LOS LIBERTADORES
FUNDACIÓN UNIVERSITARIA

MODELO DE CLASIFICACIÓN MACHINE LEARNING PARA PRONOSTICAR SECUELAS FÍSICAS EN PACIENTES POST COVID

MACHINE LEARNING CLASSIFICATION MODEL TO PREDICT PHYSICAL SEQUELS IN POST-COVID PATIENTS

José Julián Figueroa Arias, jjfigueroaa@libertadores.edu.co
José John Fredy González Veloza, jjgonzalezv02@libertadores.edu.co

RESUMEN

El Covid 19 es un virus infeccioso que produce un síndrome respiratorio agudo severo, y entre los síntomas más frecuentes son los síntomas respiratorios, la fiebre y también síntomas gastrointestinales. Una de las características de este virus es que luego del periodo de recuperación, en algunos casos se presentan secuelas físicas tales como tos, pérdida de olfato, dolores musculares, dolor de cabeza, etc, secuelas que han ocasionado víctimas fatales en todo el mundo. Por lo tanto, en este estudio se realizó un modelo de clasificación machine learning para pronosticar secuelas físicas en pacientes postcovid, como muestra se utilizó información de 1436 observaciones de pacientes del Hospital Universitario Departamental de Nariño quienes resultaron positivos para covid 19, luego de la recuperación de estos pacientes se obtuvo información sobre la variable de interés para esté estudio que fue la presentación de secuelas físicas post covid. Se obtuvo que el modelo con mejores métricas de desempeño fue el de árboles de clasificación con auc de 0.73. Se concluye que el modelo de clasificación es útil para identificar los posibles casos de individuos con secuelas postcovid y de esa manera gestionar las acciones hospitalarias para disminuir complicaciones y víctimas fatales después del periodo de recuperación causado por el virus Covid – 19.

Palabras clave: Covid 19, secuelas físicas, machine learning, modelo de clasificación.

ABSTRACT

Covid 19 is an infectious virus that produces a severe acute respiratory syndrome, and among the most frequent symptoms are respiratory symptoms, fever and also

gastrointestinal symptoms. One of the characteristics of this virus is that after the recovery period, in some cases there are physical sequelae such as coughing, loss of smell, muscle pain, headache, etc., sequelae that have caused fatalities throughout the world. Therefore, in this study, a machine learning classification model was carried out to predict physical sequelae in postcovid patients, as a sample, information was obtained from 1436 observations of patients from the Nariño Departmental University Hospital who were positive for covid 19, after recovery. Information was obtained from these patients on the variable of interest for this study, which was the presentation of post-COVID physical sequelae. It was found that the model with the best performance metrics was the classification tree with auc of 0.73. It is concluded that the classification model is useful to identify possible cases of individuals with post-covid sequelae and thus manage hospital actions to reduce complications and fatalities after the recovery period caused by the Covid-19 virus.
Keywords: Covid 19, physical sequelae, machine learning, classification model.

INTRODUCCIÓN

El primero de diciembre de 2019, en la ciudad de Wuhan, China se detecta el primer paciente sintomático del virus SARS-CoV-2, las manifestaciones clínicas son similares a las de una neumonía viral. La OMS, en colaboración con la Organización Mundial de Sanidad Animal y la Organización de las Naciones Unidas para la Alimentación y la Agricultura, denominaron a la enfermedad Covid-19 (OMS, 2020). Diversos estudios argumentan que el Covid – 19 presenta varias rutas de transmisión, entre las cuales se encuentran las rutas por contacto directo tales como inhalación de gotas gruesas y pequeñas, tos y estornudos. También se encontró que la transmisión puede producirse mediante el contacto con personas que tienen el virus Covid 19, pero que no presentan sintomatología asociada. (Parra, Bermúdez, Peña & Rueda. 2020). El Covid 19 es un virus infeccioso que produce un síndrome respiratorio agudo severo, y entre los síntomas más frecuentes son los síntomas respiratorios, la fiebre y también síntomas gastrointestinales. También existen factores en el individuo que pueden afectar el impacto en el organismo, sin embargo, aún no se encuentran muchos estudios al respecto. Por lo tanto, actualmente aún existe falta de información clínica sobre los aspectos a tener en cuenta a la hora de abordar este virus en personas contagiadas. Además, en algunos casos se ha encontrado que luego de periodo de recuperación se presentan secuelas físicas que pueden llevar a que los síntomas iniciales se mantengan por un tiempo que va desde una semana a un año, secuelas que han ocasionado víctimas fatales (Coureaux, Cuevas, 2021).

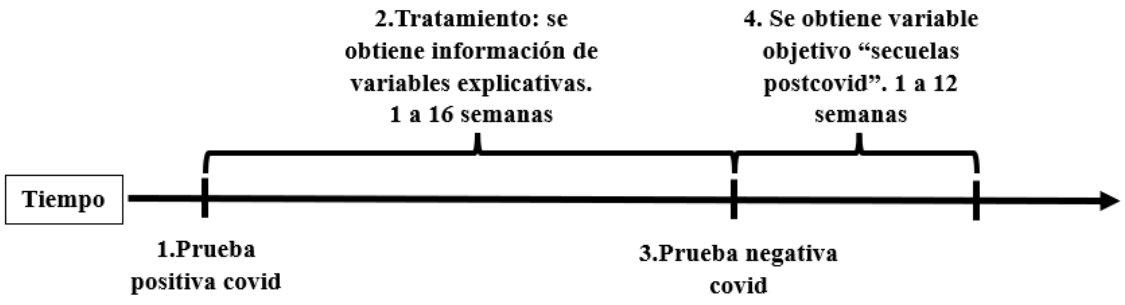
En cuanto a las secuelas de las personas que han tenido el virus Covid 19 se encuentra que a gran cantidad de la población que estuvo infectada con el virus, tuvieron secuelas, entre las cuales se encuentran las físicas tales como tos, perdida de olfato, dolor en el pecho, saturación baja y dolor de garganta, También se encuentran las secuelas psicológicas, entre las que se encuentran la ansiedad y depresión.

En Colombia se han presentado 6.309.716 millones de casos de los cuales se han recuperado 6.137.878. los casos que han tenido desenlace fatal son 141.837. Específicamente en el departamento de Nariño se han presentado 107.959 casos, de los cuales han fallecido 3426. Sin embargo, no se cuenta con información sobre secuelas causadas por Covid -19 y tampoco estudios que permitan identificar que variables influyen en la aparición de estas. (Instituto Departamental de Salud de Nariño, 2022).

Por lo tanto, este estudio nace debido a la necesidad de identificar a individuos que probablemente tendrán secuelas postcovid y gestionar tratamientos adecuados a los pacientes, favoreciendo la recuperación del paciente, generando que se reincorpore a sus labores cotidianas. Una de las herramientas utilizadas en el campo médico para el pronóstico de variables es el Machine Learning (Arash, Mina, Mehmet & Shira, 2020, Abdel, 2021). Por lo tanto, este estudio tuvo como objetivo el desarrollo de un modelo que permita pronosticar pronosticar secuelas físicas en pacientes postcovid, como base de datos se utilizó información de 1436 observaciones de pacientes del Hospital Universitario Departamental de Nariño quienes resultaron positivos para covid 19, luego de la recuperación de estos pacientes se obtuvo información sobre la variable de interés para esté estudio que fue la presentación de secuelas físicas post covid, esta información se tomó con cohorte a 30 de Julio de 2022. Para la construcción del modelo de inteligencia artificial ya que permite inferir reglas para establecer predicciones de manera automática (Geron, 2019, Dong 2021).

Entre las variables utilizadas en este estudio se encuentran, variables de características personales del individuo, el tipo de medicamento utilizado para el tratamiento, tipo de remedio casero utilizado para el tratamiento, y el tipo de síntoma presentado durante el periodo de latencia del Covid 19. A continuación en el gráfico 1, se muestra una línea de tiempo que sigue un paciente desde el momento que da positivo en covid hasta que se diagnostica con o sin secuelas postcovid. Este diagrama a su vez presenta el momento en el cual se recopila la información de este paciente y como debería capturarse en caso de que el modelo se utilice en producción.

Gráfico No1. Diagrama de obtención de datos.



METODOLOGÍA

Datos

Los datos utilizados en el presente estudio fueron brindados por el Hospital Universitario Departamental de Nariño a partir de la aplicación de 13232 pruebas de COVID 19, de las cuales 10536 resultaron negativas y 2698 resultaron positivas, teniendo en cuenta la revisión de los datos solamente se tiene la información completa de las 67 variables en 1436 observaciones. Para este estudio se tomaron en cuenta las siguientes fases: a. selección de las variables relevantes. b. análisis descriptivo. 3. construcción de modelos y su evaluación de desempeño.

Procesamiento y modelación.

En cuanto a la construcción de los modelos y su evaluación, se realizó la preparación requerida para este paso, inicialmente se realizó una partición de la base de datos, una base de datos de entrenamiento (80%) y testeo (20%). En este caso la variable objetivo Secuelas Poscovid. De la totalidad de las observaciones el 22% equivalente a 317 personas tuvieron secuelas, mientras que el 78% no tuvieron secuelas.

En cuanto a la modelación se seleccionaron las variables más relevantes a partir de un proceso de future selección, entre los parámetros que se tienen en cuenta son la baja varianza y por correlación, para este caso el método más efectivo fue la selección de variables que tuvieran baja varianza. Posteriormente se realizó la búsqueda del mejor modelo, teniendo en cuenta su métrica de auc. Las tres fases del análisis se realizaron en Python (v. 3.6) con la librería pycaret (https://pycaret.org/) y Scikit Learn (https://scikit-learn.org/stable/) y se pueden consultar en el siguiente enlace [https://github.com/alruizzo/missing_persons]. Finalmente se realizó la construcción de una tabla de probabilidades distribuida en 10 deciles para la toma de decisiones de gestión de la población respecto al fenómeno estudiado.

RESULTADOS

Entre las variables predictoras importantes que se seleccionan para este estudio se encuentran la edad, sexo, síntomas, medicamentos y remedios caseros. Al observar la Figura 1a referente a la variable edad comparada con la variable objetivo secuelas postcovid, se puede identificar que el pico máximo de frecuencia de secuelas se encuentra en los 40 años. Por su parte variable sexo comparada con la variable objetivo (Figura 1b). Indica que existe una cantidad similar tanto de hombres como de mujeres en la base de datos ya que el 46% son hombres y el 54% son mujeres. Sin embargo, al comparar esta

variable encontramos que el 25% de las mujeres presentaron secuelas respecto al 19% que fueron hombres.

Figura 1a. Variable edad

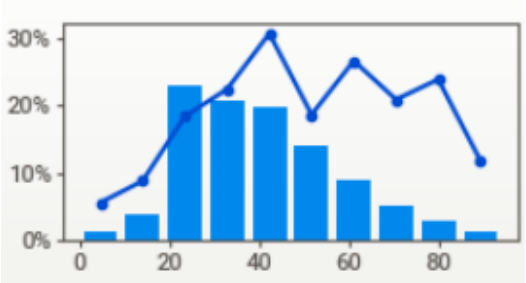
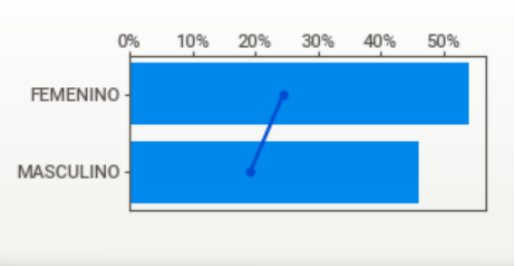


Figura 1b. Variable sexo



Teniendo en cuenta la figura 2a, el 78.67% de la población que tuvo resultado de prueba positivo para Covid-19 tuvo síntomas, lo anterior equivalente a 1132 personas, por su parte el 21.06 % de la población no tuvieron síntomas equivalentes a 307 personas, lo que quiere decir que fueron casos asintomáticos. Al compararla con la variable secuelas se encuentra que el 27% de los casos sintomáticos presentaron secuelas, por su parte el 4% de los casos asintomáticos presentaron secuelas. Por su parte al analizar la figura 2b se encontró que el 58% utilizaron medicamentos como medida de tratamiento de los síntomas presentados. El 29% de las personas que tomaron medicamentos tuvieron secuelas postcovid.

Figura 2a. Variable síntomas

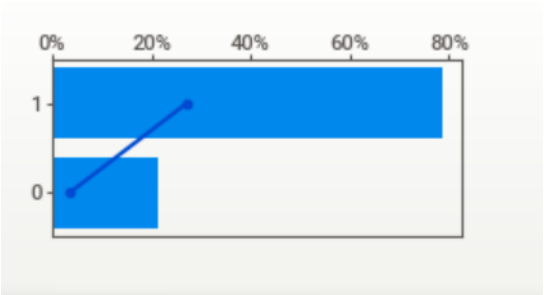
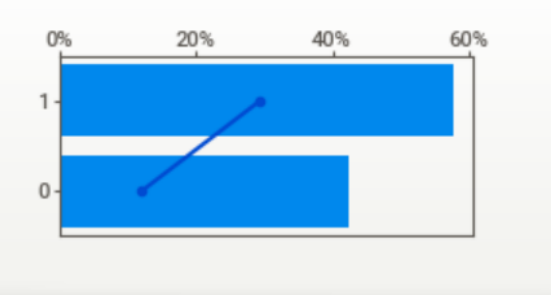
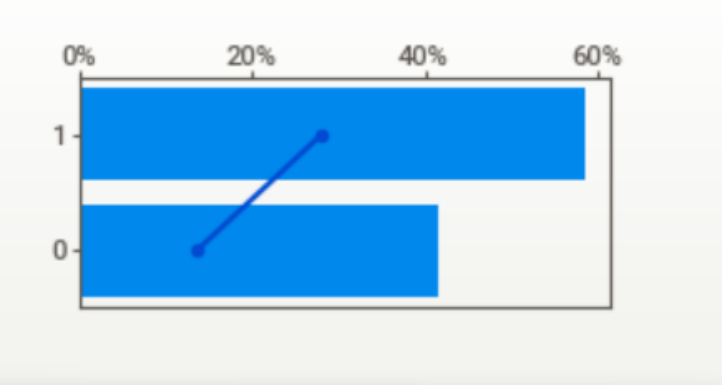


Figura 2b. Variable medicamento



En la figura 3, se observa que del total de la población que tuvo resultado positivo para Covid 19, el 58.5% equivalente a 842 personas, utilizó remedios caseros para el tratamiento de la sintomatología asociada. El 28% de la población tuvo secuelas y el 14% no tuvo secuelas postcovid

Figura 3. Variable remedio casero



Las variables que fueron seleccionadas fueron 51 y se eliminaron 15 variables para proceder a la modelación.

Desempeño y comparación de modelos

A continuación, en la tabla 1 se muestran los resultados organizados por nivel de desempeño más alto en AUC. Se puede observar que el mejor modelo clasificado por AUC es Extra Trees Classifier (et).

Tabla 1. Desempeño de los mejores modelos según los datos de prueba.

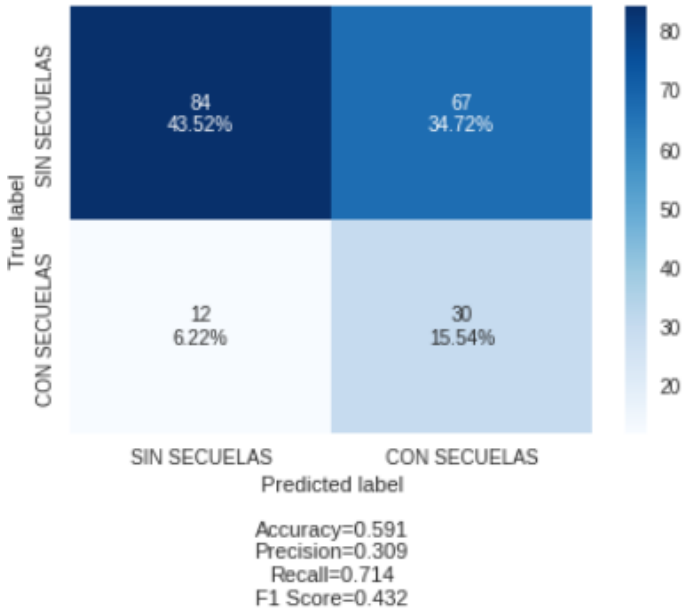
No	Modelo	Accuracy	AUC	Sensibilidad	Precisión	F1
1	Extra Trees Classifier	0.7881	0.7336	0.1129	0.5552	0.1758
2	Logistic Regression	0.7686	0.7052	0.2191	0.4478	0.2912
3	Linear Discriminant Analysis	0.7634	0.6923	0.2956	0.4574	0.3528
4	Random Forest Classifier	0.7725	0.6828	0.0176	0.0900	0.0287
5	Naive Bayes	0.3031	0.6552	0.9287	0.2307	0.3695
6	K Neighbors Classifier	0.7868	0.5895	0.1364	0.5850	0.2170
7	Light Gradient Boosting Machine	0.7738	0.5627	0.0592	0.4250	0.0994
8	Extreme Gradient Boosting	0.7361	0.5336	0.1239	0.3476	0.1225
9	Decision Tree Classifier	0.7361	0.5309	0.1651	0.3821	0.1692
10	Ada Boost Classifier	0.7439	0.5173	0.1415	0.4981	0.1546

Teniendo en cuenta que el modelo con AUC más alto es Extra Tress Classifier, se regulariza el modelo para evitar el sobreajuste, en este caso se configuró el modelo de árboles de clasificación con un máximo de10 nodos terminales. Luego se construyó una tabla de valores (Score Card) dividida en 10 deciles por nivel de probabilidad de tener secuelas postcovid, lo anterior con el fin de identificar el umbral de población que se debe gestionar para mejorar las métricas de sensibilidad y precisión. Lo anterior se realizó sobre las predicciones del modelo en la base de datos de testeo. Teniendo en cuenta el propósito del estudio en donde se requiere una métrica de sensibilidad aceptable, se encuentra lo siguiente: El umbral de 0,22 que refiere una probabilidad entre el 22% y el 100% de tener secuelas postcovid, es decir en este caso se gestiona los últimos 5 deciles, es decir el 50% de la población (Tabla 2). A continuación, en la figura 4 se muestra la matriz de confusión de las predicciones realizadas por el modelo ajustado al umbral de 0,22 y también sus métricas de precisión, sensibilidad, exactitud y f1.

Tabla 2. Score card

Decil	Probabilidad	Y0	TasaAcumY0	Y1	TasaAcumY1
1	(-1.0, 0.09218]	18	1.000.000	0	1.000.000
2	(0.09218, 0.1284]	16	0.880795	0	1.000.000
3	(0.1284, 0.16514]	21	0.774834	2	1.000.000
4	(0.16514, 0.18934]	10	0.635762	4	0.952381
5	(0.18934, 0.2178]	16	0.569536	6	0.857143
6	(0.2178, 0.23818]	17	0.463576	4	0.714286
7	(0.23818, 0.2612]	15	0.350993	4	0.619048
8	(0.2612, 0.29416]	14	0.251656	8	0.523810
9	(0.29416, 0.33664]	11	0.158940	7	0.333333
10	(0.33664, 1.0]	13	0.086093	7	0.166667

Figura 4. Matriz de confusión



DISCUSIÓN DE RESULTADOS

Teniendo en cuenta los resultados anteriormente mostrados se puede observar la herramienta de machine learning es útil en el pronóstico de variables asociadas a Covid 19 (Vahdat, 2020; Andariesta, Wasesa, 2022; Fidan, Yuksel, 2022; Boussen, S, 2021). En este caso el modelo que mejor desempeño obtuvo en este estudio para pronosticar la variable de interés "secuelas post covid" es "Extra Tres Classifier" ya que obtiene el AUC de 0.7336, el más alto sobre todos los modelos analizados. Teniendo en cuenta el fenómeno estudiado, se considera que es mejor identificar correctamente a las personas que efectivamente van a tener secuelas con respecto a las personas que no van a tener secuelas pero que el modelo predice que si van a tener secuelas. Sin embargo, esto aumentaría los costos de gestión de la población analizada.

Si se desea obtener una métrica de sensibilidad aún más alta, según la tabla de probabilidades construida en este estudio, se tendría que gestionar una mayor cantidad de personas, lo que implica un coste mayor para la entidad de salud que realicé esta tarea, por ejemplo, si se quisiera tener una sensibilidad superior al 90% se tendría que gestionar el 70% de la población. Por lo tanto se propone para la gestión de la población, una priorización de la población teniendo en cuenta inicialmente el decil con mayor probabilidad hasta llegar al decil con menor probabilidad, para atender inicialmente a las personas con mayor probabilidad de tener secuelas postcovid.

CONCLUSIONES

Teniendo en cuenta el objetivo general del presente estudio, el cual fue la construcción de un modelo machine learning que pronostique personas con secuelas poscovid, se encuentra

que el modelo más efectivo según la base de datos utilizada es árboles de clasificación con una puntuación AUC de 0.73.

Por otra parte, al analizar la tabla de puntuaciones del modelo con respecto a los deciles, puede concluir lo siguiente, si se prioriza el 50% de la población, se tendrá una efectividad del 71% de las personas que tendrán secuelas, sin embargo, en estos deciles se gestionaría también el 46% de personas que realmente no tendrán secuelas. Se concluye que este modelo es significativo para pronosticar secuelas postcovid siempre y cuando se realiza priorización de la población con base a deciles de probabilidad.

Finalmente se sugiere tener en cuenta para estudios similares, tratar de que la tasa del evento sea superior al 30%, lo anterior para lograr un balance en los datos y por lo tanto obtener mejores métricas de desempeño en modelos más avanzados. Se sugiere en el momento de la modelación, utilizar herramientas que permitan una correcta selección de las variables a intervenir en la construcción de los mismos, lo anterior para generar un mayor nivel de predicción del modelo, contribuyendo al pronóstico e implementación de acciones necesarias para el mejoramiento de su calidad de vida de los pacientes.

REFERENCIAS BIBLIOGRÁFICAS

Andariesta D, Wasesa, M (2022) Machine learning models for predicting international tourist arrivals in Indonesia during the COVID-19 pandemic: a multisource Internet data approach. J Tourism Futures

Abdel, B. (2021) FSS-2019-nCov: a deep learning architecture for semi-supervised few-shot segmentation of COVID-19 infection. Knowl-Based Syst 212:106647

Alyasseri, Z (2021) Review on COVID-19 diagnosis models based on machine learning and deep learning approaches. Expert Syst e12759.

Arash, H., Nima, N., Mehmet, U & Shiva, T. (2022) Machine learning applications for Covid 19 outbreak management. Neural Computing and Applications (2022) 34:15313 – 15348.

Afshar, P (2021) COVID-CT-MD, COVID-19 computed tomography scan dataset applicable in machine learning and deep learning. Scientific Data 8(1):1–8

Boussen, S. (2021) Triage and monitoring of COVID-19 patients in intensive care using unsupervised machine learning. Comput Biol Med 142:105192

Coureaux, L; Cuevas, M. (2021) Relación causa - efecto entre manifestaciones bucales y pacientes con la COVID-19 MEDISAN, vol. 25, núm. 5, 2021, Septiembre-Octubre, pp. 1216-1226. Centro Provincial de Ciencias Médicas.

De Siqueira Santos, S (2022) Machine learning and network medicine approaches

Dong, C (2021) Non-contact screening system based for COVID-19 on XGBoost and logistic regression. Comput Biol Medicine 141:105003

Géron, A. (2019). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems. O'Reilly Media, Inc.

Fidan H, Yuksel M (2022) A comparative study for determining Covid-19 risk levels by unsupervised machine learning methods. Expert Syst Appl 190:116243

Indumathi, N (2022) Prediction of COVID-19 outbreak with current substantiation using machine learning algorithms. In: Intelligent interactive multimedia systems for e-healthcare applications, Springer pp. 171–190

Masum, M (2022) Comparative study of a mathematical epidemic model, statistical modeling, and deep learning for COVID-19 forecasting and management. Socio-Econ Plann Sci 80:101249

Organización Mundial de la salud (2020) Los nombres de la enfermedad por coronavirus (COVID-19) y del virus que la causa. Disponible en:https://www.who.int/es/emergencies/diseases/novel-coronavirus2019/technical-guidance/naming-the-coronavirus-disease-(covid-2019)-and-thevirus-that-causes-it

Parra Sanabria EA, Bermúdez Bermúdez M, Peña Vega CP, Rueda Jiménez A. (2020) Manifestaciones orales y maxilofaciales asociadas a la COVID-19.

Rahman, M; Paul, K; Hossain, M; Ali, G; Rahman, M & Thill, J (2021) Machine learning on the COVID-19 pandemic, puman mobility and air quality: a review. In IEEE Access 9:72420–72450. https://doi.org/10.1109/ACCESS.2021. 3079121 79.

Shah V et al (2021) Diagnosis of COVID-19 using CT scan images and deep learning techniques. Emerg Radiol 28(3):497–505

Vahdat S (2020) The role of IT-based technologies on the management of human resources in the COVID-19 era. Kybernetes