International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Integrated Machine Learning Approaches for Early Infectious Disease Diagnosis

Sweet Mercy Pacolor [*,a] , Thelma Palaoag[b]

Submitted: 05/02/2024 **Revised**: 13/03/2024 **Accepted**: 19/03/2024

Abstract: In the reality of global health issues, infectious diseases are complex and challenging, requiring innovative approaches for accurate and timely diagnosis. The primary goal of this research is explore a machine learning algorithm and to determine the key features and biomarkers that may be used in the early detection of infectious diseases. The research design for this study was qualitative data. The main data source was gathered through an open-ended questionnaire for thematic analysis. A systematic search was conducted in the references for the use of machine learning to infectious diseases. Healthcare professionals make representation for the Focus Group Discussion (FGD). MAXQDA software was used to examine the FGD data using context analysis. Primary and secondary data collection are also part of the study. Following infection control policies is considered one of the best practices of healthcare facilities to alleviate the impact of infectious diseases. Key features and biomarkers of infectious diseases help the healthcare professional make a diagnosis earlier and prevent severe complications. The identified key features and biomarkers depend on the stage of infection, symptoms, and diagnosis. The findings of this research study could lead to more improvements in public health plans and advancement in the management of infectious diseases. ML algorithms include eXtreme Gradient Boosting (XGBoost), random forest (RF), long short-term memory (LSTM), support vector machine (SVM), and convolutional neutral network (CNN), improves accuracy in diagnosing infectious diseases. Continuing research studies can be used to develop more hybrid prediction models.

Keywords: *Key feature, biomarkers, machine learning algorithms, healthcare professional, MAXQDA*

Introduction

A harmful microorganism such as virus, bacteria, parasite, or fungi can cause an infectious disease. The possible danger of fast spread and severe implications, continue to create difficult challenges to public health. Effective management, treatment and control of epidemics depend on the timely and accurate diagnosis of such infections. Traditional methods encounter limitations related to speed, sensitivity, and scalability, emphasizing the need for innovative solutions. In recent years, combining machine learning (ML) techniques with detection of infectious disease diagnosis has become more possible and promising for these challenges.

Despite significant advancements in treatment and prevention over the last few decades, infectious diseases continue to be the leading cause of mortality and morbidity, worsening the living condition of countless millions of people worldwide. Often posing a challenge to a physician's diagnostic skill, infections must be considered while making differential diagnosis of disorders involving all organ systems. (Kasper, et al., 2005)

One of the problems observed by the researchers is the collection and reporting of data, which hinders the ability to track diseases. Additionally, there are gaps in healthcare infrastructure such as hospital, clinics, and laboratory facilities. There is also lack of public awareness and education about infectious diseases, coupled with limited resources including financial, human, and medical supplies. Identifying the different key features and biomarker of infectious diseases can lead to timely interventions and treatments, potentially reducing the spread and impact of these infectious diseases.

According to the Regional Unified Health Research Agenda for Region 8 (RUHRA), the data from the DOH Regional Office VIII reports that acute upper respiratory tract infections account for more than half of all the causes of morbidity in the region. Follow next by hypertensive cardiovascular disease (13%), pneumonia (9%), and pulmonary disorders (8%). Pneumonia comes in second with 21% of all fatalities, while hypertensive cardiovascular disease results for 23% of all deaths. Trauma/injuries and accidents come in third, making up 13% of the total. (https://region8.healthresearch.ph/index.php/ruhra, 2024). The data shows how cases morbidity and mortality are distributed throughout Eastern Visayas, Philippines. The findings shows that the total morbidity of the region is greatly affected by respiratory infections, particularly acute upper respiratory tract infection. Furthermore, the two top causes of deaths are hypertensive cardiovascular disease and pneumonia, suggesting the need of public

[a]: Samar State University, Philippines. [b] University of the Cordilleras, Philippines
E-mail: sweetmercy.pacolor@ssu.edu.ph.*

health interventions and preventive actions to address these health concerns.

The aim of this study is to identify the key features and biomarkers that can be used in the early detection of infectious disease as well as discover machine learning algorithms. Machine learning (ML), enable prediction of data through analysis. These three types of machine learning are the supervised, unsupervised, and reinforcement learning. In supervised learning, the model is trained to predict the dependent variable and achieve the accuracy using given data. Unsupervised learning groups the data using the training set and finds patterns in the not label data. The reinforcement learning method also teaches a machine to make accurate decisions by trial and error.

This research may explore the potential for adopting machine learning methods. Considering accurate diagnosis enables immediate action, timely diagnosis is essential in reducing the impact of infectious diseases. The use of machine learning algorithms allows a chance to improve the process of early detection and enhance the ability to address global outbreaks of infectious diseases. It is essential to comprehend and prevent to the transmission of infectious diseases.

Methodology

The research design for this study was qualitative data. The main data source was gathered through an open-ended questionnaire for thematic analysis. The questions asked were thematic with key features and biomarkers of infectious disease. A systematic review was utilized. A systematic search was conducted in the references for the use of machine learning to infectious diseases. By conducting a systematic review, researchers aim to identify highly efficient and accurate methods necessary used for parameters models. By using machine learning techniques allows software applications to make more precise predictions. Through the systematic review, researchers examined various articles related to machine learning algorithms, including eXtreme Gradient Boosting (XGBoost), random forest (RF), long short-term memory (LSTM), support vector machine (SVM), and convolutional neutral network (CNN), improves accuracy in diagnosing infectious diseases. In the selection of respondents, random sampling was utilized by the researcher. A total of eight respondents were healthcare professionals.

The healthcare professionals make representation for the Focus Group Discussion (FGD). The focus group discussion was done in one setting. The questions asked in the focus group discussion were thematic along with key features and biomarkers of infectious disease. Using MAXQDA software, context analysis was used to examine the FGD data. It made use of the word frequency query, which examined the most often used terms in the discussion using cluster analysis and a word cloud tab. Words that appear often are displayed in larger font sizes in the word cloud analysis. Primary and secondary data collection are also part of the study. By interviewing healthcare professionals, primary data was gathered. Secondary data was collected from published research on infectious diseases and machine learning.

The confidentiality of their responses was guaranteed to the research respondents. In addition, the respondents also had the option to deny participation in the study or to voluntarily accept after being informed about the processing of the data and the use of computers. In the data-collecting phase, all respondents – 100% of them answered the open-ended questions via Google Forms. The gathered data was organized, examined, and interpreted using MAXQDA software in a thematic analysis.

Results and discussion

When harmful substance enters your body, you might develop infectious diseases. Bacteria, fungi viruses and parasites are the most likely culprits. It is transmitted from person to person through dirty food, drink, or insect bites. Table 1 lists the many infectious diseases; it also includes the different key features and biomarkers for specific infectious diseases. The key feature refers to the signs and distinctive symptoms that are associated with specific infectious diseases. The biomarkers provide important data on the condition of the disease and how it responds to treatment, and they are crucial in the diagnosis and detection of infectious diseases.

Infectious Disease	Key Features	Biomarker

COVID-19	• Respiratory Symptoms
	  • Difficulty Breathing
	  • Cough
	  • Shortness Breath
	• Fatigue
	• Fever
	• Sore Throat
	• Severe Respiratory Distress
	• Muscle or Body Aches
		• SARS-CoV-2 RNA (PCR Test)
		  • RT-PCR test
		• D-Dimer

	• Gastrointestinal Symptoms • Loss of Smell and Taste	
Dengue Fever	• Skin Rash (Petechiae) • Low Platelet Count • Sudden Onset Fever • Joint and Muscle Pain • Severe Headache • Mild Bleeding • Fatigue • Nausea and Vomiting • Pain Behind the Eyes • Dengue Hemorrhagic Fever (DHF)	• Detection of Dengue Virus RNA (PCR Test) • Platelet Count • Hematocrit (HCT) Levels • Activated Partial Thromboplastin Time (aPTT)
Filariasis	• Chills • Headaches • Skin Lesions • Severe Lymphedema of Limbs (Elephantiasis) • Fever	• Lymphatic Filariasis Serum (+) • Malayi Microfilariae (MF) culture

HIV/AIDS	• Asymptomatic Stage (Acute HIV Infection)
	  • Flu-Like Symptoms
	  • Swollen Lymph Node
	  • Rash
	• Weight Loss
	• Diarrhea
	• Cough
	• If not treated
	  • Tuberculosis
	  • Cryptococcal Meningitis
	  • Severe Bacterial Infection
	  • Hepatitis B, C
		• HIV Antibodies
		• HIV Antigen
		• HIV Nucleic Acid Testing (NAT)

Influenza	• Sudden Onset of Symptoms • Respiratory Symptoms • Rhinitis • Cough • Muscle or Body Pain	• RT-PCR • Molecular assay • Nucleic Acid Amplification Test
Leprosy	• Skin Lesions • Loss of Eyebrow • Loss of Eyelashes • Discolored Skin • Deformation of Fingers and Feet • Painless ulcer on the sole of feet	• Skin Smear Examination • Skin Biopsy • Nerve Biopsy
Malaria	• Flu-Like Symptoms • Extreme Fatigue • Difficulty of Breathing • Dark or Bloody Urine • Jaundice (Yellowish color of eyes and skin)	• Blood Smear Examination • PCR • IFH or ELISA • CBC - MP (+)
Rabies	• Headache • Anxiety • Flu-Like Symptoms	• Direct Detection of Rabies Virus • Serum and Spinal Fluid Tested for Rabies Virus

	• Agitation • Weakness, Discomfort, Prickling, or Itching Sensation at the site of the bite • Cerebral Disfunction • Confusion	
Schistosomiasis	• Fever • Cough • Muscle Pain • Blood in Stool or Urine • Chills • Abdominal Pain	• Detection of Parasite Egg in Stool or Urine Specimen
Tuberculosis	• Prolong Cough • Night Sweats • Fatigue • Weight Loss • Fever • Weakness	• Chest X-Ray • Sputum AFB • Tuberculin Skin Test (Mantoux Test)

Table 1. Key Features and Biomarkers of Infectious Disease

The table 1 shows the infectious diseases with their corresponding key features and biomarkers. These provide fast intervention, correct diagnosis, and treatment of infectious diseases. COVID-19 differ from minor respiratory signs to severe respiratory distress. SARS-CoV-2 RNA and D-Dimer levels are biomarkers of COVID-19 patient. To improve patient care and stop the disease from spreading, it is important to comprehend how key features and biomarkers interact. Dengue fever is another infectious disease that is prevalent in tropical areas and is brought on by the dengue virus, which is spread by mosquitoes. Some of the key features include skin rash, joint and muscle pain, low platelet count, dengue hemorrhagic fever and sudden onset fever. Understanding this key feature prevents complications. Several biomarkers like the detection of dengue virus RNA, platelet count, hematocrit level, and activated partial thromboplastin time. When diagnosed early and treated properly, dengue fever cases can lead to better patient outcomes.

Figure 1. Word cloud on the key feature of infectious disease.

Figure 1 shows the analysis of the word cloud on the key feature of infectious disease as perceived by healthcare professional respondents. The terms that are most popular, as shown by the word cloud analysis are symptom, fever, cough, loss, and pain. As seen in Figure 1 symptom is important key feature of infectious disease. Symptom are what a person experiences, they are indicators of sickness. These signs show in many conditions and illnesses. It is essential to learn about the ways in which the body communicates health issues and to seek solutions. Fever and cough are the most common key feature of the identified infectious disease in the

region. Fever is a key feature of Covid-19, filariasis, schistosomiasis and tuberculosis, while for dengue fever, it presents as sudden onset fever, and dengue haemorrhagic fever is considered by severe symptoms. Loss of smell and taste is a key feature of Covid-19, while weight loss is related with HIV/AIDS and tuberculosis, and loss of eyebrow and loss of eyelashes is trait of leprosy. Muscle or body pain is an important feature for Covid-19, dengue fever, influenza, and schistosomiasis, while pain behind the eyes is specific to dengue fever, and abdominal pain is a symptom of schistosomiasis. Furthermore, it is important to understand the essentials of key features of each disease.

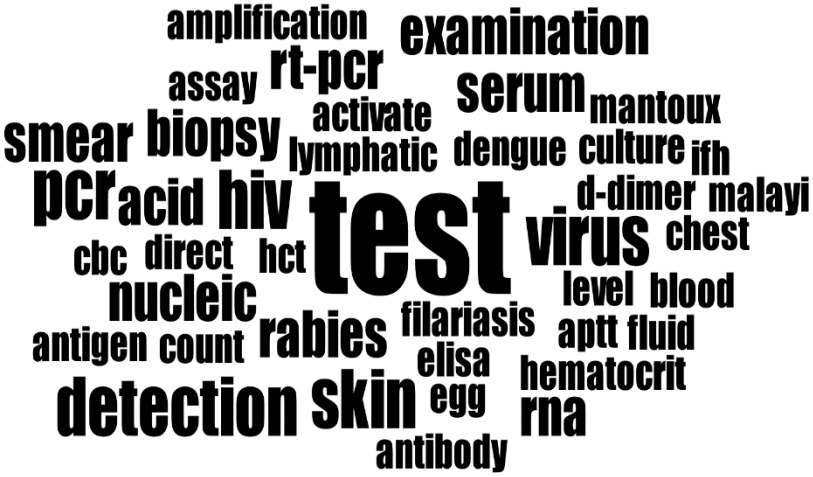

Figure 2. Word cloud on the biomarkers of infectious disease.

Figure 2 shows the analysis of the word cloud on the biomarkers of infectious disease as perceived by the healthcare professional respondents. The most often used terms as shown in the word cloud analysis are test, detection, HIV, PCR, and skin. Testing may involve immunologic tests, blood tests, oral swabs, urine tests, and microscopic examinations. These tests are important for accurately diagnosing infections and determining the underlying cause of a patient's symptoms. This accuracy is critical in distinguishing between similar diseases with different treatment approaches. Detection, HIV, PCR, and skin are the most common biomarkers of the identified infection disease in the region. Detection of dengue virus RNA is a biomarker of dengue fever; direct direction of rabies virus indicates rabies and detection of parasite egg in stool serves as a biomarker for schistosomiasis. HIV antibodies, HIV antigen and HIV nucleic acid testing are biomarkers of HIV/AIDS, while PCR test are used for Covid-19, dengue fever, influenza, and malaria. The skin smear examination and skin biopsy are utilized for diagnosing leprosy, and the tuberculin skin test is used for tuberculosis. Moreover, the use of biomarkers enables an improved understanding of disease caused, as well as how the body responds to medical interventions and fights against diseases.

CNN	XGBoost	SVM	LSTM	RF	ARIMA	LASSO	Other
Pneumonia	Hepatitis B	Influenza	COVID-19	Dengue Fever	Dengue Fever	Chickenpox	COVID-19 (KNN)
COVID-19	COVID-19	COVID-19	Dengue Fever	COVID-19	COVID-19	Dengue Fever	Dengue Fever (PSO-ANN)
Dengue Fever	Q-Fever	Hepatitis E	Hepatitis B	Zika	Hepatitis E	Malaria	Crimean-Congo Hemorrhagic Fever (Gaussian Process)
Hepatitis B	Hand, Foot, and Mouth Disease	-	Hepatitis E	Hand, Foot, and Mouth Disease	-	Hand, Foot, and Mouth Disease	
	-	-	Malaria	-	-	-	-

Table 2. Machine learning techniques for infectious disease

The machine learning techniques for infectious disease are shown in Table 2. It indicates that an infectious disease can have various ML models for diagnosis. Selecting the most effective machine learning model can be difficult, particularly when utilizing related diseases. For example, different research has focused on identifying pneumonia and COVID-19. There is a connection between these two diseases. Therefore, using machine learning techniques, a

single model that can diagnose COVID-19 and pneumonia might be developed. Patients with both diseases may benefit significantly from this approach.

Convolutional neutral network (CNN) are models that use image processing techniques and obtain significant accuracy. Using CNN, the accuracy percentage for Pneumonia is 97.92%. CNN was also used in the process to recognize COVID-19 from chest x-ray images. eXtreme gradient boosting (XGBoost) improves performance compared to other existing ML method by using clinical and demographic data. It performed better for all diseases, including Q-Fever and Hepatitis B, with a more accurate diagnosis, and had 90% accuracy rate for COVID-19. Support vector machine (SVM) predictions regarding influenza achieved an average accuracy of 86.7% using news text data. When it comes to mapping the COVID-19, the SVM model likewise demonstrates high forecast accuracy. It also examined using polynomial regression, and ML algorithm based on a geographic information system (GIS). When it comes to COVID-19 case prediction, long short-term memory (LSTM) exhibits high accuracy. However, it performs medium-low accuracy when it comes to long-term prediction and estimating the actual evidence of COVID-19 infection. Furthermore, this is the best machine learning model for predicting hepatitis E. The random forest (RF) approach is more accurate than other models when it comes to handling inputs such as demographic, temporal, meteorological, and dengue monitoring data. Additionally, it demonstrated the COVID-19 result with highest accuracy as well as an efficient model for both disease prediction and actual evidence of COVID-19 infection. The autoregressive integrated moving average (ARIMA) is the model that does the best for long-term predictions for predicting an outbreak of dengue fever. The ARIMA model was used to evaluate COVID-19 in pattern of virus infestation and empirical indicators of infection. Utilizing the least absolute shrinkage and selection operator (LASSO) model for prediction, short-term forecasts typically perform longer-term predictions. Using this method, it would be possible to predict hand, foot, and mouth disease (HFMD), chicken pox, malaria, dengue fever.

The SVM and ARIMA models were applied to the COVID-19 forecast and observations for widely patterns. RF. ARIMA, and LSTM analysis of the COVID-19 was used to determine the empirical indication of the infection. Forecasting techniques for recurring infectious disease outbreaks might be greatly improved by LASSO when combined with LSTM to provide a more accurate dengue fever prediction model. For forecasting hepatitis E, LSTM performed the best and is the most appropriate model. To analyze and forecast hand, foot, and mouth disease, two machine learning algorithm were used: RF and XGBoost. The LSTM model demonstrated an average accuracy for prediction of 87.3% when applied successfully to the prediction of malaria. With the use of an RF model, it was possible to forecast human infection with the zika virus, which spread by mosquito bites.

Conclusion

The study highlights the key features and biomarkers of infectious disease based on thematic analysis. The identified key features and biomarkers of infectious diseases give knowledge and understanding in assessing the different types of infectious diseases, especially to the community. It is important in every healthcare facility to have infectious control programs, policies, and guidelines. Following infection control policies is considered one of the best practices of healthcare facilities to alleviate the impact of infectious diseases. Healthcare professionals and healthcare facilities serve an essential part in preventing the spread of infectious diseases within communities. Awareness of the infection control policies helps guarantee the welfare of patients, healthcare professionals, and the community. A primary aspect of any ML method is the data. Many studies indicate that their developed models require additional validation thus, addressing the issue can be achieved through the process of data standardization and normalization. Using ML algorithms for early epidemic detection is a method that can effectively monitor the outbreak of major infectious diseases. Early detection of infectious diseases with machine learning algorithms is an innovative approach, and integrating these models improves accuracy. Utilizing ML models makes forecasting future trends in disease rate and spread over time within specific geographic areas. ML is a useful tool that has made significant innovations in its applications, including the ability to address real-world healthcare issues. The findings of this research study could lead to more improvements in public health plans and advancement in the management of infectious diseases. ML algorithms include eXtreme Gradient Boosting (XGBoost), random forest (RF), long short-term memory (LSTM), support vector machine (SVM), and convolutional neutral network (CNN), improve accuracy in diagnosing infectious diseases. A combination of multiple ML algorithms could be excellent for handling high-dimensional data. Continuing research studies can be used to develop more hybrid prediction models.

References

[1] Absar, N., Uddin, N., Khandaker, M. U., & Ullah, H. (2022). The efficacy of deep learing baed LSTM model in forecasting the outbreak of contagious diseases . Infectious Disease Modelling, 170-183.

[2] Ahsan, M., Luna, S., & Siddique, Z. (2022). Machine-learning-based disease diagnosis: a comprehensive review. Healthcare, 1-30.

[3] Alballa, N., & Al-turaiki, I. (2021). Machine learning approaches in COVID-19 diagnosis, mortality, and severity risk prediction: a review. Informatics in Medicine Unlocked, vol. 24, article 100564.

[4] Alqaissi, E. Y., Alotaibi, F. S., & Ramzan, M. S. (2022). Modern machine-learning predictive models for diagnosing infectious diseases. Computational and Mathematical Methods in Medicine, 1-13.

[5] Benedum, C. M., Shea, K. M., Jenkins, H. E., Kim, L. Y., & Markuzon, N. (2020). Weekly dengue forecasts in Iquitos, Peru; San Juan Puerto Rico; and Singapore. PLoS Neglected Tropical Diseases.

[6] Bhavsar, K. A., Abugabah, A., Singla, J., AlZubi, A. A., Bashir, A. K., & Nikita. (2021). A comprehensive review on medical diagnosis using machine learning. Computers, Materials and Continua, vol. 67, no. 2, 1997-2014.

[7] Borkenhagen, L. K., Allen, M. W., & Runstadler , J. A. (2021). Influenza virus genotype to phenotype predictions through machine learning: a systematic review. Emerging Microbes and Infections, vol. 10, no. 1, 1896-1907.

[8] Boyles, T., Stadelman, J., & Ellis, P. (2021). The diagnosis of tuberculosis meningitis in adults and adolescents: protocol for a systematic and individual patient data meta-analysis to inform a multivariable prediction model. Well-come Open Research, vol. 4.

[9] Chen, Y., Chu, C. W., Chen, M., & Cook, A. R. (2018). The utility of LASSO-based models for real time forecasts of endemic infectious diseases: A cross country comparison. Journal of Biomedical Informatics vol. 81, 16-30.

[10] Elaziz, M. A., Hosny , K. M., Salah, A., Darwish, M. M., & Lu, S. (2020). New machine learning method for image based diagnosis of COVID-19. PLoS One, vol. 15, no. 6, 1-18.

[11] Gambhir, S., Malik, S. K., & Kumar, Y. (2017). PSO-ANN based diagnostic model for the early detecion of dengue disease. New Horizons in Translational Medicine, vol. 4, 1-8.

[12] Harris, M., Qi, A., & Jeagal, L. (2019). A systematic review of the diagnostic accuracy of artificial intelligence-based computer programs to analyze chest x-rays for pulmonary tuberculosis. PLoS One, vol. 14, no.9, 1-19.

[13] https://region8.healthresearch.ph/index.php/ruhra. (2024, January 14). Retrieved from https://region8.healthresearch.ph/: https://region8.healthresearch.ph/index.php/ruhra

[14] Kar, P., & Karna, R. (202). A review of the diagnosis and management of hepatitis E. Current Treatment Options in Infectious Diseases, vol. 12 no. 3, 310-320.

[15] Kasper, D. L., Fauci, A. S., Longo, D. L., Braunwald, E., Hauser, S. L., & Jameson, L. J. (2005). 16th Edition Harrison's Principles of Internal Medicine. United States of America: The McGraw-Hill Companies, Inc.

[16] Marquez, E., & Barron, V. (2019). Artificial intelligence system to support the clinical decision for influenza. IEEE Int. Autumn Meeting on Power, Electronics and Computing (ROPEC), Ixtapa Mexico, 1-5.

[17] Meraj, S. S., Yaakob, R., Azman, A., Rum, S. M., & Nazri, A. A. (2019). Artificial intelligence in diagnosing tuberculosis: a review. Internationla Journal on Advanced Science, Engineering and Information Technology, vol. 9, no. 1, 81-91.

[18] Nahid, A. A., Sikder, N., Bairagi, A. K., Razzaque, M. A., & Masud, M. (2020). A novel method to identify pneumonia through analyzing chest radiographs employing a multichannel convolutional neutral network. Sensors, vol. 20, no. 12, 1-18.

[19] Santangelo, O., Gentile, V., Pizzo, S., Giordano, D., & Cedrone, F. (2023). Machine learning and prediction of infectious diseases: a systematic review. Machine learning and knowledge extraction, 175-198.

[20] Stokes, K., Castaldo, R., & Federici, C. (n.d.). The use of artificial intelligence systems in diagnosis of pneumonia via signs and symptoms: a systematic review. Biomedical Signal Processing and Control, vol. 72 article 103325, 2022.

[21] Tian, X., Chong, Y., Huang, Y., Guo, P., & Li, M. (2019). Using machine learning algorithms to predict hepatitis B surface antigen seroclearance . Computationla and Mathemaical Methods in Medicine, vol. 2019, 1-7.

[22] Tran, N. K., Albahra, S., & May, L. (2022). Evolving applications of artificial intelligence and machine learning in infectious diseases testing. Clinical Chemistry, vol. 68, 125-133.

[23] Wynants, L., Van Calster, B., & Collins, G. S. (2020). Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. bmj, vol. 369.

[24] Yan, L., Zhang, H. T., Goncalves, J., Xiao, Y., & Wang, M. (2020). Ann interpretable mortality prediction model for COVID-19 patients. Nature Machine Intelligence, vol. 2, no. 5, 283-288.