level of confidence in the positive and negative labels, in that order. The discovery of n-grams, on the other hand, is complicated and proportional to the count of feature attributes. As a result, lowering the attribute count optimizes the number of n-grams needed in the learning phase.

CVR Journal of Science and Technology, Volume 22, June 2022
DOI: 10.32377/cvrjst2204

C. Feature Optimization

The values projected to an attribute with significant confidence in the labels positive and negative are produced as ideal attributes. The t-test [24], a distribution diversity assessment measure, was used to determine the degree of diversity between the values of an attribute in records classified as positive and the values of the same attribute in records labeled as positive and the values of respective attribute in records labelled as negative. The mathematical model of feature optimization explored in following description. The algorithm is explained as follows.

Step 1: for each attribute a_i of the attribute list aL,

$$\forall_{\substack{i=1 \\ |aL|}} \{a_i \exists a_i \in aL\} \tag{1}$$

Let the vectorsaf_i^+, af_i^- denote the values of the attribute a_i in positive and negative label setstD_+ , tD_-

Step 2: mean of the values listed in the vectoraf_i^+

$$\langle af_i^+ \rangle = \frac{1}{|af_i^+|}\left(\sum_{j=1}^{|af_i^+|}\{e_j \exists e_j \in af_i^+\}\right) \tag{2}$$

Step 3: mean of the values listed in the vectoraf_i^-

$$\langle af_i^- \rangle = \frac{1}{|af_i^-|}\left(\sum_{j=1}^{|af_i^-|}\{e_j \exists e_j \in af_i^-\}\right) \tag{3}$$

Step 4: deviation of the values listed in the vectoraf_i^+

$$\sigma(af_i^+) = \frac{1}{|af_i^+|}\left(\sum_{j=1}^{|af_i^+|}\left\{\sqrt{\left(\langle af_i^+ \rangle - e_i\right)^2}\right\}\right) \tag{4}$$

Step 5: deviation of the values listed in the vectoraf_i^-

$$\sigma(af_i^-) = \frac{1}{|af_i^-|}\left(\sum_{j=1}^{|af_i^-|}\left\{\sqrt{\left(\langle af_i^- \rangle - e_i\right)^2}\right\}\right) \tag{5}$$

Step 6: The t-score is calculated by ordering the vectors af_i^+, af_i^-representing attributea_i values toward positive and negative labels.

$$ts = \frac{(\langle af_i^+ \rangle - \langle af_i^- \rangle)}{\sqrt{\sigma(af_i^+) + \sigma(af_i^-)}} \tag{6}$$

D. Mortality Span Prediction

The coefficients used to scale the mortality scope will be discovered using the mathematical model below. The mean and root-mean-square-deviation (RMSD) of the positive and negative confidence of the diversified n-gram feature values are the coefficients. The required steps for this process are:

Step 1: The mean of predictive confidence of the n-gram feature values of the positive label.

$$\langle pc_+ \rangle = \frac{1}{|pc_+|}\left(\sum_{i=1}^{|pc_+|}\{c_i \exists c_i \in pc_+\}\right) \tag{7}$$

Step 2: The mean of predictive confidence of the n-gram feature values of the negative label.

$$\langle pc_- \rangle = \frac{1}{|pc_-|}\left(\sum_{i=1}^{|pc_-|}\{c_i \exists c_i \in pc_-\}\right) \tag{8}$$

Step 3: The deviation$RMSD$ of the predictive confidence of the n-grams of the positive label.

$$pc_+^\sigma = \frac{1}{|pc_+|}\left(\sum_{j=1}^{|pc_+|}\left\{\sqrt{\left(\langle pc_+ \rangle - c_j\right)^2} \exists c_j \in pc_+\right\}\right) \tag{9}$$

Step 4: The deviation of the predictive confidence of the n-grams of the negative label.

$$pc_-^\sigma = \frac{1}{|pc_-|}\left(\sum_{j=1}^{|pc_-|}\left\{\sqrt{\left(\langle pc_- \rangle - c_j\right)^2} \exists c_j \in pc_-\right\}\right) \tag{10}$$

Step 5: The lower-bound of the positive prediction scale is the difference between mean and respective deviation of the positive confidence of n-gram feature values.

$$lb_+ = \langle pc_+ \rangle - pc_+^\sigma \tag{11}$$

Step 6: The cumulative of mean and deviation denotes the upper-bound of the prediction scale of the positive label.

$$ub_+ = \langle pc_+ \rangle + pc_+^\sigma \tag{12}$$

CVR Journal of Science and Technology, Volume 22, June 2022
DOI: 10.32377/cvrjst2204

Step 7: The lower-bound of the negative prediction scale is the difference between mean and respective deviation of the positive confidence of n-gram feature values.

$$lb_{-} = \langle pc_{-} \rangle - pc_{-}^{\sigma} \tag{13}$$

Step 8: The cumulative of mean and deviation denotes the upper-bound of the prediction scale of the negative label.

$$ub_{-} = \langle pc_{-} \rangle + pc_{-}^{\sigma} \tag{14}$$

Step 9: Coefficient of n-grams has been returned, corresponding upper and lower-bounds in the form of Heuristics regression.

$$\text{Return } (ss, ssl, ssu) \tag{15}$$

E. Prediction of Label

The input record r was given to predict if the mortality span is positive or negative. Collect all the n-gram feature values of all optimal n-gram features from the given record r. Find the mean of the positive and negative confidence $< c_{+} >, < c_{-} >$of all n-gram feature values of the given input record towards training corpus.

Further, predict the mortality scope of the record as follows,

1. $if\,(\langle c_{+} \rangle \geq ub_{+})$

   //If the mean confidence$\langle c_{+} \rangle$of the n-gram feature values discovered from the given input record r is greater than or equals to positive upper-boundub_{+}, the mortality scope against the given record$r$ is positive

2. $if\,(\langle c_{+} \rangle \geq \langle pc_{+} \rangle \&\& \langle c_{-} \rangle < \langle pc_{-} \rangle)$

   // The condition mean confidence$\langle c_{+} \rangle$is greater than the mean of the positive confidence$\langle pc_{+} \rangle$and mean confidence$\langle c_{-} \rangle$is lesser than the negative confidence$\langle pc_{-} \rangle$ also recommends labelling the given record as positive (prone to mortality).

3. $if \langle c_{+} \rangle \geq lb_{+} \&\& \langle c_{-} \rangle < lb_{-}$

   //The condition that denotes mean$\langle c_{+} \rangle$of the confidence discovered for positive label of given record r is greater than lower-bound$lb_{+}$of the positive label prediction scale and mean$\langle c_{-} \rangle$of the negative confidence of the n-gram feature values of given record r is lesser than the lower-bound$lb_{-}$of the negative label prediction scale

Under the conditions contradicting to above stated conditions, the given input record shall be labeled as negative (not prone to mortality scope).

V. Results & Discussion

Using demographic and pathological parameters of covid19 positive patients, an experimental investigation was done on a projected model of this contribution and other corresponding methods. Furthermore, the accuracy, specificity, f-measure, precision, and Mathew's correlation coefficient (MCC) have all been used to evaluate the performance of the projected model for this contribution.

The dataset used in the experiment comprises 1000 entries divided evenly between two labels: positive and negative. Each entry in the collection contains anonymized patient demographic and pathological information. There are 31 features in total, including the label. Among the 30 attributes, the optimal feature selection algorithm identified 17 as ideal.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{16}$$

$$Specificity = Selectivity = \frac{TN}{TN + FP} \tag{17}$$

$$Sensitivity = Recall = \frac{TP}{TP + FN} \tag{18}$$

$$Precision = \frac{TP}{TP + FP} \tag{19}$$

$$F - measure = \frac{2 * TP}{2 * TP + FP + FN} \tag{20}$$

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{21}$$

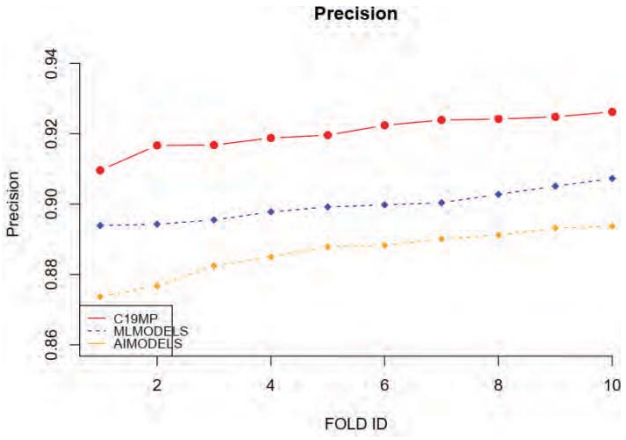

Figure 2. Precision obtained from tenfold cross validation

E-ISSN 2581 – 7957
P-ISSN 2277 – 3916

CVR Journal of Science and Technology, Volume 22, June 2022
DOI: 10.32377/cvrjst2204

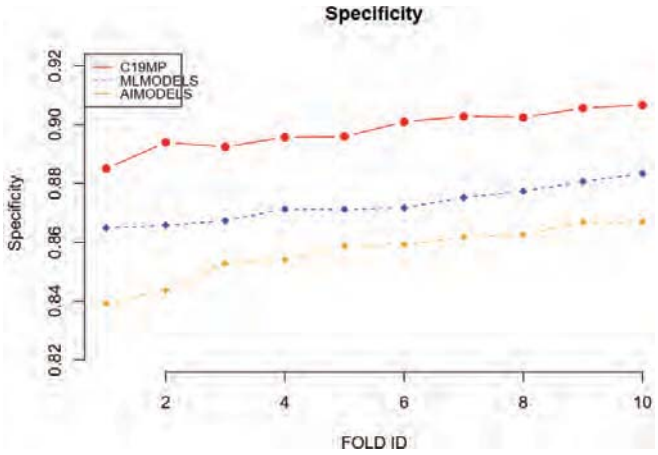

Figure 3. Specificity obtained from tenfold cross validation

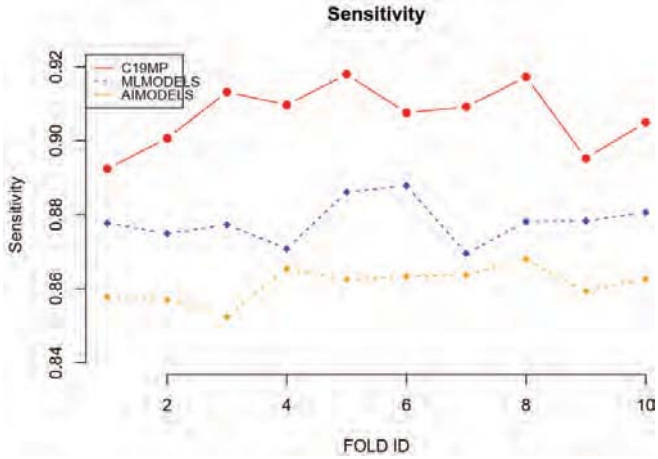

Figure 4. Sensitivity obtained from tenfold cross validation

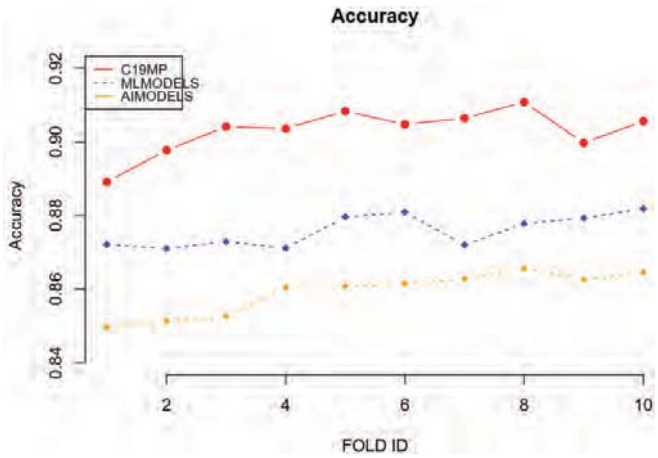

Figure 5. Accuracy obtained from tenfold cross validation

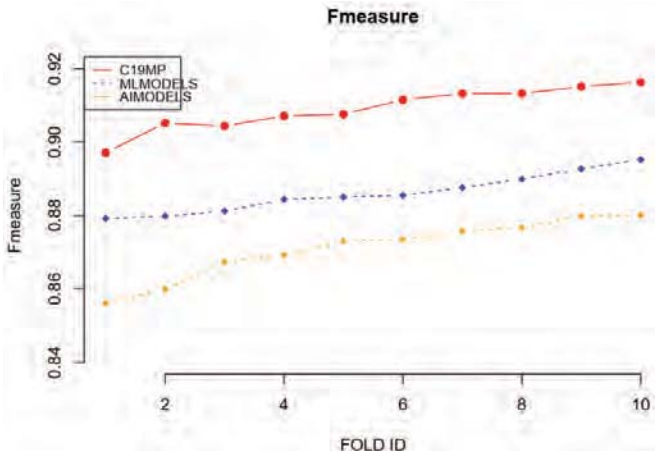

Figure 6. F-measure obtained from tenfold cross validation

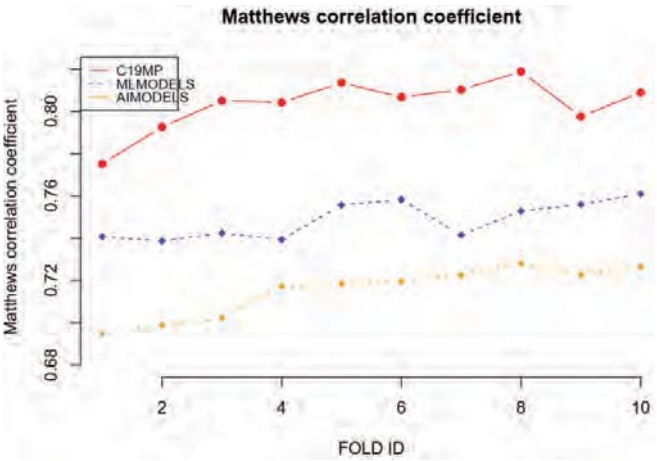

Figure 7. MCC obtained from tenfold cross validation

Figure 2 shows a graph depicting the comparison of Precision across 10 folds of cross validation observed from the suggested and contemporary models. The Precision for the ML-MODELS, AI-MODELS, and suggested C19MP, respectively, is 0.89961±0.004257, 0.88624±0.006422, and 0.9203±0.004796.

Figure 3 shows a graph depicting the comparison of Specificity over ten folds cross validation of C19MP, AI-MODELS, and ML-MODELS. In that order, the average specificity of the projected C19MP model, corresponding ML-MODELS, and AI-MODELS models is 0.89812±0.006397, 0.87287±0.005889 and 0.85659±0.008803.

In figure 4, the Sensitivity of the contribution C19MP, which indicates real positive rate, was compared to the sensitivity observed from counterpart models ML-MODELS and AI-MODELS. In that order, the average sensitivity of the ML-MODELS, AI-MODELS, and C19MP is 0.87811±0.005502,0.86118±0.004322, and 0.90675±0.008177.

Figure 5 shows a graph depicting the comparison of Accuracy observed from tenfold cross validation of C19MP, AI-MODELS, and ML-MODELS. The average accuracy of the contribution C19MP model, as well as its ML-MODELS and AI-MODELS counterparts are 0.903±0.005864, 0.87585±0.004175 and 0.85919±0.005471.

In figure 6, the F-measure obtained from tenfold cross validation of the proposed and contemporaneous models. ML-MODELS have an average F-measure of 0.88605±0.005087, while AI-MODELS have an average F-measure of 0.87115±0.007662. The contribution C19MP has an average F-measure of 0.90907±0.00561.

In figure 7, the Mathews Correlation Coefficient (MCC) observed from the contribution C19MP, counterpart models ML-MODELS and AI-MODELS were compared. The average MCC of the ML-MODELS and AI-MODELS obtained from tenfold cross validation is 0.74865±0.008393 and 0.71515±0.011289, respectively. This contribution's average MCC is 0.80326±0.011731.

VI. Conclusions

A novel statistical strategy for predicting the mortality scope of a patient who has tested positive for covid-19 has been proposed. Unlike current models, which solely examine demographic features, the proposed model reflects both demographic and pathology report features when performing supervised learning. A novel feature optimization metric based on the distribution diversity method t-test has been presented to reduce the process complexity of the learning phase. The tenfold cross validation performed on the recommended dataset demonstrated the relevance of C19MP's contribution. By comparing modern methodologies with the past, the suggested model C19MP has been scaled. Future research can incorporate cross-media features, which are a combination of demographic, picture, and signal formats, to increase the accuracy of the mortality scope prediction of COVID-19 positive patients.

References

[1] Sohrabi C, Alsafi Z, O'Neill N, Khan M, Kerwan A, Al-Jabir A, Losifidis C, Agha R. World health organization declares global emergency: a review of the 2019 novel coronavirus (COVID-19). Int J Surg 2020. doi:10.1016/j.ijsu.2020.02.034.

[2] Zhu N, Zhang D, Wang W, Li X, Yang B, Song J, et al. A novel coronavirus from patients with pneumonia in China, 2019. N Engl J Med 2020.

[3] Tan W., Zhao X., Ma X., Wang W., Niu P., Xu W., et al. A novel coronavirus genome identified in a cluster of pneumonia cases–Wuhan, China 2019– 2020China CDC Weekly2020; 2(4):61-62.

[4] Ivanov, D. Predicting the impacts of epidemic outbreaks on global supply chains: A simulation-based analysis on the coronavirus outbreak (COVID-19/SARS-CoV-2) case. Transp. Res. Part E Logist. Transp. Rev. 2020, 136, doi:10.1016/j.tre.2020.101922.

[5] Li, Yun, et al. "Individual-Level Fatality Prediction of COVID-19 Patients Using AI Methods." Frontiers in Public Health 8 (2020): 566.

[6] Ardakani AA, Kanafi AR, Acharya UR, Khadem N, Mohammadi A. Application of deep learning technique to manage COVID-19 in routine clinical practice using CT images: results of 10 convolutional neural networks. ComputBiol Med 2020;121:103795. 2020 https://doi.org/10.1016/j.compbiomed.2020.103795.

[7] Ozturk T, Talo M, Yildirim EA, BalogluUB, Yildirim O, Rajendra Acharya U. Automated detection of COVID-19 cases using deep neural networks with Xray images. ComputBiol Med 2020:103792. doi:10.1016/j.compbiomed.2020. 103792.