Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
The 3rd International Workshop on Statistical Methods and Artificial Intelligence (IWSMAI)
March 22-25, 2022, Porto, Portugal

Artificial Intelligence approaches to predict COVID-19 infection in Senegal

Abdoulaye Diallo\textsuperscript{a,b,}\textsuperscript{*}, Gaoussou Camara\textsuperscript{a}, Fodé Camara\textsuperscript{a}, Aminata Mboup\textsuperscript{b}

\textsuperscript{a}University Alioune Diop of Bambey, Bambey, Senegal
\textsuperscript{b}Institute for Health Research, Epidemiological Surveillance and Training (IRESSEF), Dakar, Senegal

Abstract

The SARS-CoV2 virus, which causes COVID-19 (coronavirus disease) has become a pandemic and has expanded all over the world. Because of increasing number of cases day by day, it takes time to interpret the data thus the limitations in terms of both treatment and findings are emerged. Due to such limitations, the need for clinical decision making system with predictive algorithms has arisen. Predictive algorithms could potentially ease the strain on healthcare systems by identifying the diseases. In this study, we design clinical predictive models that estimate, using artificial intelligence and data, which patients are susceptible to receive a COVID-19 disease. To evaluate the predictive performance of our models, accuracy, AUROC, and scores calculated. From 12,727 individuals, models were tested with basic information (sex, age) and the patient’s type of case, which is the combination of their symptoms, their travel during the last 14 days, their contact with an infected person or their participation in a festival requiring a gathering. We used 5 machine learning algorithms (LR, SVM, k-NN, RF, XGBoost) and 1 deep learning algorithm (ANN). Our models were validated with train-test split approach. The experimental results indicate that our predictive models identify patients that have COVID-19 disease at an accuracy of 73% and AUC of 69%. It is observed that predictive models trained on patients’ basic information and type of case could be used to predict COVID-19 infection in Senegal and can be helpful for medical experts to optimize the resources efficiently.

© 2022 The Authors. Published by Elsevier B.V.
This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)
Peer-review under responsibility of the Conference Program Chairs.

Keywords: SARS-CoV2; Coronavirus; Machine Learning; Deep Learning; Artificial Intelligence

\* Corresponding author. Tel.: +221 77 828 16 67.
E-mail address: abdoulaye.diallo7@uadb.edu.sn
1. Introduction

On 31 December 2019, the virus SARS-CoV2, which causes coronavirus diseases (COVID-19) was detected in Wuhan, China and since December 2019, it has spread all over the world [1]. World Health Organization (WHO) declared the COVID-19 outbreak is now pandemic, it will be essential to provide tools, mechanisms, and resources to quickly identify those at most risk of infirmity, and mortality. The novel coronavirus disease 2019 (COVID-19) pandemic caused by the SARS-CoV-2 continues to pose a critical and urgent threat to global health. The outbreak in early December 2019 in the Hubei province of the People’s Republic of China has spread worldwide. As of October 2020, the overall number of patients confirmed to have the disease has exceeded 39,500,000, in more than 180 countries, though the number of people infected is probably much higher [2]. More than 1,110,000 people have died from COVID-19 [3]. This pandemic continues to challenge medical systems worldwide in many aspects, including sharp increases in demands for hospital beds and critical shortages in medical equipment, while many healthcare workers have themselves been infected. Thus, the capacity for immediate clinical decisions and effective usage of healthcare resources is crucial. The most validated diagnosis test for COVID-19, using reverse transcriptase polymerase chain reaction (RT-PCR), has long been in shortage in developing countries. This contributes to increased infection rates and delays critical preventive measures [2]. The SARS-CoV2 affects various people in different ways. Yet over 80% infected people develop mild to moderate illness and recover without hospitalization [4,5]. Most common symptoms are fever, dry cough, and tiredness and in general, these symptoms begin as mild in all patients. However, severe symptoms such as chest pain or pressure, loss of speech or movement, and shortness of breath may be seen in a minority of patients [6,7]. Those who become more seriously ill are more likely to be older and male, with progressively more risk with each decade over the age of 50 [5]. In addition to these, the people with medical problems like diabetes, cancer, cardiovascular disease, and chronic respiratory disease are more likely to develop serious illness [4]. Although, there are many ongoing clinical trials evaluating potential treatments. These precautions are not for the treatment, yet they can protect people from the disease and slow the transmission of COVID-19[8]. In Senegal, all diagnostic laboratory tests for COVID-19 were performed according to criteria determined by the Ministry of Health and Social Action. All individuals who were tested for SARS-CoV-2 via RT-PCR assay of a nasopharyngeal swab. While subject to change, the criteria implemented during the study period included the presence and severity of clinical symptoms, possible exposure to individuals confirmed to have COVID-19, certain geographical areas, and the risk of complications if infected. All patients are categorized on a type of case which is the combination of their symptoms, their travel during the last 14 days, their contact with an infected person or their participation in a festival requiring a gathering. In addition, all negative and positive COVID-19 cases in this dataset were confirmed via RT-PCR assay. In this study, we provide a prediction model for COVID-19 infection by developing and applying various Artificial Intelligence (AI) application models. Six various AI algorithms models are designed and used on laboratory findings of patients. Each model was trained on data of 12,727 individuals tested in Senegal. Performance of the models are measured with accuracy and Area Under the ROC curve (AUROC). Thus, the best model can be implemented globally for effective screening and prioritization of testing for the virus in the general population.

2. Data and Methods

2.1. Artificial Intelligence applications models

Artificial Intelligence (AI) based algorithms learn from the historical data to provide predictions for the future outcomes. Machine learning (ML) and deep learning (DL) algorithms can be considered as a subset of the AI [9]. In this study, we develop and evaluate clinical predictive models to determine the COVID-19 infection with laboratory findings. To evaluate the study, we trained six different model types: Artificial Neural Network (ANN), Gradient Boosting (XGBoost), Random Forest (RF), k Nearest Neighbors (kNN), Logistic Regression (LR) and Supports Vectors Machine (SVM). ANN is an information processing approach that is inspired by the biological nervous system of human brain. It is composed of neurons, activation functions, input, output, and hidden layers. Gradient boosting is a ML technique for regression, classification and other tasks, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees[10,11]. When a decision tree is the weak learner, the resulting algorithm is called gradient boosted trees, which usually outperforms random forest[10,11,12].
RF are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time. For classification tasks, the output of the random forest is the class selected by most trees. For regression tasks, the mean or average prediction of the individual trees is returned [13,14]. RF correct for decision trees' habit of overfitting to their training set[15]. RF generally outperform decision trees, but their accuracy is lower than gradient boosted trees. However, data characteristics can affect their performance [16,17]. In statistics, the k-nearest neighbors algorithm (k-NN) is a nonparametric classification method first developed by Evelyn Fix and Joseph Hodges in 1951 [18], and later expanded by Thomas Cover[19]. It is used for classification and regression. In both cases, the input consists of the k closest training examples in data set. k-NN is a type of classification where the function is only approximated locally and all computation is deferred until function evaluation. Since this algorithm relies on distance for classification, if the features represent different physical units or come in vastly different scales then normalizing the training data can improve its accuracy dramatically[20,21]. LR (Logistic Regression) is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist. In regression analysis, LR is estimating the parameters of a logistic model (a form of binary regression) [22]. Mathematically, a binary logistic model has a dependent variable with two possible values, such as pass/fail which is represented by an indicator variable, where the two values are labeled "0" and "1". In ML, SVMs are supervised learning models with associated learning algorithms that analyze data for classification and regression analysis. SVMs are one of the most robust prediction methods, being based on statistical learning frameworks. Given a set of training examples, each marked as belonging to one of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier (although methods such as Platt scaling exist to use SVM in a probabilistic classification setting).

2.2. Data Description

The dataset is about 12,727 individuals who were tested for SARS-CoV-2 via RT-PCR assay of a nasopharyngeal swab. It contains initial records, on a daily basis, of these patients who were tested for COVID-19 in Senegal. In addition to the test date and result, various information is available, including clinical symptoms, sex and age. However, patients are categorized into types of case which are defined based on their symptoms, their travel during the last 14 days, their contact with an infected person or their participation in a festival requiring a gathering. According to the note of the Senegalese Ministry of Health and Social Action, dated on July-30-2021, we have extracted:

✓ **Suspected case:**

  a) a person who meets the clinical and epidemiological criteria:
  - clinical criteria:
    - Acute appearance of fever and cough;
    - OR
    - Acute appearance of at least three of the following signs or symptoms: fever, cough, general weakness / tired, headache, myalgia, sore throat, runny nose, dyspnea, nausea, vomiting, diarrhea, altered mental status.
  - epidemiological criteria:
    - reside or work in an area at high risk of transmission of the virus: closed residential settings, healthcare settings, including healthcare institution or within the community; at any time within 14 days of appearance of symptoms.
    - OR
    - reside or travel to an area of community transmission at any time within 14 days of onset of symptoms.

  b) a patient with severe acute respiratory disease (SARI): severe acute respiratory infection with a history of fever $\geq 38^\circ C$; and cough; with appearance in the last 10 days; and requiring hospitalization.
✓ **Probable case:**
- a suspected case with chest imaging showing results suggestive of COVID-19 disease;
- a person presenting with anosmia (loss of smell) or ageusia (loss of taste) of recent appearance, in the absence of any other identified cause;
- a patient who meets the above criteria AND who is a contact of a probable or confirmed case or linked to a COVID-19 cluster;
- death not otherwise explained, in an adult with respiratory distress preceding death AND being a contact of a probable or confirmed case linked to a COVID-19 cluster.

The table 1 shows the characteristics of the dataset and the features used by the model:

Table 1. Characteristics of the dataset and the features used by the model

| (#) Features | Values (code) | Total (%) | (0) Negative RT_PCR (%) | (1) Positive RT_PCR (%) |
|--------------|--------------|-----------|-------------------------|------------------------|
|              |              | n = 12 727 (100) | n = 9 142 (71.83) | n = 3 585 (28.17) |
| (1) Sex      | male (0)     | n | % | n | % | n | % |
|              | female (1)   | 5 791 | 45.50 | 4 157 | 32.66 | 1 634 | 12.84 |
| (2) Age      | child (0)    | 797 | 6.26 | 672 | 5.28 | 125 | 0.98 |
|              | young (1)    | 4 985 | 39.17 | 3 987 | 31.33 | 998 | 7.84 |
|              | adult (2)    | 4 672 | 36.17 | 3 146 | 24.72 | 1 526 | 11.99 |
|              | old (3)      | 2 273 | 17.86 | 1 337 | 10.51 | 936 | 7.35 |
| (3) Suspected| false (0)    | 5 293 | 41.59 | 4 515 | 35.48 | 778 | 6.11 |
|              | true (1)     | 7 434 | 58.41 | 4 627 | 36.36 | 2 807 | 22.06 |
| (4) Probable | false (0)    | 7 434 | 58.41 | 4 627 | 36.36 | 2 807 | 22.06 |
|              | true (1)     | 5 293 | 41.59 | 4 515 | 35.48 | 778 | 6.11 |

The following list describes each of the dataset’s features used by the model:

A. Basic information:
- Sex (male/female)
- Age (child: [0-14 ans]; young: [15-39 ans]; adult: [40-64 ans]; old: >= 65 ans)

B. Type of case:
- Suspected: (true / false)
- Probable (true / false)

3. Results and discussion

For the development of the six models, we split the dataset into training and testing sets. Although k fold cross-validation approach is frequently used in artificial intelligence in health studies especially in cases of relatively small samples, it generates less clearly the results in clinical applications [23]. 80% of the data was used for the models’ training and 20% for their testing. For each model, we calculated the accuracy and the Area under the ROC curve (AUROC) as can be seen on the Table 2.
Table 2. Evaluation results of all AI application models with train-test split approach

| AI approaches | Accuracy | AUROC  |
|---------------|----------|--------|
| kNN           | 0.658    | 0.556  |
| SVM           | 0.715    | 0.5    |
| LR            | 0.722    | 0.667  |
| RF            | 0.719    | 0.693  |
| XGBoost       | 0.727    | 0.691  |
| ANN           | 0.729    | 0.690  |

According to this table, the accuracy and the AUROC results of all artificial intelligence models were respectively reached at least 65.8% and 55.6% and above. The clinical predictive performance of all algorithms was better with an AUROC of 0.69 and accuracy of 0.73 for the best-performing algorithms, which were XGBoost and ANN models. RF was observed as the second-best model with an accuracy of 0.72 and an AUROC of 0.69. AUROC is used in the classification analysis to determine which of the used models predicts the classes best. In general, an AUROC score of 0.5 means that there is no discrimination. An AUROC score between 0.6 and 0.8 is considered acceptable, a score between 0.8 and 0.9 is considered excellent and more than 0.9 is considered outstanding [24]. Logistic Regression (LR), Random Forest (RF), Gradient Boosting (XGBoost) and Artificial Neuron Network (ANN) can be used for clinical prediction of COVID-19 in Senegal since their AUROC score is between 0.6 and 0.8. In addition, AUROC score has a vital role on medical researches, since it has a meaningful interpretation for disease classification from healthy subjects [25,26]. However, XGBoost and ANN are the best models when their accuracy are gretter than other models’ accuracy. Accuracy is a research characteristic, which provides a way to know how close are the sample parameters to population characteristics [27]. By measuring the accuracy of the models, the researcher can prove that the research is generalizable, reliable, and valid [28]. Thus, in this study, only these two-evaluation metrics were considered. To compare our results with previous studies, we list three studies on the Table 3.

Table 3. Comparison of results with previous work

| References | AI Technique | Classifier | Accuracy | AUROC  |
|------------|--------------|------------|----------|--------|
| [23]       | Machine Learning | SVM       | 80%      | -      |
| [29]       | Machine Learning | SVM, RF   | -        | 0.87   |
| [30]       | Machine Learning | XGBoost   | -        | 0.66   |
| this work  | Machine Learning | XGBoost   | 73%      | 0.69   |
| this work  | Deep Learning  | ANN       | 73%      | 0.69   |

In the study of [23,29,30], authors used machine learning techniques. As seen in Table 3, best classification was obtained with SVM classifier in these studies. Yet, in this research, we did not use only machine learning. We developed five different machine learning application models and one deep learning application model, and reached better accuracy and AUROC scores than machine learning classifiers except the XGBoost which has the same scores.

4. Conclusion

In this study, the prediction of COVID-19 in Senegal was carried out with artificial intelligence models based on the patients’ basic information and their type of case which is the combination of their symptoms, their travel during the last 14 days, their contact with an infected person or their participation in a festival requiring a gathering. Classification was carried out and the performances of the models were measured with accuracy and AUROC. To validate the models, we applied a train-test split approach. The best accuracy and AUROC (values) were obtained.
with ANN (73%, 0.69) and XGBoost (73%, 0.69) models. Further studies need to be carried out with image data from CT scans of patients tested for COVID-19. With these data, other deep learning algorithms such as Convolution Neurons Network (CNN), Long-Short Term Memory (LSTM), Recurrent Neural Networks (RNN), CNNLSTM, and CNNRNN can be used to detect COVID-19 infection in Senegal and Africa with scores between 0.8 and 0.9 or better. Various real-time RT-PCR protocols have been proposed for the diagnosis of COVID-19 [31]. RT-PCR tests performance is impacted by several factors that are difficult to measure, such as low levels of shedding during incubation and early infection, variability in the site of specimen acquisition, and sufficiency of sample collected [32 – 34]. In the light of all these flaws, these modelling techniques reveal the importance for early detection of COVID-19 infection and to start treatment without delay. In conclusion, we found evidence to suggest that artificial intelligence application models can be applied to predict COVID-19 infection. Based on our study’s results, we conclude that healthcare systems should explore the use of predictive models that assess individual COVID-19 risk in order to improve healthcare resource prioritization and inform patient care.

Acknowledgements

We would like to thank Professor Souleymane Mboup, President and Founder of IRESSEF, who gave us the opportunity to work on Senegal’s program to support the response to the Covid-19 pandemic.

References

[1] World Health Organization, Report of the WHO-China joint mission on coronavirus disease (COVID-19). 2020 https://www.who.int/docs/default-source/coronaviruse/who-china-joint-mission-on-covid-19-final-report.pdf.
[2] Zoabi, Y. Deri-Rovoz, S. & Shomron, N. Machine learning-based prediction of COVID-19 diagnosis based on symptoms. 2021 https://doi.org/10.1038/s41746-020-00372-6.
[3] Dong, E., Du, H. & Gardner, L. An interactive web-based dashboard to track COVID-19 in real time. Lancet Infect. Dis.. https://doi.org/10.1016/S14733099(20)30120-1 (2020).
[4] World Health Organization, Health topics, coronavirus. 2020 https://www.who.int/health-topics/coronavirus#tab=tab_3.
[5] National Institute of Infection Diseases, Field briefing: diamond princess COVID-19 cases. 2020 https://www.niid.go.jp/niid/en/2019-ncov-e/9407-covid-dp-fe-01.html.
[6] Del Rio C, Malani PN. Novel coronavirus – important information for clinicians. J Am Med Assoc 2019;323(11):2020. doi:10.1001/jama.2020.1490
[7] Wang D, et al. Clinical characteristics of 138 hospitalized patients with 2019 novel coronavirus-infected pneumonia in Wuhan, China. J Am Med Assoc 2020;323(11):1061-1069. doi:10.1001/jama.2020.1585.
[8] Alakus, B.T., Turkoglu, I. Comparison of deep learning approaches to predict COVID-19 infection. 2020 https://doi.org/10.1016/j.chaos.2020.110120
[9] Gavriloa, Y. Artificial Intelligence vs Machine Learning vs Deep Learning: Essentials. 2020. https://serokeli.io/blog/ai-ml-dl-difference
[10] Piriyonesi, S. Madeh; El-Diraby, Tamer E. (2020-03-01). "Data Analytics in Asset Management: Cost-Effective Prediction of the Pavement Condition Index". Journal of Infrastructure Systems. 26 (1): 04019036. doi:10.1061/(ASCE)IS.1943-555X.0000512. ISSN 1943-555X.
[11] Hastie, T.; Tibshirani, R.; Friedman, J. H. (2009). "10. Boosting and Additive Trees". The Elements of Statistical Learning (2nd ed.). New York: Springer. pp. 337–384. ISBN 978-0-387-84857-0. Archived from the original on 2009-11-10
[12] Piriyonesi, S. Madeh; El-Diraby, Tamer E. (2021-02-01). "Using Machine Learning to Examine Impact of Type of Performance Indicator on Flexible Pavement Deterioration Modeling". Journal of Infrastructure Systems. 27 (2): 04021005. doi:10.1061/(ASCE)IS.1943-555X.0000602. ISSN 1076-0342.
[13] Ho, Tin Kam (1995). Random Decision Forests (PDF). Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, 14–16 August 1995. pp. 278–282. Archived from the original (PDF) on 17 April 2016. Retrieved 5 June 2016.
[14] Ho TK (1998). "The Random Subspace Method for Constructing Decision Forests" (PDF). IEEE Transactions on Pattern Analysis and Machine Intelligence. 20 (8): 832–844. doi:10.1109/34.709601.

[15] Hastie, Trevor; Tibshirani, Robert; Friedman, Jerome (2008). The Elements of Statistical Learning (2nd ed.). Springer. ISBN 0-387-95284-5.

[16] Piryonesi S. Madeh; El-Diraby Tamer E. (2020-06-01). "Role of Data Analytics in Infrastructure Asset Management: Overcoming Data Size and Quality Problems". Journal of Transportation Engineering, Part B: Pavements. 146 (2): 04020022. doi:10.1061/JPEODX.0000175

[17] Piryonesi, S. Madeh; El-Diraby, Tamer E. (2021-02-01). "Using Machine Learning to Examine Impact of Type of Performance Indicator on Flexible Pavement Deterioration Modeling". Journal of Infrastructure Systems. 27 (2): 04021005. doi:10.1061/(ASCE)IS.1943-555X.0000602. ISSN 1076-0342.

[18] Fix, Evelyn; Hodges, Joseph L. (1951). Discriminatory Analysis. Nonparametric Discrimination: Consistency Properties (PDF)(Report). USAF School of Aviation Medicine, Randolph Field, Texas.

[19] Altman, Naomi S. (1992). "An introduction to kernel and nearest-neighbor nonparametric regression" (PDF). The American Statistician. 46 (3): 175–185

[20] Piryonesi S. Madeh; El-Diraby Tamer E. (2020-06-01). "Role of Data Analytics in Infrastructure Asset Management: Overcoming Data Size and Quality Problems". Journal of Transportation Engineering, Part B: Pavements. 146 (2): 04020022. doi:10.1061/JPEODX.0000175

[21] Hastie, Trevor. (2001). The elements of statistical learning: data mining, inference, and prediction: with 200 full-color illustrations. Tibshirani, Robert., Friedman, J. H. (Jerome H.). New York: Springer. ISBN 0-387-95284-5. OCLC 46809224

[22] Tolles, Juliana; Meurer, William J (2016). "Logistic Regression Relating Patient Characteristics to Outcomes". JAMA. 316 (5):533–4. doi:10.1001/jama.2016.7653. ISSN 0098-7484. OCLC 6823603312. PMID 27483067

[23] Batista, A.F., Miraglia, J.L., Donato, T.H.R., and Filho, A.D.P.C., COVID-19 diagnosis prediction in emergency care patients: a machine learning approach, medRxiv, 2020. doi: 10.1101/2020.04.04.20052092.

[24] Mandrekar JN. Receiver operating characteristic curve analysis for medical diagnostic test evaluation. J Thoracic Oncol 2010;5(9):1315–16. doi:10.1097/JTO.0b013e3181ec173d.

[25] Hajian-Tilaki K. Receiver operating characteristic (ROC) curve analysis for medical diagnostic test evaluation. Caspian J Intern Med 2013;4(2):627–35.

[26] Kamarudin AN, Cox T, Kolamunnage-Dona R. Time-dependent ROC curve analysis in medical research: current methods and applications. BMC Med Res Methodol 2017;17(53). doi:10.1186/s12874-017-0332-6.

[27] Wynants L, Calster BV, Collins GS, Riley RD, Heinzle G, Schuit E, et al. Prediction model for diagnosis and prognosis of covid-19: systematic review and critical appraisal. BMJ 2020;369. doi:10.1136/bmj.m1328.

[28] Pierce R. Evaluating information: validity, reliability, accuracy, triangulation. Res Methods Polit 2008:79–99. doi:10.4135/9780857024589.

[29] Jiang X, Coffee M, Bari A, Wang J, Jiang X, et al. Towards an artificial intelligence framework for data-driven prediction of coronavirus clinical severity. Compu Mater Continua 2020;63(1):537–51. doi:10.32604/cmc.2020.010691

[30] Schwab, P., Schütte, A.D., Dietz, B., and Bauer, S. “predCOVID-19: a systematic study of clinical predictive models for coronavirus disease 2019, arXiv:2005. 08302, 2020.

[31] World Health Organization. Laboratory testing for coronavirus disease 2019 (COVID-19) in suspected human cases: interim guidance. https://www.who.int/publications-detail/laboratory-testing-for-2019-novel-coronavirus-in-suspected-human-cases-20200117. (Updated on March 19, 2020).

[32] Wölfel R, Corman VM, Guggemos W, et al. Virological assessment of hospitalized patients with COVID-2019. Nature 2020;581:465–9. doi:10.1038/s41586-020-2196-x.

[33] Wang, W., Xu, Y., Gao, R., and et al. “Detection of SARS-CoV-2 in different types of clinical specimens,” JAMA, 323(18), 1843–4, 220. doi: 10.1001/jama.2020.3786.

[34] Fang Y, Zhang H, Xie J, et al. Sensitivity of chest CT for COVID-19: comparison to RT-PCR. Radiology 2020. doi:10.1148/radiol.2020200432.