2022 IEEE 7th International conference for Convergence in Technology (I2CT)
Pune, India. Apr 07-09, 2022

2022 IEEE 7th International conference for Convergence in Technology (I2CT) | 978-1-6654-2168-3/22/$31.00 ©2022 IEEE | DOI: 10.1109/I2CT54291.2022.9824470

Machine Learning based COVID-19 Mortality Prediction using Common Patient Data

Shubham Agrawal
Department of Information Technology
National Institute of Technology Karnataka
Mangalore, India
shubham050300@gmail.com

Nagamma Patil
Department of Information Technology
National Institute of Technology Karnataka
Mangalore, India
nagammapatil@nitk.edu.in

***Abstract*—COVID-19 was declared a pandemic in 2020, and it caused havoc worldwide. The fact that it is unpredictable adds to its lethality. The world has already seen various COVID-19 infection waves, subsequent waves being even more deadly. Many patients are asymptomatic initially but suddenly develop breathing problems. More than four million people have died due to COVID-19. It is necessary to forecast a patient's likelihood of dying so that appropriate precautions can be implemented. In this study, a COVID-19 mortality prediction model which uses machine learning is proposed. Most of the current research work requires several patient features and lab test results to predict mortality. However, we suggest a simpler and more efficient technique that relies solely on X-rays and basic patient information such as age and gender. Several ensemble-based models were evaluated and compared using a variety of metrics, and the best method was able to achieve a classification accuracy of 92.6% and AUPRC of 0.95.**

***Index Terms*—COVID-19, Machine learning, Ensemble learning, classification**

I. Introduction

A new kind of lung infection was first detected in December 2019. It started spreading and soon became a global pandemic [1]. The World Health Organization (WHO) termed it COVID-19 (Coronavirus Disease 2019) after determining that it was a new variant of coronavirus.

COVID-19 is highly unpredictable [2]; around 5 million people have lost their lives. New variants of COVID-19 are emerging [3], which are even more deadly. This brings in a need for a system that can potentially detect and warn about the chances of death of a patient. This can help hospital staff to focus on such patients and be proactive. Such a system will also reduce the load on hospital staff working tirelessly for COVID-19 infected patients.

Machine learning is an emerging field that can tackle this problem. Machine learning has been successfully used in the medical domain [4]. It can extract features from images that can be used for classification, annotation, etc. Existing work shows that Deep learning can help extract COVID-19 specific features from chest X-ray images [5]. Moreover, ML can be utilized to extract important information and analyse patient data. There are many studies that have tried to use this data for COVID-19 mortality prediction [6]. However, most of them rely on a lot of lab tests and other patient features that are difficult to obtain.

In this study, machine learning is used to create an efficient model that can predict the chances of mortality of patients. The primary goal of this research is to predict the patient mortality with as few features as possible and which does not depend on patient data that is either hard to get or which takes time to obtain. The proposed model requires a chest X-ray image of the patient and two other basic features like age and sex. Firstly, two scores are calculated from a model proposed by [7], which denotes lung involvement and opacity in the lungs. Then these four features are passed to a machine learning model. Various models are compared for this study. This method shows promising results that can help reduce the number of fatalities and provide a more accurate diagnosis.

There are many studies that aim to predict patient mortality. However, these models require a lot of input features [8]. This can include patient details like demographic data or lab test results. Acquiring all this information is difficult, and it takes time. In order to obtain survival prediction results of a patient early, it is critical to develop a model that can use basic features that are easy to obtain to predict mortality. In this paper, we predict patient mortality using just three things-patient's age, sex, and chest X-ray image. We also propose the use of ensemble techniques as they can combine the prediction capability of various weak models and overall produce a better model. The results obtained using ensemble techniques outperforms existing models even though a significantly less number of features are used in predicting the results.

The following are the study's main contributions:

- Selecting minimal and easy-to-get patient data for predicting mortality.
- Comparing several machine learning models which use ensemble techniques.
- Proving that mortality prediction is possible using common patient data, which can help to triage faster.

In the following paper, we first discuss the relevant work in the field of COVID-19 detection and mortality prediction using machine learning. Then section 3 provides details about the proposed approach and details of the dataset. In section 4, the experimental evaluation and its analysis is provided. Finally, the conclusion is discussed in the final section.

978-1-6654-2168-3/22/$31.00 ©2022 IEEE

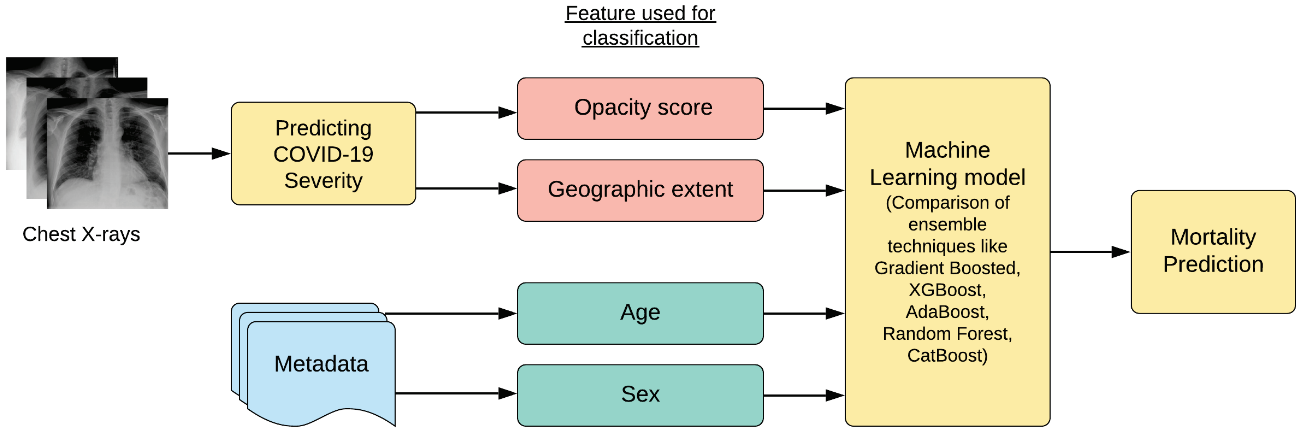

Feature used for classification
Chest X-rays
Predicting COVID-19 Severity
Opacity score
Geographic extent
Metadata
Age
Sex
Machine Learning model
(Comparison of ensemble techniques like Gradient Boosted, XGBoost, AdaBoost, Random Forest, CatBoost)
Mortality Prediction

Fig. 1. Proposed framework for mortality prediction.

II. Related Work

After COVID-19 became a pandemic [9], there has been a lot of research to detect and then classify COVID-19 using radiography chest x-ray imaging data and Machine learning. A fully automated system to identify COVID-19 using CT images was developed by S. Hu et al. [10].

Deep learning has recently brought a revolution in the medical domain. Deep learning is a sub-field of Machine learning that can model complex relationships among the different features and tries to gain knowledge from it for predictions [11] [12]. Deep learning has previously been successfully used for various tasks like speech recognition, image classification, image retrieval, and many more. Nowadays, Machine learning models are being used in almost all aspects of medical imaging [13].

Many studies have helped combat the pandemic, from analysing the impact of COVID-19 and its possible treatment to finding the presence of COVID-19 in samples of blood [14]. Several proposed models can find the presence of COVID-19 from chest X-ray images. However, with multiple variants of COVID-19 developing, there is a need to be able to detect COVID-19 patients that are in serious condition and need more assistance. For this, ML models can be utilised to predict the mortality chances in patients. Serious patients can then be given immediate treatment and focus. This is like adding an extra step with a more specialised focus on COVID-19 out of all the lung-related diseases.

Survival chances of severe COVID-19 patients was predicted by [15] based on factors like risk and demographics. More than 300 features were studied, and out of them, only three main features were identified using XGBoost. They were high sensitivity C-reactive protein (Hs-CRP), lymphocyte and lactic dehydrogenase (LDH). This proposed method was able to attain a precision of 95% and an accuracy of 90%in predicting risk of death.

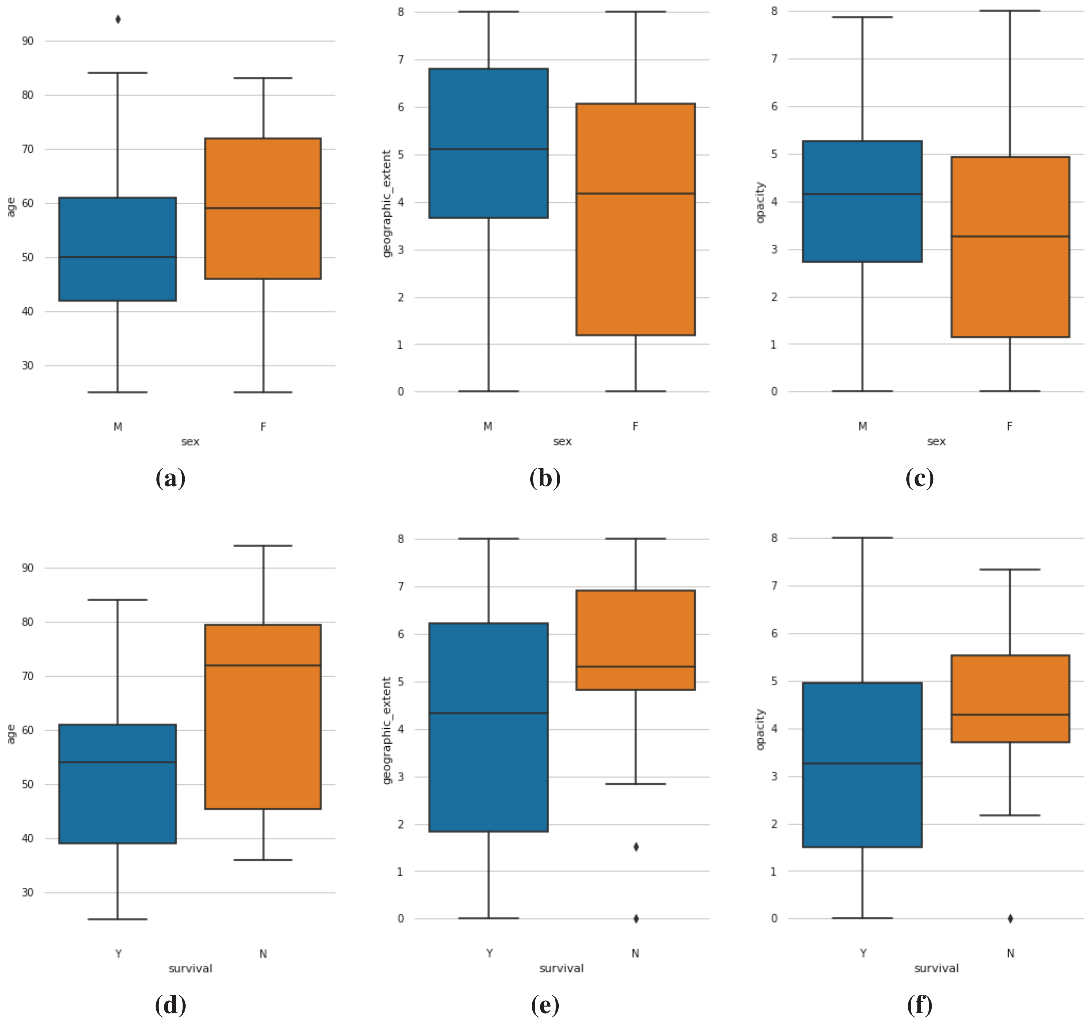
Sun et al. [16] used Support Vector Machine (SVM) to develop a model that can predict severe COVID-19 cases. The authors in this study found out the most significant laboratory and clinical features to distinguish between severe and mild cases. With an accuracy of 75.5%, the proposed model was able to predict severe patients.

Authors in [17] analysed multiple classification algorithms like Random Forest (RF), Linear Regression (LR) and Extreme Gradient Boosting (EGB). Moreover, twenty clinical features were found out to be significant for predicting critical patients. Among all the models, Random Forest [18] outperformed all other models with a 95% accuracy. However, one of the limitations in all these studies is the availability of patient features which is used to predict this risk. The features that are used in these models are not cheap to obtain and are not easily available everywhere.

J. Prada et al. [19] suggested a two-layer technique for predicting patient mortality risk. Both Convolution Neural Networks and Machine Learning were used in this study. The model used to predict the mortality risk from X-ray images is named COVID-CheXNet. Machine learning models like XGBoost, SVMs, Logistic Regression, Random Forest and Neural Nets were compared.

Experts were used to score the opacity degree and involvement of lung in the public COVID-19 dataset [7]. A model was then developed to score these factors using the developed dataset. This model was first trained on a huge non-COVID-19 dataset for better results. This uses the TorchXRayVision library, which belongs to the DenseNet model. The proposed regression model was able to predict the geographic extent on a scale of 0-8 and lung opacity on a range of 0-6.

In this work, we utilise deep learning models proposed by [7] to obtain patient features like geographic extent and lung opacity. This is then combined with basic patient features like age, sex. Among the four features used in the proposed method, only an X-ray is required as the clinical data, which is easy to obtain. It makes the prediction model easy to use and less time-consuming. Various ensemble techniques are compared in this study to propose the best one.



(a)

(b)

(c)

(d)

(e)

(f)

Fig. 2. Box plots for age, geographic extent and opacity based on sex and survival type.

III. Methodology

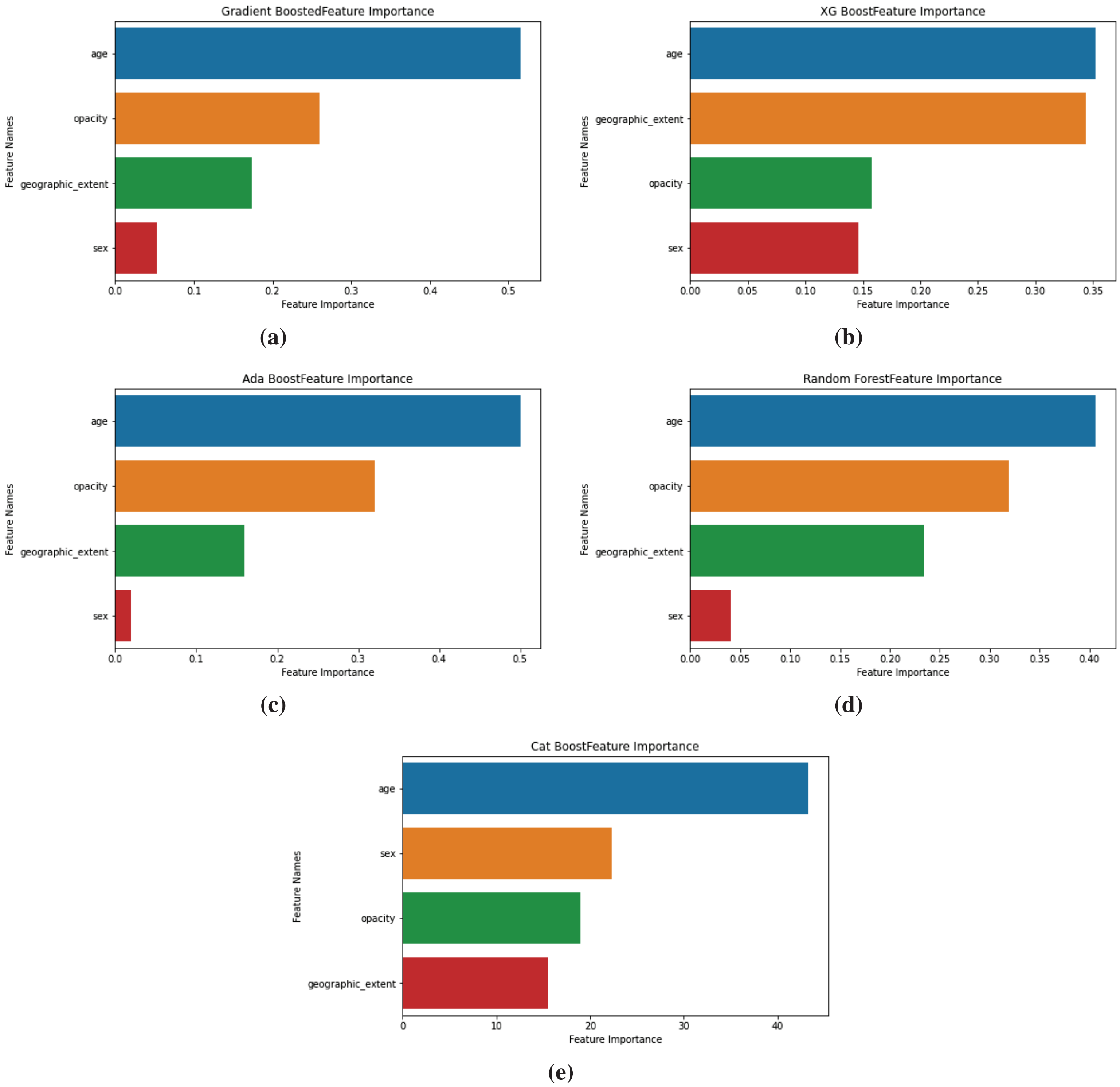
A. Dataset Details

The proposed model utilizes two kinds of data: chest X-ray images and patient meta-data, which involves age and sex. Chest X-ray images were obtained from [20]. Authors have compiled data from various open-source resources. This involves chest X-ray images for various lung diseases. The maintainer of this repository has also developed metadata that contains data related to the patients whose chest X-rays they have collected. This metadata file is used to collect other data required for the proposed method.

From this metadata, only positive COVID-19 cases were filtered. Many patients did not have their survival outcome mentioned in this data; therefore, those also had to be filtered. Finally, data containing age, sex, and survival outcome was extracted for positive cases. Then images corresponding to these cases were obtained from the dataset. A total of 136 cases were obtained, out of which 102 patients had survived, and 34 did not survive. One reason for the limited size of this data is that this dataset was developed during early stages, and not many patient's survival outcome was noted.

B. Model

Firstly, to get all the patient features, we use a chest X-ray image to obtain two main features. These are the level of opacity and lung involvement. These are calculated with the help of models as proposed by [7]. Authors here used TorchXRayVision [21] [22] library for DenseNet [23] model. This involved pre-training the model with seven X-ray datasets without any COVID-19 images. This helped the model to understand general representations and features of lungs. Then a small dataset of COVID-19 images was labelled by expert radiologists. This was used to train the model further to predict 18 common findings such as effusion, lung opacity, etc. Linear Regression was used to calculate opacity and lung involvement using a set of 18 predicted features from the model. Finally, the model predicts lung opacity with 0.78 Mean Absolute Error and Geographic Extent with a Mean Absolute Error of 1.14.



(a)

(b)

(c)

(d)

(e)

Fig. 3. Feature importance of different models.

The other two features, age and sex were retrieved from the metadata. These four features were then passed on to various machine learning models to get the most optimal model for the task of mortality prediction. Figure 1 shows all the steps of the proposed framework. To get insights about the dataset and the features used, feature importance was calculated and plotted. Different ensemble techniques [24] used in this study are:

- Gradient Boosted Classifier: This technique can combine the predictive power of different weak models and combine them to create a relatively stronger model. This internally uses decision trees.
- XGBoost (eXtreme Gradient Boosting): It is an easy to use open-source library that provides an efficient implementation of the stochastic gradient boosted algorithm.
- Ada Boost: It is one of the first types of boosting algorithm. It uses decision stumps, which are decision trees with a single split. This technique is efficient for binary classification.
- Random Forest: A group of decision trees is used in this technique. What makes this efficient is that for each tree, this algorithm can select a subset of features (either input samples/rows or input features/columns) at every tree node and thus obtain the best possible combination.
- CatBoost: It is a library that is an efficient and improved version of gradient boosting provided by Yandex. It is also known as category gradient boosting because of its

ability to select categorical input variables.

IV. Results and Analysis

For all the machine learning models, the training and testing data was split as 70:30 ratio. This means a total of 95 patient records was used for training, and 41 records were used for testing the models. The model was developed in python language using the Keras library, which uses the TensorFlow backend. Scikit-learn library was used for ensemble techniques like Gradient Boosted classifier, Ada Boost, and Random Forest. XGBoost and CatBoost have their own library.

Box plots for age, geographic extent, and opacity are shown according to sex and survival type in Figure 2. The minimum and maximum age in the dataset are 25 and 94 respectively, with characteristics of 54.49 ± 16.35. The figure suggests that the higher age group has less chances of survival as the median age of not surviving is more than 70 and the median age of surviving is less than 55. Similarly, patients with high geographic extent and opacity have less chances of survival.

Various standard metrics were used to analyse and compare various models. The classification models were compared using accuracy, precision, recall, F1-score, and area under the curve (AUC). Further, feature importance was calculated and plotted to understand the significance of every feature in predicting the output.

TABLE I
Results for different ensemble techniques.

Ensemble Model Name	**Accuracy**	**Precision**	**Recall**	**F1-score**	**AUC**
Gradient Boosted	0.83	0.84	0.83	0.83	0.92
XGBoost	0.83	0.84	0.83	0.83	0.89
Ada Boost	0.85	0.85	0.85	0.85	0.86
Random Forest	0.88	0.87	0.88	0.88	0.93
CatBoost	0.93	0.93	0.93	0.93	0.95

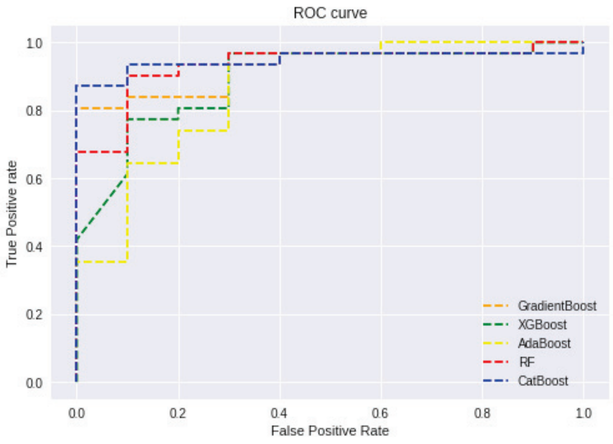


Fig. 4. Area Under the Curve (AUC) plots for different ensemble techniques.

Table I shows a comparison of all the different ensemble techniques tested in this study. It can be observed that CatBoost outperforms all the other techniques and attains a classification accuracy of 93% with 95% area under the curve. AUC of all the models are plotted in Figure 4. Finally, feature importance was calculated for each model and compared, as shown in Figure 3. In all of the models, age is clearly the most important element. However, the second most important feature differs among models.

V. Conclusion

Most of the current mortality prediction models require a large number of patient details and test results as input features. However, this is not feasible in current times as the time required to triage a patient is critical and should be as little as possible. In this study, we show by using simple and easy-to-get patient features also, we can predict patient mortality. Further, we propose the use of ensemble techniques as that can improve the classification power of various weak classifiers and prove a better overall result. Various metrics were used to compare different ensemble techniques, and CatBoost was found out to give the best result with a classification accuracy of 93% and 95% area under the curve. As expected, age was found out to be the major contributing factor to predict the survival chances of a patient.

We plan to test the model on a larger dataset in future research. As metadata like patient's age, sex is a limiting factor while obtaining the dataset, it is crucial to note such features while developing a COVID-19 related dataset.

References

[1] A. Spinelli and G. Pellino, "Covid-19 pandemic: perspectives on an unfolding crisis," *Journal of British Surgery*, vol. 107, no. 7, pp. 785–787, 2020.

[2] H. Seligmann, N. Vuillerme, and J. Demongeot, "Unpredictable, counter-intuitive geoclimatic and demographic correlations of covid-19 spread rates," *Biology*, vol. 10, no. 7, p. 623, 2021.

[3] D. Duong, "What's important to know about the new covid-19 variants?" 2021.

[4] S. M. Anwar, M. Majid, A. Qayyum, M. Awais, M. Alnowami, and M. K. Khan, "Medical image analysis using convolutional neural networks: a review," *Journal of medical systems*, vol. 42, no. 11, pp. 1–13, 2018.

[5] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. Langlotz, K. Shpanskaya *et al.*, "Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning," *arXiv preprint arXiv:1711.05225*, 2017.

[6] C. An, H. Lim, D.-W. Kim, J. H. Chang, Y. J. Choi, and S. W. Kim, "Machine learning prediction for mortality of patients diagnosed with covid-19: a nationwide korean cohort study," *Scientific reports*, vol. 10, no. 1, pp. 1–11, 2020.

[7] J. P. Cohen, L. Dao, K. Roth, P. Morrison, Y. Bengio, A. F. Abbasi, B. Shen, H. K. Mahsa, M. Ghassemi, H. Li *et al.*, "Predicting covid-19 pneumonia severity on chest x-ray with deep learning," *Cureus*, vol. 12, no. 7, 2020.

[8] M. Pourhomayoun and M. Shakibi, "Predicting mortality risk in patients with covid-19 using machine learning to help medical decision-making," *Smart Health*, vol. 20, p. 100178, 2021.

[9] . World Health Organization *et al.*, "Statement on the second meeting of the international health regulations (2005) emergency committee regarding the outbreak of novel coronavirus (2019-ncov)," 2020.

[10] S. Hu, Y. Gao, Z. Niu, Y. Jiang, L. Li, X. Xiao, M. Wang, E. F. Fang, W. Menpes-Smith, J. Xia *et al.*, "Weakly supervised deep learning for covid-19 infection detection and classification from ct images," *IEEE Access*, vol. 8, pp. 118 869–118 883, 2020.

[11] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep learning*. MIT press Cambridge, 2016, vol. 1, no. 2.

[12] L. Deng and D. Yu, “Deep learning: methods and applications,” *Foundations and trends in signal processing*, vol. 7, no. 3–4, pp. 197–387, 2014.

[13] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez, “A survey on deep learning in medical image analysis,” *Medical image analysis*, vol. 42, pp. 60–88, 2017.

[14] M. Kukar, G. Gunčar, T. Vovko, S. Podnar, P. Černelč, M. Brvar, M. Zalaznik, M. Notar, S. Moškon, and M. Notar, “Covid-19 diagnosis by routine blood tests using machine learning,” *Scientific reports*, vol. 11, no. 1, pp. 1–9, 2021.

[15] L. Yan, H.-T. Zhang, Y. Xiao, M. Wang, Y. Guo, C. Sun, X. Tang, L. Jing, S. Li, M. Zhang *et al.*, “Prediction of criticality in patients with severe covid-19 infection using three clinical features: a machine learning-based prognostic model with clinical data in wuhan,” *MedRxiv*, 2020.

[16] L. Sun, F. Song, N. Shi, F. Liu, S. Li, P. Li, W. Zhang, X. Jiang, Y. Zhang, L. Sun *et al.*, “Combination of four clinical indicators predicts the severe/critical symptom of patients infected covid-19,” *Journal of Clinical Virology*, vol. 128, p. 104431, 2020.

[17] S. S. Aljameel, I. U. Khan, N. Aslam, M. Aljabri, and E. S. Alsulmi, “Machine learning-based model to predict the disease severity and outcome in covid-19 patients,” *Scientific Programming*, vol. 2021, 2021.

[18] G. Biau and E. Scornet, “A random forest guided tour,” *Test*, vol. 25, no. 2, pp. 197–227, 2016.

[19] J. Prada, Y. Gala, and A. Sierra, “Covid-19 mortality risk prediction using x-ray images.” *International Journal of Interactive Multimedia & Artificial Intelligence*, vol. 6, no. 6, 2021.

[20] J. P. Cohen, P. Morrison, L. Dao, K. Roth, T. Q. Duong, and M. Ghassemi, “Covid-19 image data collection: Prospective predictions are the future,” *arXiv 2006.11988*, 2020. [Online]. Available: https://github.com/ieee8023/covid-chestxray-dataset

[21] J. P. Cohen, J. Viviano, M. Hashir, and H. Bertrand, “Torchxrayvision: A library of chest x-ray datasets and models,” *2020*, 2020.

[22] J. P. Cohen, M. Hashir, R. Brooks, and H. Bertrand, “On the limits of cross-domain generalization in automated x-ray prediction,” in *Medical Imaging with Deep Learning*. PMLR, 2020, pp. 136–155.

[23] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[24] D. Opitz and R. Maclin, “Popular ensemble methods: An empirical study,” *Journal of artificial intelligence research*, vol. 11, pp. 169–198, 1999.