Using Machine Learning to Predict the Death Risk in COVID-19 Patients

GIL-VERA, VICTOR DANIEL
Faculty of Engineering
Luis Amigó Catholic University
Transversal 51A # 67B 90
COLOMBIA

Abstract: - The COVID-19 pandemic caused a worldwide health crisis resulting in millions of deaths and infections, which led to the collapse of the intensive care units of many clinics and hospitals despite the strategies implemented by governments to prevent its proliferation, such as strict quarantines, social distancing, teleworking, among others. Predictive models are very useful to identify the mortality of infected patients. The objective of this study was to analyze several models used to categorize the patient's risk of passing away. According to the study's findings, the accuracy of the various models—logistic regression, K-nearest neighbors, support vector machines, Naive Bayes, decision trees, and random forest—was high (> 0.70), with random forests taking the lead (Accuracy=0.92), indicating that the models are reliable for predicting the risk of death in COVID-19 infection patients.

Key-Words: COVID-19, Databases, Limitations, Machine Learning, Predictive Models.
Received: March 16, 2022. Revised: October 9, 2022. Accepted: November 3, 2022. Published: December 1, 2022.

1 Introduction

The World Health Organization (WHO) proclaimed the COVID-19 virus a global pandemic on March 11, 2020, after it first appeared in Wuhan City, Hubei Province, China, at the end of 2019 [1].

The pandemic generated by the rapid proliferation of the virus generated major social, economic, cultural and health problems, including the death of millions of people. The collapse of intensive care units (ICU) led to the adoption of extreme strategies such as the prioritization of certain types of patients. Faced with this complex situation, predictive models are very useful tools to support health systems in decision-making and in formulating strategies to reduce the deaths of infected patients [2].

This study compares various supervised learning models (Logistic Regression, KNN, SVM, Naive Bayes, Decision Trees, and Random Forest) to estimate the likelihood that a patient with a viral infection will die. A free database of 566,602 sick patients from Mexico was used for this project; 70% of the sample was used for training and the other 30% for validation.

The Python programming language and the Google Colaboratory work environment were used. The rest of the paper is divided as follows: section 2 presents the review of the scientific background, section 3 presents the methodology used for the construction of the models, and section 4 presents the results and discussion. Finally, the paper concludes.

2 State of the art

Bibliometrics was performed with the bibliometrix package of the R-Cran 4.2.1 software. The keywords used were COVID-19, Predict*, and Mortality. The databases used were Scopus and WoS. The search period was from January 2020 to June 2022. The search equation used was:

1) TITLE-ABS-KEY ("COVID-19" AND "PREDICT*" AND "MORTALITY") AND (LIMIT-TO (DOCTYPE, "ar")) AND (LIMIT-TO (SUBJAREA, "COMP")) > 2019

Figure 1 shows the number of publications per year.

There is evidence of a constant increase, which is due to the availability of information on

infected patients that allows us to analyze the evolution of the virus worldwide.

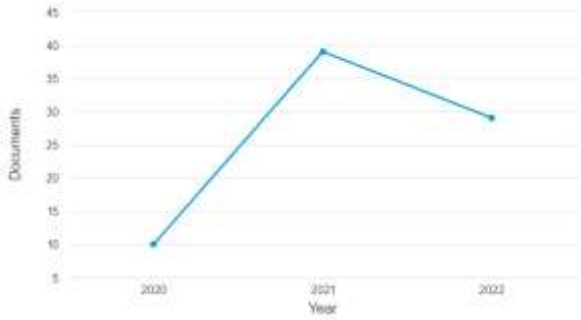

Fig 1. Publications by year

Figure 2 presents the Top 10 leading authors, most of whom have only two publications. Three of the top ten authors and their respective institutional affiliations are; Alnumay, Waleed S., King Saud University, Riyadh (Saudi Arabia), Jain, Rachna C., Bharati Vidyapeeth's College of Engineering, (New Delhi, India) and Homayounieh, Fatemeh Harvard Medical School, (Boston, USA).

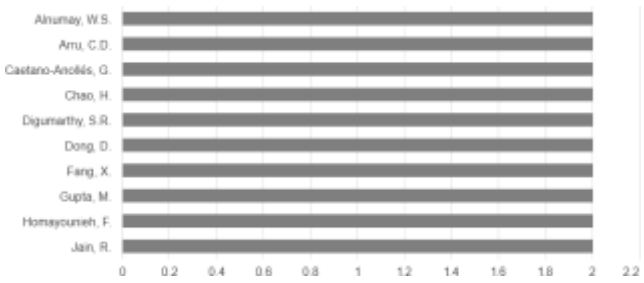

Fig 2. Main authors

The top 10 countries with the most publications are shown in Figure 3, led by the USA (18 publications), China (12 publications),

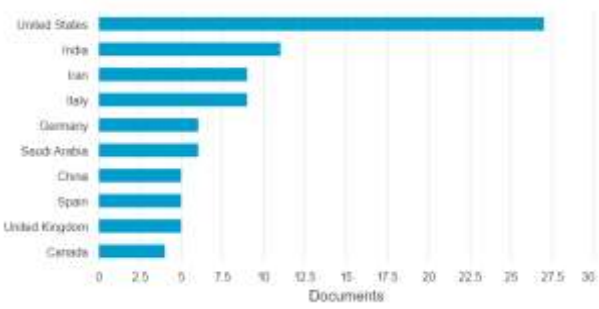

Figure 3. Main countries

Fig 4. Wordcloud

A wordcloud was constructed from the abstracts of the articles, as presented in Figure 4. The main words that stand out are; "COVID patients", "Deep learning" and "Sars COV". Other third and fourth and fifth order words emphasized relate to forecasting, prediction and supervised learning models. Table 1, presents the main descriptions of the documents analyzed.

Table 1. General information

Description	Results
Search time	2020:2021
Sources	51
Documents	66
Authors	32
Publications with only author	3
Publications with multiple authors	63
Colaboration index	4.81

For the review of the state of the art, a systematic literature review (SLR) was performed, which allowed us to identify a wide variety of documentation and literature and to learn more about predictive models that have been implemented in real life. Searches were conducted in the WHO database and Scopus. The following research questions were used:

- Q1. What models have been created, and how have they been applied, to forecast the probability of death in COVID-19-infected patients?

- Q2. What methods are employed by the prediction models to forecast the likelihood that COVID-19-infected individuals would pass away?

-Q3. Which databases or sources of data were used to create the present models?

The selection criteria for the publications finally selected were: academic level or quality

of the document, contribution to knowledge, clarity in writing and exposition of ideas, number of references (>15), originality and that they helped to answer the research questions. The following is a response to each of the research questions posed:

- Q1. What models have been created, and how have they been applied, to forecast the probability of death in COVID-19-infected patients?

The health crisis generated by COVID-19 has led the medical industry to develop and implement new technologies to monitor and control the spread of the virus [3]. To facilitate decision-making, the use of predictive models has increased massively, due to their great usefulness [4]. During the pandemic, different predictive models have been implemented to observe the behavior of COVID-19, to predict what will happen in the future and implement actions to reduce the mortality rate. Some of the models developed are prognostic, diagnostic and risk models [5].

Within the prediction models that have been developed, they have used information related to characteristics, comorbidities, habits, and climates, among others. In [6] a predictive model was developed considering different demographic characteristics such as age, gender, enzymes (ACE, ARB) and chronic diseases of a group of patients under treatment. These characteristics and the extraction of other variables were used as parameters for the model. They evaluated different deep learning and machine learning methodologies or approaches, to review the behavior of the data in each.

The results obtained and the implementation helped to assess the risk of death and the level of severity. The model allowed us to analyze each patient and implement early treatment to decrease the risk of contagion, complications, or death. In [7], they proposed regression models that predict the mortality rate using machine learning models and analyzed their relationship with climatic variables.

In [8], they used machine learning and evaluated its relationship with dietary habits in different countries, and developed a model that evaluated the relationship between the mortality rate of infected people and the type of diet. They conclude that obesity is a risk factor for death in this type of patient.

In [9] they developed a model based on regularization and cross-validation, using death data. They performed a comparison with different models to know the level of accuracy of the developed model and to evaluate the stability of the data handled. In [10] they built a predictive model based on artificial intelligence (AI) to help hospitals, make decisions about people who needed to receive priority care (hospitalization), classify patients when the system collapsed by overcrowding and eliminate delays in service delivery.

The model they developed considered different demographic and physiological characteristics and information on patients' symptomatology, and they adjusted to obtain as a result the probability and risk of death that a person has.

In Spain, they developed a predictive model for COVID-19 infections and deaths, based on Gompertz curves with data only on deaths and infections to limit the noise of external variables or unrelated data [4]. In Nigeria, they implemented a model to predict the daily incidence of COVID-19 based on the gender of individuals and conclude that males are at higher risk of infection [11].

In Pakistan, they developed a quadratic model with demographic information, which was used to predict the trend in cases of deaths from the virus. This model served as a guide to evaluating measures adopted by the Pakistani government; quarantines, social distancing, use of masks and smart lock-ups [12].

- Q2. What methods are employed by the prediction models to forecast the likelihood that COVID-19-infected individuals would pass away?

- Artificial Neural Networks (ANN): in [14] This technique was applied for the creation of a model for the detection and prediction of COVID-19 in rural areas. They classified images and different pre-trained architectures to be implemented in Deep-Learning models with lung imaging data. They trained and classified patient datasets to obtain a final prediction [13]. ANNs were also employed in other research to identify COVID-19-positive patients from hemograms, symptomatology and comorbidities.

- SIR model: in [15] they used this model in the observation of virus transmission, scale prediction, prediction of mortality and recovery rates. They established the following general

classification for all types of patients: Susceptible (S) (the patient did not contract the disease, but can be infected due to transmission from infected persons), Infected (I) (the patient has contracted the disease), Recovered / Deceased (R) (the disease can lead to one of two fates: either the patient survives and thus develops immunity to the disease or dies). These also employed a support vector machine (SVM) and an attribute reducer to minimize irrelevant and redundant information increase prediction accuracy [15].

-Machine Learning: [16] employed machine learning in the construction of a predictive model for prevent the mortality in an intensive care unit (ICU) patients, they used data from infected patients, which were divided into a training and test sample. They used different analyses for the selection of potential risks and five machine learning methods; AdaBoost, GBDT, XGBoost and CatBoost and regression methods. The model performance was evaluated with ROC/AUC (area under the curve) and through calibration, they analyzed the predicted and actual risk.

In [16] they propose a machine learning-based approach to aid safety and mitigate risks of COVID-19 in patients with poor medical history. They employed logistic regression and gradient augmentation decision trees to evaluate the behavior of the data in both approaches. They obtained better accuracy with the gradient augmentation decision tree model.

In [16], they developed a model with the same technique to predict the intubation of patients diagnosed or suspected of COVID-19. They distributed the data into several identifiers, positively and negatively labeled the registry of intubated patients according to ICU length of stay, and evaluated the accuracy through the ROC/AUC metric. They conclude that the developed model is acceptable with a ROC/AUC = 0.86.

-Surface Learning: in [16] they developed an expert system called COVIDC, this used images of chest CT scans using a web server. They used the classical support vector machine (SVM) model for diagnosis and prediction of patient severity. They used a Random Forest (RF) model to classify and optimize the features and a gradient boosting machine (XGBoost) to combine the trees by splitting or differentiating them into strong and weak.

-Cascade model: in [20] they built a system for personal monitoring of patients with COVID-19 using the Waterfall methodology. In summary, the system they developed is made up of the following components: An App that identifies the parameters for detection and/or prediction, real-time databases and follow-up tests to analyze the limitation and response of the system [20].

-SEAIHRDS mathematical model: in [21] they used this model to divide a population of infected patients into groups, to study different behaviors in the selected variables, which allowed them to simulate the progression of the pandemic. They also evaluated data to extract different rates related to the pandemic (death, contagions) and concluded that further research using artificial intelligence (AI) and Deep Learning (DL) is needed to mitigate the effect of health crises [22].

The lack of tools and information at the beginning of the pandemic contributed to the increase in infections and fatalities worldwide [23]. Due to the need to find different strategies that could help mitigate the pandemic, multiple investigations have been carried out to model and predict characteristics, however, being such a recent issue, accessing real patient information has become a difficult task [24], [25]. In addition, the data needed to feed the models are unstable, with large variations and differences from one country to another [26], [27].

Most of the models that have been built have been made with a specific country or population in mind, which limits their scope and application in other contexts [28]. Predicting the lethality rate in depth is a complex task, the data are variable, which can generate misleading results and limit the accuracy of the predictions [29].

-Q3. Which databases or sources of data were used to create the present models?

For the construction of the prediction models, databases made available by different organizations have been found. The dataset considered in an investigation in Italy was made available by the Italian civil protection department. https://github.com/pcm-dpc/COVID-19

From this database, information such as the number of recovered persons, ICU patients, discharged patients and daily deaths can be extracted [30]. In Nigeria, also considering the variable of daily deaths, COVID-19 incidence data by gender from April 11, 2020, to September 12, 2020, were used. This information was

obtained from the Nigerian Control Center and the data are available at: https://ncdc.gov.ng/diseases/sitreps [11].

Patients with solid tumors who had been diagnosed with COVID-19 infection and admitted to 32 hospitals in China between December 17, 2019, and March 18, 2020, were sampled for a study on the severity of the involvement of cancer patients and COVID-19. The real-time test with patients older than 18 years of age and the histological confirmation of a solid tumor were two of the inclusion criteria. In addition, a collection of patient data was made, which included, among other things, demographic information, cancer features, and smoking history. Benign tumors in patients led to their dismissal [31].

Another study conducted in China was applied to the Zhongnan Hospital of Wuhan University, where patients in critical condition were sampled, where the criteria for patient selection were as follows: respiratory failure requiring mechanical ventilation, shock, complications with another organ failure, requiring intensive care.

All patients were divided into two groups, survivors and non-survivors [32]. A study was conducted where data from Union Hospital was used, where through several selection criteria data was taken, some of these criteria were: patients aged 14 years or older and patients who were diagnosed with pneumonia. Patients were labeled as survivors or non-survivors [33].

For an advanced AI system, 284 COVID-19 images, 281 community-acquired pneumonia images, 293 secondary pulmonary tuberculosis images, and 306 healthy control images from local hospitals were used [34]. Finding quality CXR images to build diagnostic systems is difficult [35], [36]. In Seattle (USA) they also developed an investigation that used radiography with patients admitted to ICU, the data came from nine different hospitals, within the characteristics of the database were demographic information, chest radiographs and CT scans [37].

In North Africa, a study was conducted to know the affectation of COVID-19, WHO information was taken as a database in a period from March 02 to March 11, 2020. This information is updated every day and three indicators were taken such as confirmed cases, deaths attributable to COVID-19 and recovered cases of COVID-19 [38]. For the area of diabetes, different studies were conducted on the relationship between COVID-19 in people with diabetes, using databases with information from several countries in the world, to implement strategies for patients with this disease [39].

3 Methodology

Google Colaboratory and the NumPy, matplotlib and Pandas libraries were used to build the classification models in Python. A database developed by [40] was used, which gathered demographic information (gender, age, race/ethnicity), personal information (active/non-active smoker, number of medications), comorbidities (Diabetes, chronic obstructive pulmonary disease, asthma, pneumonia, immuno-suppressed, hypertension, another disease, cardiovascular disease, obesity and chronic renal failure). The variables available in the dataset are presented in Table 2.

Table 2. Database description

Variable	Description
Gender	Male, Female
Age	Integer
Type of patient	Outpatient, inpatient
Death by COVID 19	Yes, No
Tubing	
Pregnant	
Immunosuppressed	
Smoker	
Contact with another infected person	
Intensive care	
Comorbidity	Pneumonia, diabetes, COPD, asthma, hypertension, cardiovascular disease, obesity, chronic renal failure, chronic renal failure, other.

The classification models were built in Python, they predict whether the patient's risk of death is high or low based on the variables in the database, identify the correlation between the variables and predict the type of risk. The codes of the developed models are presented below:

```
Model - Random Forest
model = RandomForestClassifier(n_estimatiors=20, max_depth=10)
model.fit(X_train, y_train)
y_pred = model.predict(X-test)
y_pred_train = model.predict(X_train)


Model - Naïve Bayes
model = GaussianNB()
model.fit(normalized_X_train, y_train)
y_pred = model.predict(X-test)
y_pred_train = model.predict(normalized_X_train)


Model - Logistic Regression
LR = LogisticRegression()
LR.fit(X_train, y_train)
LR.score(X_train, y_train)
y_pred = LR.predict(X-test)
y_pred_train = model.predict(X_train)


Model - Support Vector Machine
model = SVC(gamma= 'auto' , kernel = 'linear')
model.fit(X_train, y_train)
y_pred = model.predict(X_test)
y_pred_train = model.predict(X_train)


Model - K-Nearest Neighbor (K-NN)
model = KNeighborsClassifier(n_neighbors=5, n_jobs=-1)
model.fit(X_train, y_train)
y_pred = model.predict(X-test)
y_pred_train = model.predict(normalized_X_train)

Model - Decision Tree
model = DecisionTreeClassifier(criterion='entropy', random_state=0)
model.fit(X_train, y_train)
y_pred = model.predict(X-test)
y_pred_train = model.predict(X_train)

Model - Artificial Neural Network
m= MLPRegressor(hidden_layer_sizes=(9,),
activation=logistic",
learning_rate=adaptative"
momentum=0.9,
learning_rate_init=0.01,
max_iter=1000,

)
```

The database was 566,602 infected Mexican patients. Seventy percent of the data were used for model training, the remaining 30% for validation. Before building the models, a normalization of the data set was performed using the Python function MinMaxScaler. In summary, the steps performed to implement the classification algorithms were:

- Import of the NumPy, matplotlib and Pandas libraries.
- Import of the patient database.
- Division of the dataset into a training sample (75%) and a test sample (25%).

4 Results and Discussion

The results presented in Table 3 show that the machine learning algorithms classify well the data set of infected patients. Random Forest offers the highest classification accuracy at 0.92.

Table 3. Model metrics

Metrics	Model						
	RF	KNN	SVM	NB	DT	LR	ANN
Accuracy	0.92	0.85	0.78	0.71	0.90	0.89	0.88
Erroneous Classification	0.08	0.15	0.22	0.29	0.10	0.11	0.12
VP	0.75	0.71	0.85	0.88	0.77	0.64	0.84
FP	0.25	0.29	0.15	0.12	0.23	0.36	0.16
VN	0.34	0.37	0.22	0.27	0.38	0.49	0.29
Precision	0.74	0.83	0.57	0.72	0.81	0.79	0.72
Prevalence	0.52	0.52	0.52	0.52	0.52	0.52	0.52

Note: Logistic Regression (LR), K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Naïve Bayes (NV), Decision Tree (DT), Random Forest (RF), Artificial Neural Networks (ANN).

5 Conclusion

In this paper, several classification models are implemented and their accuracy is compared. The classification accuracy of the different classifiers, such as logistic regression, KNN, SVM, Naïve Bayes, decision tree and random forest, exceeds 70%, with Random Forest with 0.92, so it can be affirmed that the models are valid for predicting the risk of death in patients infected with COVID-19.

The development of the prediction models reported in various research studies has been an option to support medical personnel and other health professionals attending the health crisis generated by COVID-19. It is necessary to mention that the perfect model does not exist; some different options and versions can be studied and implemented according to the scale and magnitude of the investigation. In addition, it should be noted that, although progress has been made and good use and implementation of technology has been found, it is essential to continue investigating and promoting the development of new models that serve at a general and not limited level, opting for the creation of global models and not reducing them to a demographic study that only allows applicability to some areas.

References:

[1] Cercas-Lobo, S. & Deniel-Rosanas, J. "COVID-19 and intracranial hemorrhage," Aten. Primaria Pract., vol. 3, no. 1, pp. 1–2, 2021, DOI: 10.1016/j.appr.2020.100078.

[2] Organizacion Panamericana de la Salud (OPS), "¿Por qué los modelos predictivos son cruciales en la lucha contra la covid-19 ?," Hoja Inf. OPS, pp. 1–5, 2020.

[3] Vaishya, R., Javaid, M., Khan, I. H., & Haleem, A. "Artificial Intelligence (AI) applications for COVID-19 pandemic," Diabetes Metab. Syndr. Clin. Res. Rev., vol. 14, no. 4, pp. 337–339, 2020, DOI: 10.1016/j.dsx.2020.04.012.

[4] Sánchez Villegas, P. & Daponte Codina, A., "Predictive models of the COVID-19 epidemic in Spain with Gompertz curves," Gac. Sanit., pp. 1–8, 2020, DOI: 10.1016/j.gaceta.2020.05.005.

[5] El-Solh, A.A, Lawson, Y., Carter, M., El-Solh, D.A & Mergenhagen, K. A. "Comparison of in-hospital mortality risk prediction models from COVID-19," PLoS One, vol. 15, no. 12 December, Dec. 2020, DOI: 10.1371/journal.pone.0244629.

[6] Kivrak, M., Guldogan, E., & Colak, C. "Prediction of death status on the course of treatment in SARS-COV-2 patients with deep learning and machine learning methods," Comput. Methods Programs Biomed., vol. 201, pp. 1–8, 2021, DOI: 10.1016/j.cmpb.2021.105951.

[7] Malki, Z., Atlam, E. S., Hassanien, A. E., G. Dagnew, G., Elhosseini, M. A. & Gad, I. "Association between weather data and COVID-19 pandemic predicting mortality rate: Machine learning approaches," Chaos, Solitons and Fractals, vol. 138, pp. 1--10, 2020, DOI: 10.1016/j.chaos.2020.110137.

[8] García-Ordás, M.T., Arias, N., Benavides, C., García-Olalla, O., & Benítez-Andrades, J.A. "Evaluation of Country Dietary Habits Using Machine Learning Techniques about Deaths from COVID-19," Healthcare, vol. 8, no. 4, pp. 1–10, 2020, DOI: 10.3390/healthcare8040371.

[9] Barigou, K., Loisel, S., & Salhi, Y. "Parsimonious predictive mortality modeling by regularization and cross-validation with and without covid-type effect," Risks, vol. 9, no. 1, pp. 1–18, 2021, DOI: 10.3390/risks9010005.

[10] Pourhomayoun, M. & Shakibi, M. "Predicting mortality risk in patients with COVID-19 using machine learning to help medical decision- making," Smart Heal., vol. 20, pp. 1–8, 2021, DOI: 10.1016/j.smhl.2020.100178.

[11] Olusola-Makinde, O., & Makinde, O. S. "COVID-19 incidence and mortality in Nigeria: gender based analysis," PeerJ, vol. 9, pp. 1–16, 2021, DOI: 10.7717/peerj.10613.

[12] Daniyal, R. O., Ogundokun, K., Abid, M. Khan, D. & Ogundokun, O. E. "Predictive modeling of COVID-19 death cases in Pakistan," Infect. Dis. Model., vol. 5, pp. 897–904, 2020, DOI: 10.1016/j.idm.2020.10.011.

[13] Qjidaa, M., Qjidaa, M., Mechbal, Y., Ben-Fares, A., Amakdouf, H., Maaroufi, M., Alami, B., & Qjidaa, H. "Early detection of COVID-19 by deep learning transfer Model for populations in isolated rural areas.," 2020 Int. Conf. Intell. Syst. Comput. Vision, ISCV 2020, 2020, DOI: 10.1109/ISCV49265.2020.9204099.

[14] Banerjee, A., Ray, S., Vorselaars, B., Kitson, J., Mamalakis, M., Weeks, S., Baker, M. & Mackenzie, L. S. "Use of Machine Learning and Artificial Intelligence to predict SARS-CoV-2 infection from Full Blood Counts in a population," Int. Immunopharmacol., vol. 86, pp. 1–8, 2020, DOI: 10.1016/j.intimp.2020.106705.

[15] Arun, S.S & Iyer, G.N. "On the Analysis of COVID19-Novel Corona Viral Disease Pandemic Spread Data Using Machine Learning Techniques," Proc. Int. Conf. Intell. Comput. Control Syst., vol. 1, pp. 1222–1227, 2020, DOI: 10.1109/ICICCS48265.2020.9121027.

[16] Pan, P., Li, Y., Xiao, Y., Han, B., Su, L., Su, M., ... & Xie, L. (2020). Prognostic assessment of COVID-19 in the intensive care unit by machine learning methods: model development and validation. Journal of medical Internet research, vol. 22, no. 11, pp. 1–16, 2020, DOI: 10.2196/23128.

[17] Gawade, P & Joshi, P.S. "Personification and Safety during pandemic of COVID19 using Machine Learning," Proc. 4th Int. Conf. Electron.

Contribution of individual authors to the creation of a scientific article (ghostwriting policy)

Victor Daniel Gil Vera has performed the normalization of the database, trained the predictive models in Python and performed the statistical analysis.

Sources of funding for research presented in a scientific article or scientific article itself

This research was funded by the Luis Amigó Catholic University and was one of the results of the research project entitled " Predicting the Death Risk in Patients with COVID-19 with Machine Learning ".