

On the Prediction of Isolation, Release, and Decease for COVID-19 Patients: A Case Study in South Korea

Tarik Alafif

*Computer Science Department
Jamoum University College
Umm Al-Qura University
Jamoum, Saudi Arabia
tkafif@uqu.edu.sa*

Reem Alotaibi

*Information Technology Department
Faculty of Computing and Information Technology
King Abdulaziz University
Jeddah, Saudi Arabia
ralotibi@kau.edu.sa*

Ayman Albassam

*Computer Science Department
Jamoum University College
Umm Al-Qura University
Jamoum, Saudi Arabia
asbassam@uqu.edu.sa*

Abdulelah Almudhayyani

*Computer Science Department
Jamoum University College
Umm Al-Qura University
Jamoum, Saudi Arabia
s44181770@st.uqu.edu.sa*

Abstract—A respiratory syndrome COVID-19 pandemic has become a serious public health issue nowadays. The COVID-19 virus has been affecting more than five millions of people worldwide. Some of them have recovered and have been released. Others have been isolated and few others have been unfortunately deceased. In this paper, we apply and compare different machine learning approaches such as decision tree models, random forest, and multinomial Logistic regression to predict isolation, release, and decease states for COVID-19 patients in South Korea. The proposed approaches are evaluated using Data Science for COVID-19 (DS4C) dataset. An analysis for DS4C dataset is also provided. Experimental results and evaluation show that random forest outperforms other approaches with 99.63% in state prediction accuracy and 99.51% in F1-score.

Index Terms—COVID-19, prediction, decision tree, random forest, multinomial logistic regression.

I. INTRODUCTION

The COVID-19 pandemic has first appeared in Wuhan, Hubei province, China in December 2019. It has become a serious public health issue worldwide as it spreads quickly. The COVID-19 virus is an infectious disease which directly affects people lungs. It is a branch of Coronaviruses family. It is believed that it has been transmitted from animals to human. The virus also transmits from human respiratory to another which is already noticed worldwide. The virus may cause mild or severe symptoms to affected people during the virus replications in their bodies. Affected people may develop symptoms such as dyspnea, fever, sore throat, cough, fatigue, and pneumonia. The virus has become a serious disease since it may cause death after developing the symptoms. Fortunately, many people have recovered and have been released while others are still isolated. While many people are recovered,

the number of deaths is very few compared to the number of recovered ones.

Machine learning (ML) algorithms play an important role in today's research because of the ability to automate complex tasks. These algorithms can learn from previous experience to predict future outcomes. Many research works have been developed and used in machine learning. Some of well known ML classifiers are decision trees (DTs), random forest (RF), and Multinomial Logistic Regression (MLR).

DT, the so called CART, is a popular supervised machine learning and predictive modelling approach [1]. It uses a simple tree structure representation to classify negative and positive examples based on their input features. DT has many advantages among other machine learning algorithms. It is easy to understand even for non-experts by displaying its graphical details since it imitates human thinking. Also, it is able to classify a large-scale categorical textual and numerical data. In addition, it doesn't require lots of data pre-processing.

DT is a classifier which can hold a discrete set of values to classify examples. The tree consists of nodes organized in top-down hierarchy in which a root node is at the top of the tree while leaf nodes are at the bottom. The leaf nodes represent the class labels while other non-leaf nodes represent the features. Each leaf node is labeled with a class and has a probability distribution over the example classes. The whole tree is split from the root node to the leaf nodes using arcs recursively. The arcs that split one node from another are labeled.

Various decision tree approaches have been proposed over the past years. ID3 [2], C4.5 [3] and CART [4] are the most common ones [5]. In order to select the best feature to split the training data, the approach iterates over all the features and then selects the ones that satisfies the splitting criterion.

Additionally, decision tree approach requires stopping rules such as purity, maximal depth, and minimal training examples at leaves.

Often a single DT is not enough for effective results as decision tree models tend to high variance, which means that small changes in the data may lead to a different tree and hence high variance in the performance. Random Forest (RF) mitigates the variance issue by generating multiple decision trees.

RF is an ensemble learning approach that is based on generating a large number of decision trees using a different subset of the training data [6, 7]. These subsets are chosen by random sampling of the original training data. The final predictions are made by taking the majority vote from all individual classification trees.

Regression approaches are commonly used to predict future events in many research fields. In machine learning context, logistic regression is a well-known technique that predicts binary outcomes. On the other hand, MLR is an extended version of a binary logistic regression approach that allows performing multi-class prediction where prediction outcome can be either dichotomous or continuous [8].

In this paper, we apply and compare different machine learning algorithms such as CARTs, RF, and MLR to predict isolation, release, and decease states for COVID-19 patients in South Korea. The proposed approaches are evaluated using Data Science for COVID-19 (DS4C) dataset. An analysis for DS4C dataset is provided. Experimental results and evaluation show that RF outperforms other approaches with 99.63% in state prediction accuracy and 99.51% in F1-score.

The remainder of this paper is organized as follows. In **Section II**, related work is reviewed. **Section III** introduces basic notations and mathematical models used throughout the paper. In **Section IV**, we briefly describe and analyze DS4C dataset. In **Section V**, we present our work. In **Section VI**, Experimental results and evaluation are provided using public DS4C dataset. Finally, we conclude our work in **Section VII**.

II. RELATED WORK

Many research works have been published recently to analyze and tackle the problem of COVID-19 virus in many fields. Some current attempts for COVID-19 analysis, classifications, and predictions used structured data such as a numerical and textual data while others used an unstructured data such as images.

Several methods were experimented using structured data in [9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21] for COVID-19 analysis and predictions. Fanelli and Piazza [9] analyzed and forecasted COVID-19 spreading in China, Italy, and France. Nadia and Hazem [10] studied the effects of sex, region, infection reason, birth year, release date, and the diseased date on the recovered and deceased cases for COVID-19 patients. Hu et al. [11] applied a stacked auto-encoder and K-means to model the transmission dynamics of the epidemics based on provinces and cities. The model was built to forecast the confirmed cases of COVID-19. Remuzzi and Remuzzi

[12] predicted the number of infected patients in Italy using an exponential curve. Tartai and Varallyay [13] attempted to predict the results of COVID-19 epidemic in a region using a logistic model. Pedersen and Meneghin [14] analyzed and predicted the spread of the COVID-19 virus using restrictions to reduce the epidemic in Italy. Peng et al. [15] proposed a SEIR model to analyze the epidemic in China. Petropoulos and Makridakis [16] forecasted the confirmed cases, death cases, and recovery cases of COVID-19 using time series forecasting approaches. Chatterjee et al. [17] and Jia et al. [18] developed mathematical models to study the impact of COVID-19 epidemic. Weissman et al. [19] proposed a SIR model to predict hospital capacity needs during the COVID-19 pandemic. Yang et al. [20] used a long term short memory neural network to predict the COVID-19 epidemic in China. Similar to [20], Vattay [21] attempted to predict the epidemic in Italy.

Other several methods were proposed in [22, 23, 24, 25, 26, 27, 28, 29, 30, 31] which apply machine and deep learning methods to detect COVID-19 from images. Barstugan et al. [22] proposed to use hand-crafted feature extractors and SVM to detect COVID-19 from abdominal Computed Tomography (CT) scans. Wang et al. [23] used Convolutional Neural Network (CNN) to classify whether the patient is infected or not from chest CT scans. Li et al. [24] applied a ResNet50 pre-trained CNN model to chest CTs to detect COVID-19 by classifying Pneumonia and other lung diseases. Different from [23] and [24], Narin et al. [25] detects COVID-19 virus from chest X-ray images using transfer learning since X-ray images are cheaper than CT images. Pre-trained CNN models were used such as ResNet50, InceptionV3, and Inception-ResNetV2. Similarly, Apostolopoulos and Mpesiana [26] applied VGG19, MobileNetV2, Inception, Xception, and Inception-ResNetV2 pre-trained CNN models to detect COVID-19 from chest X-ray images. [28] designed a COVID-Net based CNN architecture to detect COVID-19 from chest X-ray images. [27] proposed to use Capsule Networks to identify COVID-19 from chest X-ray images. [29] applied deep features and Support vector machine to classify COVID-19 infected X-ray images. Recently, [30] estimated the confidence in the deep learning based methods for detecting COVID-19 from X-ray images. Monte-Carlo dropweights based Bayesian CNN was used to improve the performance of the classification. Wang and Wong [31] proposed a machine learning framework to detect COVID-19 from X-ray images.

Different from the aforementioned works to tackle the COVID-19 problem from different perspectives, we apply and compare different machine learning approaches such as CARTs, RF, and MLR to predict isolation, release, and decease states for COVID-19 patients in South Korea. The proposed approaches are evaluated using DS4C dataset to measure the performance of the predictions.

III. PRELIMINARIES

Let D a set of patients records such that $D = \{X_1, X_2, \dots, X_n\}$, where n is the number of patients. Each

patient record X is belonging to one of the classes j that is used for the classification task such that $D = \{(X_1, j_1), (X_2, j_2), \dots, (X_n, Y_n)\}$. Each patient record X is composed of several input features such that $X = \{x_1, x_2, \dots, x_m\}$, where m is the total number of patient features. Then, each feature is a predictor variable of the class j such that $X = \{(x_1, x_2, \dots, x_m, j)\}$, Y is the class variable taking values either in a set of classes $C = \{1, 2, \dots, k\}$, where k is the total number of classes.

Next, we show the basic mathematical equations used in the following approaches:

A. CARTs

In CARTs, Gini impurity is used to measure the purity in the decision tree models by knowing how good the split of input features in each node [32, 33]. Then, Gini impurity is computed for the examples in the training data with the class j as shown in Equation 1:

$$\text{Gini}(D|x) = 1 - \sum_{j=1}^k P_j^2 \quad (1)$$

where x , k and P_j are the examined feature, the number of classes and the probability of the j^{th} class, respectively.

The Gini Index is computed for each input feature partition and the average Gini Index for the examined feature x is defined as:

$$\text{Gini}(D|x) = \sum_{i=1}^{\text{vals}(x)} \frac{|D_i|}{|D|} \text{Gini}(D_i|x) \quad (2)$$

where x , D_i and $\text{vals}(x)$ are the examined feature, a data partition and number of discrete values in feature x , respectively. The feature with the lowest Gini Index is chosen as the splitting feature.

B. RF

A RF consists of a collection of decision trees $T = \{1, \dots, ntree\}$. Each decision tree is trained on a different subset of the data using bootstrap sampling (bagging) out of D . Final predictions are made using the majority votes. The RF employs random feature selection for each node of every tree in the forest, meaning that the splitting feature x is chosen at random.

C. MLR

A logistic regression is defined as a generalized linear model that is usually used to classify binary outcomes [8]. Ridge estimator is proposed to improve model prediction [34]. Logistic regression approach is slightly modified to deal with multi-class classification that is called the MLR, in case of predicting k outcome classes for n instances with m observations, the parameter matrix of regression coefficients B to be calculated will be an $m * (k - 1)$ matrix. e is the exponential function. Class j probability with the exception of the last class is shown in Equation 3:

$$P_j(X_i) = \frac{e^{X_i B_j}}{\sum_{j=1}^{k-1} e^{X_i B_j} + 1} \quad (3)$$

The last class probability is shown in Equation 4:

$$1 - \sum_{j=1}^{k-1} P_j X_i = \frac{1}{\sum_{j=1}^{k-1} e^{X_i B_j} + 1} \quad (4)$$

The (negative) multinomial log-likelihood is shown in Equation 5::

$$L = - \sum_{j=1}^n \left\{ \sum_{j=1}^{k-1} Y_{ij} * \ln(P_j(X_i) + (1 - (\sum_{j=1}^{k-1} Y_{ij})) * \ln(1 - \sum_{j=1}^{k-1} P_j(X_i))) \right\} + \text{ridge} * (B^2) \quad (5)$$

D. Evaluation Metrics

The accuracy for multi-class classification is defined as the average number of correct predictions for all classes as shown in Equation 6. Error rate is the complement of accuracy.

$$\text{accuracy} = \frac{1}{k} \sum_{j=1}^k \frac{TP_j + TN_j}{TP_j + FP_j + TN_j + FN_j} \quad (6)$$

where TP_j , TN_j , FP_j and FN_j represent true positive, true negative, false positive and false negative for the j^{th} class respectively.

F1-score is the average of precision and recall as shown in Equations 7:

$$\begin{aligned} \text{precision}_j &= \frac{TP_j}{TP_j + FP_j} \\ \text{recall}_j &= \frac{TP_j}{TP_j + FN_j} \\ \text{F1-score} &= \frac{1}{k} \sum_{j=1}^k \frac{2 * \text{precision}_j * \text{recall}_j}{\text{precision}_j + \text{recall}_j} \end{aligned} \quad (7)$$

where precision_j and recall_j represent precision and recall for the j^{th} class, respectively.

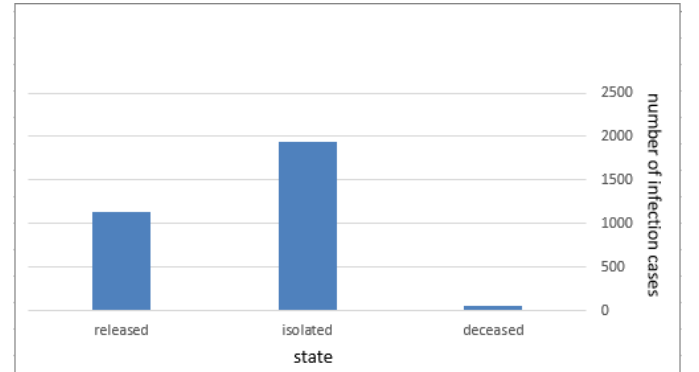


Fig. 1. The the number of the samples distribution for each state in DS4C dataset.

the patients records. For DT models, we use the maximum DT depth of 3, 5 and 10 in our work. Figure 2 shows the generated DT architecture after training using the depth of 3. The DT architectures with the maximum depth of 5 and 10 generate a very large tree which consist of many branches, many leaf nodes and non-leaf nodes. A generalization capability is lost if the the depth is more than 10 as seen in Figure 4. For RF model, the model is trained with the default parameters according to randomForest package in R [6]. The default parameters in the randomForest algorithm: *ntree*, the number of classification tree (the default value is 500) and *mtry*, the number of features tested at each node (the default values is $\sqrt{m}$), where m is the total number of features).

For MLR model, the model-building process is carried out using weka tool with a ridge estimator which is considered as a stable implementation. The predictive model is fed with the full set of features. The Quasi-Newton method is used to perform optimization task. Ridge values of 1×10^8 is used in the log likelihood calculation.

VI. EXPERIMENTAL RESULTS

Tree-based approaches are implemented in R using *rpart* and *randomForest* packages [6, 36] while MLR approach was built using the Weka version 3.8.4 of machine learning tool [37]. A single laptop 1.6 GHz Dual-Core Intel Core i5 CPU with 8 GB of RAM is used to train and test the models. 10-fold cross validation is applied for training and testing the models using DS4C dataset. A seed state is set to 1234.

Table I shows the accuracy, error rates, and *F1 – score* for the tested models. For DT models, we choose three different tree depths: 3, 5, and 10. In order to obtain the effect of the maximum tree depth parameters on the experimental results, we show the performance curve to verify the effectiveness of depth parameters on the models. Figure 4 shows the models accuracy versus different tree depths. It clearly shows that the DT models accuracy stops improving at the maximum depth of 10.

TABLE I
PREDICTION ACCURACY, ERROR RATES AND F1-SCORE SCORE OF THE ALGORITHMS

Algorithms implemented	Accuracy	Error rate	F1-score
DT (Depth = 3)	77.61%	22.39%	50.17 %
DT (Depth = 5)	84.51%	15.49%	70.26 %
DT (Depth = 10)	86.39%	13.61 %	71.52 %
MLR	89.03%	10.97%	77.51%
RF	99.63%	00.37%	99.51%

In addition, RF and MLR approaches are also compared. Figure 5 shows the effect of the number of generated trees in the RF on the classification error rate. We can see that the error rate becomes stable when the number of trees is increased. The curve in black color represents out-of-bag (OOB) error rate and the other colors represent misclassification error rate curves. The out-of-bag error is estimated internally during building the DTs. The RF approach outperforms other ap-

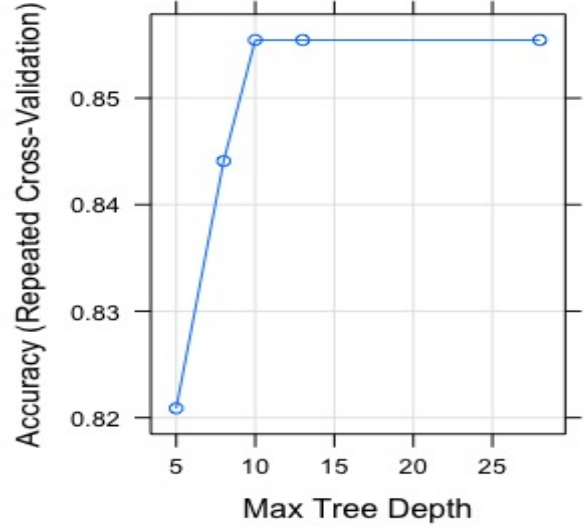


Fig. 4. Maximum tree depth tuning based DT.

proaches with 99.63% in state prediction accuracy and with a margin of 10.6%.

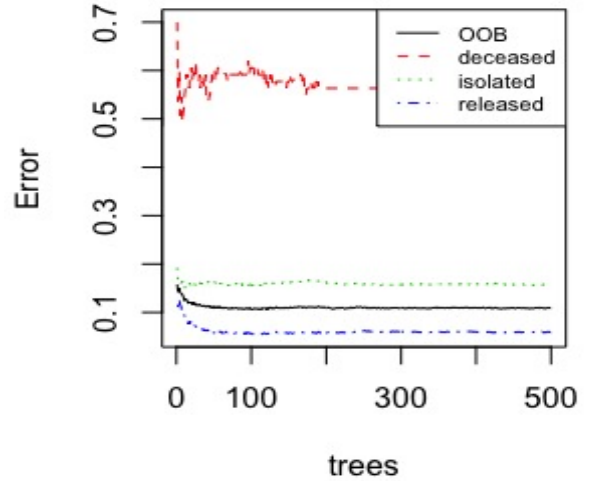


Fig. 5. RF error rates with an increase in the number of classification trees. The number of trees and the estimated error rates are shown on x-axis and y-axis respectively.

VII. CONCLUSION

In this paper, we apply and compare different machine learning approaches such as DT models, RF, and MLR to predict isolation, release, and decrease states for COVID-19 patients in South Korea. The proposed approaches are

evaluated using DS4C dataset. An analysis for DS4C dataset is provided. Experimental results and evaluation show that RF outperforms other approaches with 99.63% in state prediction accuracy and 99.51% in F1-score.

ACKNOWLEDGMENT

This work is funded by Research and Development Grants Program for National Research Institutions and Centers (GRANTS), Target Research Program, Infectious Disease Research Grant Program, King Abdulaziz City for Science and Technology (KACST), Kingdom of Saudi Arabia, grant number (5-20-01-007-0001).

REFERENCES

- [1] L. Rokach and O. Z. Maimon, *Data mining with decision trees: theory and applications*. World scientific, 2008, vol. 69.
- [2] R. Quinlan, "Induction of decision trees," *Machine Learning*, vol. 1, no. 1, pp. 81–106, Mar. 1986. [Online]. Available: <http://dx.doi.org/10.1023/A:1022643204877>
- [3] —, *C4.5: Programs for Machine Learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1993.
- [4] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*. Chapman and Hall, New York, 1984.
- [5] A. Priyam, G. Abhijeeta, A. Rathee, and S. Srivastava, "Comparative analysis of decision tree classification algorithms," *International Journal of current engineering and technology*, vol. 3, no. 2, pp. 334–337, 2013.
- [6] A. Liaw and M. Wiener, "Classification and regression by randomforest," *R News*, vol. 2, no. 3, pp. 18–22, 2002. [Online]. Available: <https://CRAN.R-project.org/doc/Rnews/>
- [7] T. K. Ho, "Random decision forests," in *Third International Conference on Document Analysis and Recognition, ICDAR 1995, August 14 - 15, 1995, Montreal, Canada. Volume I*. IEEE Computer Society, 1995, pp. 278–282. [Online]. Available: <https://doi.org/10.1109/ICDAR.1995.598994>
- [8] D. W. Hosmer, *Applied logistic regression*, 3rd ed., ser. Wiley series in probability and statistics. Hoboken, N.J: Wiley.
- [9] D. Fanelli and F. Piazza, "Analysis and forecast of covid-19 spreading in china, italy and france," *Chaos, Solitons & Fractals*, vol. 134, p. 109761, 2020.
- [10] A.-R. Nadia and A.-N. Hazem, "Data analysis of coronavirus covid-19 epidemic in south korea based on recovered and death cases," *Journal of Medical Virology*, 2020.
- [11] Z. Hu, Q. Ge, L. Jin, and M. Xiong, "Artificial intelligence forecasting of covid-19 in china," *arXiv preprint arXiv:2002.07112*, 2020.
- [12] A. Remuzzi and G. Remuzzi, "Covid-19 and italy: what next?" *The Lancet*, 2020.
- [13] D. Tátrai and Z. Várallyay, "Covid-19 epidemic outcome predictions based on logistic fitting and estimation of its reliability," *arXiv preprint arXiv:2003.14160*, 2020.
- [14] M. G. Pedersen and M. Meneghini, "Quantifying undetected covid-19 cases and effects of containment measures in italy," *ResearchGate Preprint (online 21 March 2020) DOI*, vol. 10, 2020.
- [15] L. Peng, W. Yang, D. Zhang, C. Zhuge, and L. Hong, "Epidemic analysis of covid-19 in china by dynamical modeling," *arXiv preprint arXiv:2002.06563*, 2020.
- [16] F. Petropoulos and S. Makridakis, "Forecasting the novel coronavirus covid-19," *PloS one*, vol. 15, no. 3, p. e0231236, 2020.
- [17] K. Chatterjee, K. Chatterjee, A. Kumar, and S. Shankar, "Healthcare impact of covid-19 epidemic in india: A stochastic mathematical model," *Medical Journal Armed Forces India*, 2020.
- [18] L. Jia, K. Li, Y. Jiang, X. Guo *et al.*, "Prediction and analysis of coronavirus disease 2019," *arXiv preprint arXiv:2003.05447*, 2020.
- [19] G. E. Weissman, A. Crane-Droesch, C. Chivers, T. Luong, A. Hanish, M. Z. Levy, J. Lubken, M. Becker, M. E. Draugelis, G. L. Anesi *et al.*, "Locally informed simulation to predict hospital capacity needs during the covid-19 pandemic," *Annals of internal medicine*, 2020.
- [20] Z. Yang, Z. Zeng, K. Wang, S.-S. Wong, W. Liang, M. Zanin, P. Liu, X. Cao, Z. Gao, Z. Mai *et al.*, "Modified seir and ai prediction of the epidemics trend of covid-19 in china under public health interventions," *Journal of Thoracic Disease*, vol. 12, no. 3, p. 165, 2020.
- [21] G. Vattay, "Predicting the ultimate outcome of the covid-19 outbreak in italy," *arXiv preprint arXiv:2003.07912*, 2020.
- [22] M. Barstugan, U. Ozkaya, and S. Ozturk, "Coronavirus (covid-19) classification using ct images by machine learning methods," *arXiv preprint arXiv:2003.09424*, 2020.
- [23] S. Wang, B. Kang, J. Ma, X. Zeng, M. Xiao, J. Guo, M. Cai, J. Yang, Y. Li, X. Meng *et al.*, "A deep learning algorithm using ct images to screen for corona virus disease (covid-19)," *MedRxiv*, 2020.
- [24] L. Li, L. Qin, Z. Xu, Y. Yin, X. Wang, B. Kong, J. Bai, Y. Lu, Z. Fang, Q. Song *et al.*, "Artificial intelligence distinguishes covid-19 from community acquired pneumonia on chest ct," *Radiology*, p. 200905, 2020.
- [25] A. Narin, C. Kaya, and Z. Pamuk, "Automatic detection of coronavirus disease (covid-19) using x-ray images and deep convolutional neural networks," *arXiv preprint arXiv:2003.10849*, 2020.
- [26] I. D. Apostolopoulos and T. A. Mpesiana, "Covid-19: automatic detection from x-ray images utilizing transfer learning with convolutional neural networks," *Physical and Engineering Sciences in Medicine*, p. 1, 2020.
- [27] P. Afshar, S. Heidarian, F. Naderkhani, A. Oikonomou, K. N. Plataniotis, and A. Mohammadi, "Covid-caps: A capsule network-based framework for identification

of covid-19 cases from x-ray images,” *arXiv preprint arXiv:2004.02696*, 2020.

- [28] L. Wang and A. Wong, “Covid-net: A tailored deep convolutional neural network design for detection of covid-19 cases from chest radiography images,” *arXiv preprint arXiv:2003.09871*, 2020.
- [29] P. K. Sethy and S. K. Behera, “Detection of coronavirus disease (covid-19) based on deep features,” *Preprints*, vol. 2020030300, p. 2020, 2020.
- [30] B. Ghoshal and A. Tucker, “Estimating uncertainty and interpretability in deep learning for coronavirus (covid-19) detection,” *arXiv preprint arXiv:2003.10769*, 2020.
- [31] F. Shi, L. Xia, F. Shan, D. Wu, Y. Wei, H. Yuan, H. Jiang, Y. Gao, H. Sui, and D. Shen, “Large-scale screening of covid-19 from community acquired pneumonia using infection size-aware classification,” *arXiv preprint arXiv:2003.09860*, 2020.
- [32] L. Rokach and O. Maimon, *Data Mining with Decision Trees: Theory and Applications*. River Edge, NJ, USA: World Scientific Publishing Co., Inc., 2008.
- [33] L. E. Raileanu and K. Stoffel, “Theoretical comparison between the gini index and information gain criteria,” *Annals of Mathematics and Artificial Intelligence*, vol. 41, no. 1, pp. 77–93. [Online]. Available: <http://dx.doi.org/10.1023/B:AMAI.0000018580.96245.c6>
- [34] S. le Cessie and J. van Houwelingen, “Ridge estimators in logistic regression,” *Applied Statistics*, vol. 41, no. 1, pp. 191–201, 1992.
- [35] Datartist, “Data Science for COVID-19 (DS4C),” <https://www.kaggle.com/kimjihoo/coronavirusdataset>, 2020, [Online; accessed 25-April-2020].
- [36] T. Therneau, B. Atkinson, and B. Ripley, “Rpart: Recursive partitioning and regression trees. r package version 4.1-8,” *R News*, 2014. [Online]. Available: <http://CRAN.R-project.org/package=rpart>
- [37] E. Frank, M. A. Hall, and I. H. Witten, “The WEKA workbench,” in *Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques"*, fourth edition ed. Morgan Kaufmann, 2016.