Prognosis Patients with COVID-19 using Deep Learning

Jose Luis Guadiana Alvare

Monterrey Institute of Technology and Higher Education

Ruben Morales-Menendez

Monterrey Institute of Technology and Higher Education

Fida Hussain (✉ fidahussain@ujs.edu.cn)

Monterrey Institute of Technology and Higher Education

Etna Rojas Flores

Monterrey Institute of Technology and Higher Education

Arturo García Zendejas

Monterrey Institute of Technology and Higher Education

Ricardo A Ramirez Mendoza

Monterrey Institute of Technology and Higher Education

Research Article

Keywords: deep learning, random forest, coronavirus, COVID-19, prognosis, mortality risk prediction

Posted Date: February 15th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-192397/v1>

License: This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

RESEARCH

Prognosis patients with COVID-19 using deep learning

Jose Luis z Guadiana Alvare, Ruben Morales-Menendez, Fida Hussain*, Etna Rojas Flores, Arturo García Zendejas and Ricardo A Ramirez Mendoza

* Correspondence:
fidahussain@ujs.edu.cn
School of Engineering and
Sciences, Tecnológico de
Monterrey, Monterrey, Mexico
Full list of author information is
available at the end of the article

Abstract

Background: Prognostics study the prediction of an event before it happens, to enable critical decision making to be more efficient. The prognostics are very useful for front line physicians to predict how a disease may affect a patient and react accordingly to save the patients' lives. The coronavirus (COVID-19) is novel and not enough knowledge about the virus' behaviour and Key Performance Indicators (KPIs) to assess the mortality risk prediction. However, using a lot of complex and expensive medical biomarkers could be impossible for many low budget hospitals. This motivates the development of a prediction model that not only maximizes performance but does so using the least number of biomarkers possible.

Methods: For the mortality risk prediction, this research work proposes a COVID-19 mortality risk calculator based on a Deep Learning (DL) model, and based on a data set provided by the HM Hospitals from Madrid, Spain. A pre-processing strategy for unbalanced classes and feature selection is proposed.

Results: The DL model is tested, and the following results are achieved include area under the curve (AUC) 0.93, F2 score 0.93, recall 1.00, accuracy, 0.95, precision 0.91, specificity 0.9279 and maximum probability of correct decision (MPCD) 0.93.

Conclusion: The MPCD score shows that the proposed DL outperforms on the every data set when evaluating even with an over-sampling technique. The benefits of imputating unavailable biomarker data are also evaluated. The results are compared against a random forest (RF) algorithm and the newly proposed methods. The results show that the proposed method is significantly best for the risk prediction of the patients with COVID-19.

Keywords: deep learning; random forest; coronavirus; COVID-19; prognosis; mortality risk prediction

Background

Coronavirus is a large family of viruses that can cause common colds and even acute severe respiratory syndrome (SARS) and the Middle East respiratory syndrome (MERS). It has been proven that the root cause is a new type of virus called 2019 Novel Coronavirus (2019-nCoV). The World Health Organization (WHO) and the Centers for Disease Control and Prevention (CDC) are closely monitoring the development of the virus and how to prevent and treat diseases caused by the 2019-nCoV [1].

Recently many algorithms have been developed to diagnose the COVID-19 outbreak. [2, 3]. A predictive model of CoViD-19 disease progression is proposed by

[4] using Multivariate Analysis (Cox proportional regression). Pourhomayoun et al. [5] proposed an ML algorithm to accurately predict the mortality risk of CoViD-19 patients. In [6], a Gaussian Process Regression (GPR) model with optimized hyper-parameters is used to predict the mortality rate in five different countries (Turkey, Spain, Sweden, France and Pakistan) and the data was obtained from [7]. Extreme Gradient Boosting (XGBoost) Classifier was proposed to model the probability of requiring mechanical ventilation within the next 24 hours, using data from the first 2 hours after admission [8]. In [9], attempted to predict the occurrence of Major Adverse Cardiac Events (MACE) in Acute Myocardial Infarction (AMI) patients, during the 1, 6 and 12 months follow up periods after hospital admission using ANN. A mortality risk calculator was developed based on XGBoost model was developed using patients from hospitals in Spain (HM Hospitals) and Italy (ASST Cremona) [10]. Deep neural network transfer model based on a Convolutional Neural Network (CNN) was proposed to diagnose a patient with CoViD-19 by analyzing their lungs' X-ray images [11]. In [12], proposed a CNN transfer learning model to diagnose CoViD- 19 patients using X-ray and CT-Scan images by using the Gradient Weighted Class Activation Mapping (GRAD-CAM) technique [13, 14]. Yadaw et al. [15] developed a mortality prediction model using an XGBoost algorithm. In [16], introduced a multivariate regression model based on clinical characteristics at presentation to predict ICU admission and mortality in CoViD-19 patients.

Early identification of the patients with CoViD-19 at risk to quickly deteriorate, and those with very low risk for critical disease, are the measures required. Moreover, not every hospital has the resources, budget, time, staff, equipment, etc., to conduct many complicated tests before needing to decide. A mortality risk calculator for CoViD-19 must not only be as accurate as possible but also use only the minimum number of features to produce an acceptable prediction. The proposed methodology will help to easy prognose the patient's survival rate.

Materials and Methods

Datasets and Pre-process

The datasets of the patients with COVID-19 have been collected from HM Hospitals Madrid, Spain. It contains the anonymized records of 2,307 patients with CoViD-19. The database was divided into six different sections, each section contained a different type of data for each patient with an ID of the patients. This raw database poses five main technical challenges: 1. Incomplete records, 2. different units, 3. Combination of categorical and numerical features. 4. irrelevant and redundant, 5. unbalanced classes. The raw and filtered datasets are shown in Table 1 and Table 2, respectively.

An RF algorithm was used to select the features with the highest predictive power to decrease the feature space, by analyzing the importance assigned to each feature by the algorithm. For this purpose, the SHAP value [17] was used to estimate the impact/weight of each input variable in prediction. The SHAP value graph is a graphical visualization of how much a feature contributes to a model's prediction. A large positive SHAP value indicates the feature is very relevant to detect positive outputs, while a large negative value is associated with negative output. The colour bar shows the feature value associated to the given SHAP value, while the thickness

of a feature's line, indicates the number of samples present in the data set for the given feature value, and the SHAP values are shown in Figure 1.

Table 1: Complete raw database description.

Section	Description
1	Demographic data: Patient ID, Age, Gender, Diagnosis
2	(Positive/Negative/Pending), Admission/Discharge date and motive, SpO2, Temperature, Heart rate, Blood Pressure, etc.
3	Prescribed medication: daily dose and duration
4	Evolution of vital signs: SpO2, Heart rate, Temperature, Blood pressure and Blood glucose values
5	Laboratory tests with date, results, and units
6	Comorbidities, these are coded based in [World Health Organization]

Table 2: Complete raw database description.

S. No	Acronym	Feature Name	S. No	Acronym	Feature Name
1		Gender	25	IGMm	IgM (Immunoglobulin M) min
2		Age	26	IGGm	IgG (Immunoglobulin G) min
3	SpO2	Oxygen saturation	27	TNIm	Troponin min
4		Kidney Failure	28	APm	Prothrombin activity min
5		Hypertension	29	TPm	Prothrombin time min
6		Diabetes	30	LIN%M	Lymphocytes % Max
7		Heart Disease	31	LINM	Lymphocytes Max
8		Respiratory distress	32	LEUCM	Leukocytes Max
9		Discharge motive	33	NEUM	Neutrophils Max
10		Patient ID	34	NEU%M	Neutrophils % Max
11	LIN%m	Lymphocyte % min	35	PLAQM	Platelet count Max
12	LINm	Lymphocytes min	36	PCRM	C-reactive protein Max
13	LEUCm	Leukocytes min	37	DDM	Dimer D Max
14	NEUm	Neutrophils min	38	CrM	Creatinine Max
15	NEU%m	Neutrophils % min	39	LDHM	LDH Max
16	PLAQm	Platelet count min	40	IL6M	Interleukin 6 Max
17	PCRM	C-reactive protein min	41	LEULCRM	Leukocytes count Max
18	DDm	D Dimer min	42	LACM	Lactate Max
19	CrM	Creatinine min	43	FERM	Ferritin Max
20	LDHm	LDH min	44	IGMM	IgM (Immunoglobulin M) Max
21	IL6m	Interleukin 6 min	45	IGGM	IgG (Immunoglobulin G) Max
22	LEULCRM	Leukocytes count min	46	TNIM	Troponin Max
23	LACm	Lactate min	47	APM	Prothrombin activity Max
24	FERm	Ferritin min	48	TPM	Prothrombin time Max

Figure 1: SHAP values of the selected features accordingly.

In this research work, biomarkers were selected including: Prothrombin activity, Creatinine, Dimer-D, Ferritin, Immunoglobulin G, Immunoglobulin M, Interleukin 6, Lactate, LDH, Leukocytes (Count), Lymphocytes (Count), Neutrophils (Count), C-reactive protein, platelets, prothrombin time, and troponin. Because sampling frequency for lab tests is inconsistent, simple time series statistical representation, such as maximum and minimum values, were chosen to represent these features. Therefore, two more features were added for every biomarker, one for the maximum and another one for the minimum values. Features 1, and 4 through 9 are categorical data [True; False], while every other feature is numerical data. Data distribution: in Figure 2, can be seen that the patient's age distribution has a mean of about 70 years old following a normal distribution. Oxygen saturation values have a mean of 92.28, with a couple of lower outliers, which suggest a more severe disease state according to literature. A clear unbalance between the classes is observed, with only

16.5% of deceased patients. As for comorbidities, there are 919 patients with none of the selected comorbidities, 398 patients with only 1 comorbidity, 148 patients with 2 of them, 35 patients with 3, 3 patients with 4 comorbidities, and no patients with every comorbidity. The most common comorbidity among patients is hypertension. Since the original data set contains a lot of missing cells, the sample size reduces as the number of features increases. Table 3 and Table 4 shows that the filtered database distribution and biomarker, respectively.

Figure 2: (a) Left side distribution of age and right side oxygen saturation (SPO2) and (b) left side comorbidities and right side multiple comorbidities.

Table 3: Filtered database distribution according to key features.

Feature	Detail	Number of patients (%)
Gender	Male	927 (61.7%)
	Kidney failure	81 (5.4%)
	Hypertension	446 (29.7%)
Comorbidities	Diabetes	194 (12.9%)
	Heart disease	69 (4.6%)
	Respiratory distress	21 (1.4%)
Discharge motive	Deceased	248 (16.5%)

Table 4: Biomarkers data shows average values for maximum and minimum features.

Name of Biomarker	Mean value min - max
LEUC ($\times 103/\mu\text{L}$)	6.03 - 9.85
LIN ($\times 103/\mu\text{L}$)	1.00 - 1.65
LIN%	13.90 - 25.03
NEU ($\times 103/\mu\text{L}$)	4.10 - 7.84
NEU%	63.23 - 78.66
PLAQ ($\times 103/\mu\text{L}$)	206.42 - 320.81
Cr (mg/dL)	0.84 - 1.09
PCR (mg/L)	38.84 - 130.69
LDH (U/L)	482.91 - 726.57
DD (ng/mL)	1194.71 - 4509.60
IL6 (pg/mL)	192.74 - 239.50
LAC (mmol/L)	1.73 - 2.29
FER (ng/mL)	1150.13 - 1526.91
TNI (ng/L)	27.01 - 36.47
AP (%)	70.58 - 80.93
TP (s)	13.70 - 16.55

Pre-processing dataset: To increase the training data availability, a pre-processing algorithm [18] was applied, it is a Greedy-like algorithm that at each iteration maximizes the number of samples by selecting the column (feature) with more rows (samples) available. Since the original data set contains a lot of missing cells, the sample size reduces as the number of features increases. The eight sub-data sets pose a trade-off between the number of features and the number of samples, since the sub-set with the most discriminatory information cannot be determined in advanced, the learning algorithm is applied to all of them. Since normalizing the data generally speeds up learning and leads to faster convergence [19] the remaining numeric features are re-scaled using the Min-Max normalization method [20]. By examining the dataset, and it created into eight sub-data with different features and illustrated in Table 5.

Table 5: The number of features in every defined sub data set.

Data subsets	Num. of features	Num. of samples
1	10	1503
2	22	1449
3	24	1434
4	26	1428
5	28	1419
6	30	1341
7	34	1291
8	48	683

Methodologies

SMOTE technique: SMOTE technique was used to balance the dataset. It generates synthetic examples in a less application-specific manner, by operating in “feature space” rather than “data space”. The minority class is over-sampled by taking each minority class sample and introducing synthetic examples along the line segments joining any/all of the k minority class nearest neighbours [21].

Decision threshold: the output of the trained model is the probability of a given sample belonging to one class or another. Intuitively, we assigned an equally spaced decision threshold for every class, e.g., 0:5 for a binary classification problem. In this research work, the Optimal Classifying Threshold Method (OCTM) algorithm was used to obtain the decision threshold value which optimizes the MPCD score [22]. The algorithm is shown in Figure 3.

Data imputation: The impact of using imputed

Figure 3: Pseudo-code of the OCTM algorithm.

data on the prediction model is evaluated for the cases where there are time or budget constraints and obtaining complex biomarker data is impossible or unfeasible [23]. To properly evaluate the proposed imputation method, impute available biomarker features were used to be able to calculate an error between imputed and real values. The dataset was split into train and test sets using a 10-Fold approach. Biomarker features were imputed on the test set using the mean value of the K most similar patients from the real biomarker data of the train set using the KNN algorithm [24]. The value of K is determined by the amount of available data. The benefit of using the imputed features will be evaluated by comparing the model’s performance against the same test set with the real biomarker data. The error of the estimated imputed data is calculated using the Root Squared Error. Finally, the benefit of adding imputed biomarkers data were measured by comparing the performance of the imputed test set against the performance of a model which only uses basic patient information, without any imputation. As we impute more features, the model’s performance will have more uncertainty and therefore a higher error. This motivates us to impute only the necessary number of features to see an improvement on the model without adding variance to the output.

Deep learning (DL): The basic idea behind DL is to stack many shallow multi-layer algorithms to obtain more abstract representation of features as the network gets deeper [25]. DL has recently gained popularity particularly in the field of computer vision but is rapidly moving towards different areas, such as diagnosis and prognosis

in the medical field [26]. The quintessential DL algorithm is the ANN. ANNs are a type of ML algorithm roughly based on the biological neurons of the brain and the way that they are interconnected with one another to learn complex abstract representations. In this research work, the DL model with binary cross-entropy as the loss function and the Adam algorithm as an optimizer to adjust the network's weights have been used. The model has 3 hidden layers with 17, 10 and 5 neurons, respectively. The mini-batch optimization technique was employed. For the binary classification problem, the sigmoid function has been used as the activation function. This model was developed using the Keras framework running on TensorFlow in python.

Random forest algorithm (RF): The RF algorithm uses an ensemble of decision trees to make a prediction [27, 28]. A decision tree fits a function (typically piece-wise constant) over domain X by recursive partitioning in a greedy way. RF Regressor was used to predict the mortality risk of the patients. The following hyper-parameters have been used to train and test the RF model which are the number of estimators = 500, maximum depth = 2 and maximum feature = 5. This model was developed using the scikit-learn library in python.

Results and Discussion

The proposed model is trained many times, where at each iteration a different fold of the data set is used as the testing set while every other fold is used for training. The final unbiased result is the average value of each evaluation metric across every fold. The proportion of the label classes distribution should be kept across every fold. This is done to avoid a fold of the model training with only positive or negative class samples, therefore getting a biased classifier. For this purpose, the Stratified K-Fold sci-kit-learn function, which keeps the proportion of the label feature across every fold. For the validation, 90% of the samples for training and 10% for testing have been used. Figure 4 shows a boxplot graph of the MPCD score for the proposed DL model with and without the SMOTE approach on every sub dataset. The best results are shown in Figure 5 of the deep learning method.

Figure 4: Results summary of the DL model for every data set (Left - Normal, Right - SMOTE).

Figure 5: Boxplot for 10-fold CV on data set 4 for the normal (left) and SMOTE (right) approaches of the DL model.

Data imputation is the process of replacing missing data with substituted values. The mean performance value of the model is the same, which suggests that the imputation process did not introduce false information to the process. Figure 6 and Figure 7 show the boxplot graph of root squared error (RSE) values for the first of the 10 folds when imputating the NEU%M and LIN%M biomarkers, and DDm and DDM, respectively. The red dots highlighted in the graph represent the patients

who had a different classification between real and amputated data. The results for the DD%_m and DD%_M imputation process are shown in Figure 8 and Figure 9. It can be seen a similar variance when comparing the amputated and real value models, suggesting a good imputation performance.

Figure 6: RSE values for imputed NEU%_m and LIN%_M features.

Figure 7: RSE values for amputated DD_m and DDM features.

Figure 8: Metrics comparison of amputating the NEU%_m and LIN%_M biomarkers (Left) against real values (Right).

Figure 9: Metrics comparison of imputation the DD%_m and DD%_M biomarkers (Left) against real values (Right).

Performance comparison For comparison purposes, the RF algorithm has been trained and tested based on the same dataset. The comparison of the results from the DL and RF models on every data set with and without the SMOTE approaches are presented in Figure 10 and Figure 11, respectively. The results show that the DL model has high prediction accuracy.

However, an increase in the variability of the results is also observed, because more samples are dropped as more features are added, as shown in Table 2. Comparing both approaches on every data set shows that the SMOTE approach yields no significant improvement for the model. The proposed DL model can make an accurate prediction even on the unbalanced data set. Further, analysing the proposed DL outcomes distributions, we can see that it is very close to the actual output distribution of the dataset. Figure 12 shows the original database class distribution on the left side, while the right side shows the DL predicted distribution. This indicates that the proposed model successfully models the data set underlying distribution.

Figure 10: Comparison of results for the DL (left) and RF (right) models.

Figure 11: Comparison of results for the DL (left) and RF (right) models using SMOTE.

Figure 12: The original database class distribution on the left side, while the right side shows the DL predicted distribution.

In Figure 5, the comparative result shows that the SMOTE approach has a Recall distribution closer to 1, but more variability for the Precision metric. This is observed on the outlier of the MPCD value near to 0:95 which could suggest a possible improvement on the model's performance when introducing more training data. The same can be said about the final data sets, where the variability of the MPCD score is bigger, but outliers with very high MPCD values are also observed. Finally, the threshold value set by the SMOTE approach gets closer to 0:50 because the proportion between classes are closer to one another. In Table 6, the proposed DL model outperforms comparatively the support vector machine (SVM), artificial neural network (ANN), XGBoost, logistic regression (LR) and RF model in practically every data set on both approaches. The RF model greatly benefits from the SMOTE approach, while the DL model appears to work better even when having unbalanced classes.

Table 6: The proposed model (DL) is compared with other models.

Feature	Key performance indicators (KPI)					
	MPCD	Recall	F2 score	Precision	AUC	Accuracy
DL	0.93	1	0.93	0.91	0.93	0.95
RF	0.88	0.95	0.89	0.9	0.89	0.93
[29]	0.77	0.91	0.87	0.85	0.87	0.89
ANN [30]	0.78	0.89	0.9	0.87	0.88	0.90
XGBoost [31]	0.81	0.93	0.94	0.90	0.90	0.91
LG [32]	0.76	0.90	0.87	0.88	0.86	0.89

The proposed model is compared to the recently published methods. The comparative results are shown in Table 6. In [2], the authors used 53 confirmed CoViD-19 patients from the Wenzhou Central Hospital and Cangnan People's Hospital in Wenzhou, China. The median age was 43 years and 62.2% were men. Common symptoms included fever (in 47 patients, 88.7%) and cough (in 32, 60.4%). Median white blood cell count ($\times 10^9/L$) was reported to be 4.8, and median Lymphocyte count ($\times 10^9/L$) was of 1.2. Out of the 53 patients 9.4% developed ARDS, 1.9% was taken into the Intensive Care Unit (ICU) and 47.5% required supplemental oxygen. They compared various ML algorithms using a 10- fold cross-validation accuracy. The top accuracy achieved was 80% from a Support Vector Machine (SVM) and a K-Nearest Neighbor (KNN) ($k=5$) algorithm. Pourhomayoun et al. [5] proposed an ML algorithm to accurately predict the mortality risk of CoViD-19 patients. 17,000 laboratory-confirmed CoViD-19 patients' dataset from 76 countries with an average of age of 56.6, from which 74.4% recovered. Data imputation techniques were used for missing values and a balanced dataset was created for training and testing the model. 112 features were available from symptoms and doctor's medical notes, and patient's demographic and physiological data. After applying different filter and wrapper methods, the feature space was reduced to 42 features. The best performance accuracy achieved was 93.75% by the ANN algorithm. Hyper-parameters

were tuned using grid search and the final architecture had two hidden layers with 10 neurons in the first layer and 3 neurons in the second layer. A sigmoid function was used as the hidden layer activation function and stochastic gradient as the optimizer with a constant learning rate and a regularization rate of $\alpha = 0.01$.

In [6], used GPR and the results were compared against a polynomial regression model using the Root Mean Square Error as a performance metric. The data was obtained from using data from March 21, 2020, to May 10, 2020, which represent 51 days. 36 days were used for training and 15 days for testing.

In [8], taken 197 patients with confirmed CoViD-19 and was obtained from five US health systems, being 51.3% male and the majority being between 30 and 80 years old (73%). For each patient, 12 features were used as input to the model. The XGBoost Classifier was shown outperformed MEWS both in every metric, specifically sensitivity (0.90) and specificity (0.58), meaning that it can detect 16% more patients at risk while simultaneously reducing false-positive alerts. In [9], the

Table 7: Comparison between ML models to predict CoViD-19 patients' mortality risk.

References Model	Input features	Performance evaluation		
		KPI	DL	RF
The proposed DL and RF models	48 input features	MPCD	0.93	0.88
		Recall	1.00	0.95
		F2 score	0.93	0.89
		Precision	0.91	0.90
		AUC	0.93	0.89
		Accuracy	0.95	0.93
		Threshold	0.38	0.20
SVM and KNN [2]	11 input features	Acc = 0.80		
ANN [5]	42 input features	Acc = 0.9375		
GPR [6]	1 feature	RMSE = 0.1306		
XGBoost Classifier [8]	12 features	Recall = 0.90		
		TPR = 0.58		
		Acc = 0.9598		
ANN [9]	51 features	Recall = 0.8125		
		TPR = 0.961		
		AUC = 0.97		
XGBoost Classifier [10]	20 features	AUC = 0.938		
		Recall = 0.9762		
CNN and Deep Transfer Learning [11]	224x224 RGB X-ray images	TPR = 0.7857		
		Acc = 0.8810		
		AUC - ROC = 0.88		
CNN and Deep Transfer Learning [12]	X-ray and CT-Scan images	Precision = 0.95		
		Recall = 0.94		
		F1 - Score = 0.95		
XGBoost Classifier [15]	3 features	Acc = 0.95		
Multivariate Regression model [16]	7 features	AUC - ROC = 0.91		
Multivariate Analysis (Cox proportional regression) [4]	4 features	AUC - ROC = 0.82		
		AUC = 0.91		

dataset consists of 10,813 patients from 52 Korean hospitals, using 51 variables for prediction. The ANN algorithm was used and got the following results Accuracy = 95.98%, Sensitivity = 81.25%, Specificity = 96.10% and AUC = 0.97.

The study comprised 2,831 patients, 711 (25.1%) of whom died during hospitalization while the remaining ones were discharged. Vitals signs and laboratory test values from only the first day upon admission were used. Two models were trained for the mortality risk calculator: one with lab test results and one without. The missing values were imputed using KNN, features missing in more than 40% were excluded and 95% confidence intervals were calculated using bootstrapping. Performance evaluation using 40 random data partitions into training and test sets was

made yielding an average AUC of 93.8% using laboratory values and 90.5% without laboratory test values [10].

The data set consisted of 284 X-ray images from which around 142 were positive for CoViD-19. A 70%-30% training-testing split was made for proper model evaluation. The nCOVnet uses the top layers of the VGG16 image classifier as a base model and then adds 5 layers as part of the transfer learning methodology. The proposed model achieved Sensitivity of 97.62%, Specificity of 78.57%, Accuracy of 88.10% and AUC-ROC of 0.88 [11]. In [12], proposed a CNN transfer learning model to diagnose CoViD-19 patients using X-ray and CT-Scan images, and the following results obtained precision = 0.95, recall = 0.94, F1 score = 0.95 and accuracy = 95%. Yadaw *et al.* [15] developed a mortality prediction model using a database (3,841 patients, 8.2% deceased) with features age, SpO2 and type of patient. The selected metrics were AUC of ROC (0.91) scores. The best results were obtained with the XGBoost algorithm. In [16], 641 hospitalized patients database had used with a median age of 60 years old, 40.1% female, 62% no critical illness, 30% were admitted to the ICU, and 82 who expired. Five significant variables predicting ICU admission were lactate dehydrogenase, procalcitonin, SpO2, smoking history, and LIN. Seven significant variables predicting mortality were heart failure, procalcitonin, lactate dehydrogenase, chronic obstructive pulmonary disease, SpO2, heart rate, and age. The mortality group uniquely contained cardiopulmonary variables. The risk score model (a multivariable regression model) yielded good accuracy with an AUC-ROC of 0.74 for predicting ICU admission and 0.82 for mortality.

Ji *et al.* [4] were used 208 patients dataset and the average age was 44, 117 of 208 patients (56.2%) male, 31 (14.9%) older than 60 years, 45 (21.6%) 40 (19.2%) patients, the clinical conditions deteriorated progressed during the observation period. Analyses showed comorbidity, older age, lower lymphocyte count, and higher lactate dehydrogenase (CALL) were independent high-risk factors for CoViD-19 progression. A nomogram achieved good concordance (0.86) and a good AUC-ROC (0.91). Using CALL score model, clinicians can improve the therapeutic effect and reduce the mortality with more accurate and efficient use of medical resources.

The comparative analysis shows that the proposed DL method gives a significantly higher result. The proposed DL model can make an accurate prediction even on the unbalanced data set. The proposed procedure can be applied to research areas such as manufacture.

Conclusion

A mortality risk calculator for CoViD-19 patients based on DL model is proposed. Its high reproduction rate demands hospitals to predict patient's evolution upon admission to efficiently manage hospitals resources. Therefore, a mortality risk calculator must not only accurately classify patients with high mortality risk but also do it using only the necessary features. This can enable hospitals to make early predictions even when only basic features are available while evaluating the benefits of later obtaining more complex biomarker features. The proposed DL model can be used for basic features. The proposed model outperformed an RF model when evaluating each of the proposed data sets. Both over-sampling and data imputation approaches were analysed. A data imputation method using the KNN algorithm was proposed and evaluated to improve MPCD.

Acknowledgements

Authors are particularly thankful to all participants for their voluntary participation in the research

Funding

This research received no external funding.

Abbreviations

Not applicable

Availability of data and materials

The data presented in this study are available on request from the corresponding author.

Ethics approval and consent to participate

Not applicable

Competing interests

The authors declare that they have no competing interests.

Consent for publication

Not applicable

Authors' contributions

J.L.G.A. and F.H.; formal analysis, methodology and investigation R.M.M and E.R.F; project administration, A.G.Z. and R.R.M.; visualization, and F.H. and J.L.G.A.; writing—original draft. All authors have read and agreed to the published version of the manuscript.

Authors' information

The authors affiliated with the School of Engineering and Sciences, Tecnológico de Monterrey, Av. E Garza Sada 2501, 64,849 Monterrey, NL México

Author details

School of Engineering and Sciences, Tecnológico de Monterrey, Monterrey, Mexico.

References

1. Sanghvi, S.K., Schwarzman, L.S., Nazir, N.T.: Cardiac mri and myocardial injury in covid-19: Diagnosis, risk stratification and prognosis. *Diagnostics* **11**(1), 130 (2021)
2. Jiang, X., Coffee, M., Bari, A., Wang, J., Jiang, X., Huang, J., Shi, J., Dai, J., Cai, J., Zhang, T.: Towards an artificial intelligence framework for data-driven prediction of coronavirus clinical severity. *CMC: Computers, Materials and Continua* **63**, 537–51 (2020)
3. Ghaffari, A., Meurant, R., Ardakani, A.: Covid-19 point-of-care diagnostics that satisfy global target product profiles. *Diagnostics* **11**(1), 115 (2021)
4. Ji, D., Zhang, D., Xu, J., Chen, Z., Yang, T., Zhao, P., Chen, G., Cheng, G., Wang, Y., Bi, J.: Prediction for progression risk in patients with covid-19 pneumonia: the call score. *Clinical Infectious Diseases* (2020)
5. Pourhomayoun, M., Shakibi, M.: Predicting mortality risk in patients with covid-19 using artificial intelligence to help medical decision-making. *medRxiv* (2020)
6. Khan, Y., Abbas, S., Truong, B.-C.: Machine learning-based mortality rate prediction using optimized hyper-parameter. *Computer Methods and Programs in Biomedicine* **197**, 105704 (2020)
7. Roser, M., Ritchie, H., Ortiz-Ospina, E., Hasell, J.: Coronavirus pandemic (covid-19). *Our World in Data* (2020)
8. Burdick, H., Lam, C., Mataraso, S., Siefkas, A., Braden, G., Dellinger, R.P., McCoy, A., Vincent, J.-L., Green-Saxena, A., Barnes, G.: Prediction of respiratory decompensation in covid-19 patients using machine learning: The ready trial. *Computers in biology and medicine* **124**, 103949 (2020)
9. Kim, Y.J., Saqlian, M., Lee, J.Y.: Deep learning-based prediction model of occurrences of major adverse cardiac events during 1-year follow-up after hospital discharge in patients with ami using knowledge mining. *Personal and Ubiquitous Computing*, 1–9 (2019)
10. Bertsimas, D., Boussiou, L., Wright, R.C., Delarue, A., Digalakis Jr, V., Jacquillat, A., Kitane, D.L., Lukin, G., Li, M.L., Mingardi, L.: From predictions to prescriptions: A data-driven response to covid-19. *arXiv preprint arXiv:2006.16509* (2020)
11. Panwar, H., Gupta, P., Siddiqui, M.K., Morales-Menendez, R., Singh, V.: Application of deep learning for fast detection of covid-19 in x-rays using ncovnet. *Chaos, Solitons and Fractals*, 109944 (2020)
12. Panwar, H., Gupta, P., Siddiqui, M.K., Morales-Menendez, R., Bhardwaj, P., Singh, V.: A deep learning and grad-cam based color visualization approach for fast detection of covid-19 cases using chest x-ray and ct-scan images. *Chaos, Solitons and Fractals* **140**, 110190 (2020)
13. Selvaraju, R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Visual explanations from deep networks via gradient-based localization. In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 618–626 (2017)
14. Rezaei, M., Razavi Bazaz, S., Zhand, S., Sayyadi, N., Jin, D., Stewart, M.P., Ebrahimi Warkiani, M.: Point of care diagnostics in the age of covid-19. *Diagnostics* **11**(1), 9 (2021)
15. Yadaw, A.S., Li, Y.-c., Bose, S., Iyengar, R., Bunyavanich, S., Pandey, G.: Clinical features of covid-19 mortality: development and validation of a clinical prediction model. *The Lancet Digital Health* **2**(10), 516–525 (2020)
16. Zhao, Z., Chen, A., Hou, W., Graham, J.M., Li, H., Richman, P.S., Thode, H.C., Singer, A.J., Duong, T.Q.: Prediction model and risk scores of icu admission and mortality in covid-19. *PloS one* **15**(7), 0236618 (2020)

17. Lundberg, S.M., Lee, S.-I.: A unified approach to interpreting model predictions. In: *Advances in Neural Information Processing Systems*, pp. 4765–4774 (2017)
18. Escobar, C., Arinez, J., Morales-Menendez, R.: Process-monitoring-for-quality-a step forward in the zero defects vision. Report 0148-7191, SAE Technical Paper (2020)
19. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167* (2015)
20. Mohamad, I.B., Usman, D.: Standardization and its effects on k-means clustering algorithm. *Research Journal of Applied Sciences, Engineering and Technology* **6**(17), 3299–3303 (2013)
21. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research* **16**, 321–357 (2002)
22. Escobar, C.A., Morales-Menendez, R.: Machine learning and pattern recognition techniques for information extraction to improve production control and design decisions. In: *Industrial Conference on Data Mining*, pp. 286–300. Springer, ??? (2017)
23. Efron, B.: Missing data, imputation, and the bootstrap. *Journal of the American Statistical Association* **89**(426), 463–475 (1994)
24. Pandey, A., Jain, A.: Comparative analysis of knn algorithm using various normalization techniques. *International Journal of Computer Network and Information Security* **9**(11), 36 (2017)
25. (2019)
26. Bakator, M., Radosav, D.: Deep learning and medical diagnosis: A review of literature. *Multimodal Technologies and Interaction* **2**(3), 47 (2018)
27. Yeşilkanat, C.M.: Spatio-temporal estimation of the daily cases of covid-19 in worldwide using random forest machine learning algorithm. *Chaos, Solitons and Fractals* **140**, 110210 (2020)
28. Zhang, N., Wei, M., Fan, J., Aldhaferi, M., Zhang, Y., Bai, B.: Development of a hybrid scoring system for eor screening by combining conventional screening guidelines and random forest algorithm. *Fuel* **256**, 115915 (2019)
29. Suthaharan, S.: Support vector machine, pp. 207–235. Springer, ??? (2016)
30. Abraham, A.: Artificial neural networks. *Handbook of measuring system design* (2005)
31. Chen, T., Guestrin, C.: Xgboost: A scalable tree boosting system. In: *Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining*, pp. 785–794 (2016)
32. Kleinbaum, D.G., Dietz, K., Gail, M., Klein, M., Klein, M.: Logistic Regression. Springer, ??? (2002)

Figures

Figure 1

SHAP values of the selected features accordingly.

Figure 2

(a) Left side distribution of age and right side oxygen saturation (SPO2) and (b) left side comorbidities and right side multiple comorbidities.

Input: $CP(CP_1, CP_2, \dots, CP_m)$:

List of conditional probabilities ordered

Output: γ

Optimal classification threshold

Initialization: set $CCTL$ as empty

List of MPCD values associated to each candidate classification threshold

1. For $i = 1$ to $m - 1$ do begin
2. $CCT_i = \frac{CP_i + CP_{i+1}}{2}$
3. estimate $MPCD_i$ at each CCT_i
4. add $MPCD_i$ to $CCTL$
5. end
6. Find p , the position of the max ($CCTL$)
7. return $\gamma = CCT_p$

Figure 3

Pseudo-code of the OCTM algorithm.

Figure 4

Results summary of the DL model for every data set (Left - Normal, Right - SMOTE).

Figure 5

Boxplot for 10-fold CV on data set 4 for the normal (left) and SMOTE (right) approaches of the DL model.

Figure 6

RSE values for imputed NEU%m and LIN%M featuresl.

Figure 7

RSE values for amputated DDm and DDM features.

Figure 8

Metrics comparison of amputating the NEU%M and LIN%M biomarkers (Left) against real values (Right).

Figure 9

Metrics comparison of imputation the DD%M and DD%M biomarkers (Left) against real values (Right).

Figure 10

Comparison of results for the DL (left) and RF (right) models.

Figure 11

Comparison of results for the DL (left) and RF (right) models using SMOTE.

Figure 12

The original database class distribution on the left side, while the right side shows the DL predicted distribution.