Proceedings of the 19th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 537–542
DOI: 10.15439/2024F8981
ISSN 2300-5963 ACSIS, Vol. 39

A novel ensemble learning technique of shallow models applied on a COVID-19 dataset

D Babuc
ORCID: 0009-0000-5126-6480
Computer Science Department, West University of Timişoara
Blvd. Vasile Pârvan 4, 300223 Timişoara, Romania
Email: diogen.babuc00@e-uvt.ro

***Abstract*—Our lives were affected by the COVID-19 pandemic. In order to face this crisis, we provided a novel ensemble learning strategy to tackle the COVID-19 prediction and classification problems. Because of their capacity to handle the complex and varied nature of COVID-19 data, a range of shallow models, including K-Nearest Neighbors, Decision Trees, Support Vector Machines, Classification and Regression Trees, and Extreme Gradient Boost, are included in our method. Using a COVID-19 dataset, each model is trained independently and then ensemble learning techniques are used to integrate the predictions of the models. We use strict model validation and hyperparameter optimization to improve performance. Comparing our ensemble method to a single model or traditional ensemble techniques, our results show considerable improvements in classification performance and prediction accuracy.**

Index Terms–Ensemble Learning, Machine Learning, COVID-19, Performance Metrics, Prediction and Classification.

I. Introduction

SINCE its appearance in late 2019, the COVID-19 pandemic has had an influence on cultures, economy, and healthcare systems across the globe [1]. Predictive modeling has become an essential process for studying and projecting the trajectory of the virus as governments and health organizations struggle to stop its spread [2]. In this paper, we investigate the creation of models that forecast the total number of COVID-19 cases worldwide. We used a dataset that runs through September 2020. Additionally, we classify nations into those with and without a higher risk of contracting SARS-CoV-2. Due to the COVID-19 pandemic's intricacy, new methods of data analysis and forecasting have been required.

Our primary focus lies in exploring the predictive potential of historical data up to September 2020. We want to capture critical phases of the pandemic's evolution. Through retrospective analysis, we aim to elucidate patterns, trends, and underlying factors influencing the spread of COVID-19 across different regions and timeframes. Using statistical indicators and machine learning techniques, we seek to construct predictive models capable of discerning the complex interplay between various epidemiological variables and forecasting the total cases of COVID-19 with precision and reliability. Through this interdisciplinary effort, we aim to contribute to ongoing global efforts to combat the COVID-19 pandemic. Our goal is to provide stakeholders with the knowledge and resources they need to effectively navigate the obstacles presented by this unprecedented public health catastrophe by using the power of prospective and retrospective predictive modeling and data-driven insights.

This work was supported by West University of Timişoara.

We hope to provide a better understanding of the dynamics of the pandemic and enable informed decision-making in the face of uncertainty (Fig. I).

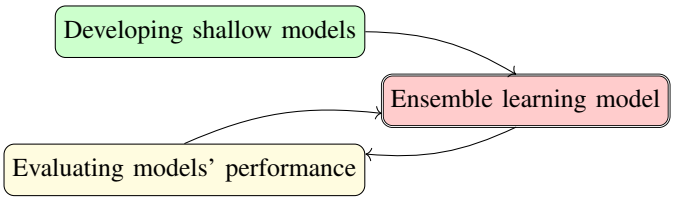

Fig. 1. Objectives for COVID-19 infection risk estimation.

Concretely, our main objectives are:

1) Developing and optimizing predictive shallow machine learning models using historical worldwide COVID-19 data up to September 2020.
2) Building an ensemble learning model, called *Renesansa*, for investigating the impact of epidemiological factors, including active cases, total tests conducted, and population demographics, on the total number of COVID-19 cases worldwide.
3) Evaluating performance of the predictive models developed through evaluation metrics and statistical indicators.

We compare the models' results with the observed data to gauge the trustworthiness and effectiveness of the selected forecasting and classification methodologies.

II. Background Information and Related Works

In this section, we will discuss the background information and previous studies that have explored various methodologies, from traditional statistical models to modern ensemble learning approaches. We want to accurately forecast transmission trends and classify disease outcomes.

A. COVID-19 Disease

Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) is a new betacoronavirus that is a member of the Coronaviridae family and the source of COVID-19 [3]. SARS-CoV-2 is a single-stranded, enveloped, positive-sense RNA virus with spike (S), envelope, membrane and nucleocapsid proteins. Viral entrance into host cells is mediated by the S protein, which binds to the angiotensin converting enzyme 2 receptor and promotes membrane fusion and viral multiplication.

A wide range of clinical signs are displayed by COVID-19, from moderate or asymptomatic sickness to severe respiratory failure [4]. Common symptoms include fever, cough, fatigue, and gastrointestinal symptoms. Severe cases are characterized by acute respiratory distress syndrome, multiorgan dysfunction, and thrombotic complications. Certain population groups, such as older adults and immunocompromised individuals, are at increased risk of severe disease and adverse outcomes.

B. COVID-19 Pandemics' Social Impact

One of the worst global health emergencies in recent memory, the COVID-19 epidemic has had a tremendous effect on civilizations all around the world. [5]. Since the new coronavirus SARS-CoV-2 first appeared in late 2019, the pandemic has quickly expanded throughout continents, overcoming geographic barriers and igniting hitherto unheard-of public health measures [6].

COVID-19 has challenged our understanding of infectious diseases and highlighted the interconnectedness of our modern world. From the outset, the pandemic has posed multifaceted challenges, ranging from containment efforts and healthcare delivery to social distancing measures and economic stability [7]. The COVID-19 pandemic has underscored the importance of rapid and coordinated responses from governments, healthcare institutions, and communities to mitigate transmission, protect vulnerable populations, and minimize the burden on healthcare infrastructure. Measures such as lockdowns, travel restrictions, mass testing, contact tracing, and vaccination campaigns have been implemented globally.

The pandemic has also exposed existing vulnerabilities and inequalities within societies, disproportionately affecting marginalized communities, low-income countries, and front-line workers [8]. Disparities in access to healthcare and socioeconomic factors have exacerbated the impact of COVID-19 on vulnerable populations. The rapid development of vaccines, diagnostic tests, therapeutics, and public health interventions has demonstrated the collective resilience and ingenuity of the global scientific community in the face of adversity.

C. Shallow Models' Results

In this section, we will analyze the results from the convex literature for each shallow machine learning model.

K-Nearest Neighbors (KNN)

Ye and colleagues [9] implemented an intelligent system for classifying the severity of COVID-19, to help clinicians in their decisions. The authors trained the model HHO-FKNN, based on KNN, considering the list of symptoms, complications degree, already existing diseases, and the immune system. They achieved an average accuracy of 94%, the Matthews' correlation coefficient of 88.91%, an average sensitivity of 90%, and an average specificity of 96.67%.

In another article [10], Hamed and coauthors focused on incomplete datasets to predict if a patient suffers from coronavirus or not, and to classify properly, using KNN bases, all the patients. They used two distances: mahalanobis and euclidean. For the mahalanobis distance, the authors obtained an average accuracy of 84%, a sensitivity of 76%, a precision of 95%, and an F1-score of 84%. For the euclidean distance, they achieved an accuracy of 88%, a sensitivity of 87%, a precision of 91%, and the F1-score of 88%.

Decision Trees (DT)

The authors of [11] calculated the performance evaluation metrics for predicting retrospectively the coronavirus considering the blood gas parameter, by using decision trees methods. They got an accuracy ratio of 65% for the correctly predicting cases and 68.2% for correctly identifying people who indeed suffer from coronavirus. When categorizing patients by cutoff values (less than 1.0, between 1.0 and 1.6, and bigger than 1.6), the achieved an accuracy of 92.7%, the metric which was the main target for the authors.

Support Vector Machines (SVM)

In the paper written by Singh and coauthors [12], the scientists tried to predict coronavirus with SVM by treating on time series data. They considered the active cases, total number of deaths, and recovered ones from January and until April 2020 with international data. They referenced to the article [13], where an accuracy of 88% and an F1-score value of 76% were achieved. The authors of [14] got an accuracy of 88.76% by using the radial basis function in SVM when classifying countries into those at risk and without risk.

Classification and Regression Trees (CART)

CART [15] have been utilized in COVID-19 prediction and classification tasks owing to their simplicity and interpretability. By recursively partitioning the data based on the most informative features, CART constructs decision trees that can effectively classify COVID-19 cases into different categories or predict outcomes such as disease severity. CART's ability to handle both numerical and categorical data makes it well-suited for analyzing heterogeneous COVID-19 datasets with diverse epidemiological variables [16].

The authors of [17] built a predictive instruction for COVID-19 pneumonia and classified pneumonia into the one provoked by COVID-19 and not provoked by it. They obtained an area under the ROC curve (ROC-AUC) of 86%, and an accuracy of maximum 95%. On the other hand, Zimmerman and colleagues [16] obtained an ROC-AUC of 76%, a sensitivity of 69%, and specificity of 78%.

Extreme Gradient Boost (XGBoost)

XGBoost [18], an ensemble learning technique, has been widely applied in COVID-19 prediction and classification tasks due to its exceptional performance and scalability. By combining the predictions of multiple weak learners, such as decision trees, XGBoost can effectively capture complex patterns in COVID-19 data and improve predictive accuracy. However, XGBoost may require careful tuning of hyperparameters and regularization techniques to prevent overfitting, especially with large-scale COVID-19 datasets [19].

In the article [20], Carvalho and colleagues built an approach which can diagnose accurately and precisely the COVID-19 for the patients with XGBoost layer added to a convolutional neural network. They obtained an accuracy of 95.07%, a recall of 95.1%, precision of almost 95%, the F1-score and the ROC-AUC of 95% both, while the Cohen's index was 90%. The second article [21] (Fang et al.) related to XGBoost analyzed statistical indicators such as mean squared errors, mean absolute errors and the R-squared coefficient to improve the prediction of the number of patients who are infected with SARS-CoV-2 only in the USA, providing an excellent R-squared, no bigger than 4.1.

D. Ensemble Learning Framework

The predictions of many base models are combined in an ensemble learning process to get a final prediction that is more accurate and dependable [22]. The idea behind ensemble learning is to leverage the diversity of individual models to compensate for their weaknesses and improve overall predictive performance. There are several ensemble learning frameworks, including bagging, boosting, and stacking, each with its advantages and disadvantages.

There are various benefits to bagging. By using diverse subsets of the training data to train numerous base models, it effectively lowers variance and overfitting. Additionally, bagging can be parallelized. However, it comes with its own set of disadvantages. As the number of base models increases, computational complexity and memory requirements may become prohibitive, particularly for large datasets [23]. Boosting, on the other hand, offers distinct advantages. It builds a strong learner iteratively by focusing on examples that are difficult to classify or have high prediction errors. Also, it is sensitive to noise and outliers, potentially leading to overfitting on irrelevant examples during training [24]. Stacking integrates predictions from multiple heterogeneous base models, leveraging the strengths of different modeling techniques. On the other hand, stacking may suffer from information leakage or overfitting if the meta-learner is trained on predictions from the same data used to train the base models [25].

III. Proposed Model

Given the complexity and heterogeneity of COVID-19 data, a diverse set of shallow models is selected to capture various aspects of the pandemic. Models such as KNN, DT, SVM, CART, and XGBoost (Fig. 2) are chosen based on their suitability for handling different types of COVID-19 data, including epidemiological, clinical, genomic, and environmental factors.

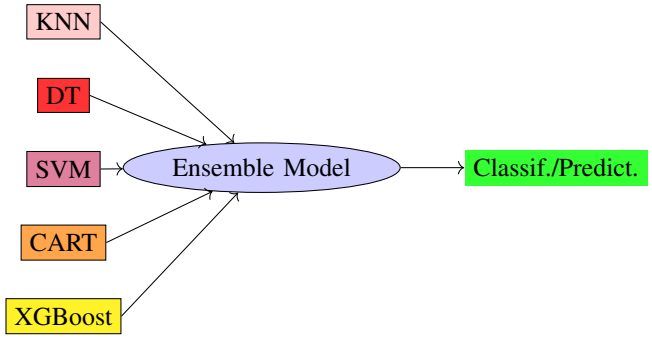

Fig. 2. Hyperparameterized ensemble learning model for classifying countries into those at risk and safe, and predicting the specific prospective *Total Cases* value for a sample.

Each selected shallow model is trained on COVID-19 data obtained from reliable source publicly available on Kaggle [26]. The training process involves preprocessing the data, selecting appropriate features, and tuning hyperparameters to optimize model performance. For instance, KNN is trained using historical COVID-19 case data to predict future transmission trends, while SVM is trained to classify patients based on clinical symptoms and demographic information. Ensemble learning techniques, including bagging, boosting, and stacking, are employed to combine the predictions of the individual shallow models. Bagging is used to aggregate predictions from multiple models to reduce variance and overfitting, boosting adapts the models iteratively to improve performance over time, and stacking integrates predictions from diverse models to capture complex relationships in COVID-19 data. Hyperparameters for both individual models and the ensemble framework are tuned using COVID-19-specific data and evaluation metrics. Grid search or Bayesian optimization techniques are applied to identify optimal hyperparameter configurations that maximize predictive performance and classification accuracy for COVID-19-related outcomes such as disease transmission, severity, etc.

The performance of the ensemble model is validated using cross-validation techniques on COVID-19 datasets. Special attention is given to account for temporal and geographical variations in COVID-19 data to ensure robustness and generalizability. Evaluation metrics such as accuracy, precision, recall, F1-score, and ROC-AUC are used to assess predictive performance and classification accuracy.

Once validated, the ensemble model is deployed into production systems or applications for real-time COVID-19 prediction and classification (Fig. 2). Integration into public health surveillance systems, decision support tools, or epidemiological models ensures that the ensemble model contributes to informed decision-making and effective public health interventions in the fight against COVID-19.

IV. Results

In this chapter, we will present our results, through which we get a possible solution for the specific total case value for Romania and also for classifying nations into those at risk and without risk. We offer some graphs and charts for results visualization.

The calculations include the statistical indicators: Mean Squared Error (MSE), Mean Absolute Error (MAE), and R-squared coefficient, and correlation coefficient. We also calculate the evaluation metrics: sensitivity, specificity, accuracy, and precision. For the selected dataset, a strong Pearson correlation is obtained between the column *Total Recovered*, and the column for the dependent variable, *Total Cases* (0.9). Between the column *Active Cases* and *Total Cases*, there is a strong correlation, of 0.72. Other columns included in the independent variable are *Population*, *Serious or Critical*, and *Total Tests*.

A. Statistical Indicators for All Selected Models

We calculate the statistical indicators for 40 different values of the selected parameters of each state-of-the-art model. The *k-neighbors* parameter is the one that varies in the KNN model. The value for MSE does not exceed 2.1 and MAE is at most 1.1. For the 40 values, the R-squared coefficient is between 0.75 and 0.98. The consistently low MSE and MAE values across different k values suggest that the KNN model is consistently accurate in its predictions for various levels of *k-neighbors* parameter. This stability in accuracy is important for identifying critical regions exactly. It is important to select an appropriate value, to balance model complexity and generalizability to ensure reliable predictions.

In the DT model, the parameter *max-depth* varies. An MSE between 1 and 4 is obtained; the MAE is not greater than 1.7. The coefficient of determination for DT is between 0.55 and 0.95. The varying R-squared coefficients indicate that certain depths might result in better explanations and stronger relationships in the data, potentially leading to more accurate predictions for critical geographic regions. The choice of an appropriate tree depth involves considering a balance between accuracy and model complexity to ensure reliable predictions.

The SVM model provides, through the regularization parameter $C = \frac{1}{10}$, an MSE of at most 7.25, and an MAE of at most 2.3. The R-squared coefficient is around 0.7 for most values. Although the explanatory strength of the model might be moderate, the consistent accuracy at different levels of regularization indicates that the model performs fairly well. It is important to consider the specific context of the study, the trade-offs between regularization and accuracy, and potential avenues for model improvement.

The CART model, with the value for *max-depth* and *min-leaf* varying, has an MSE between 5 and 11, and an MAE of at most 2.6. The R-squared coefficient ranges between 0.88 and 0.98. It is important to select appropriate combinations of *max-depth* and *min-leaf* for dependable forecasts, to strike a balance between generalizability and model complexity.

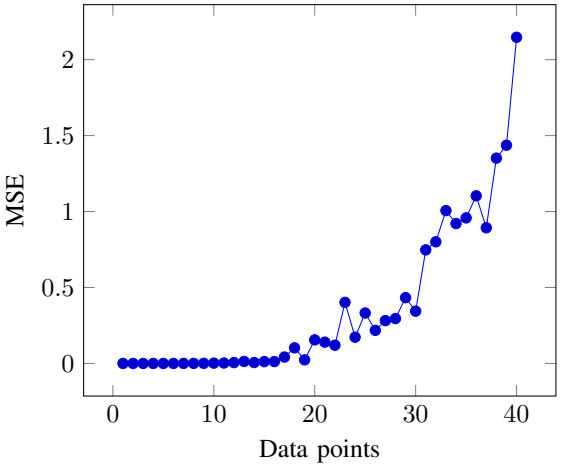

Fig. 3. The MSE indicator for the *Renesansa* ensemble model.

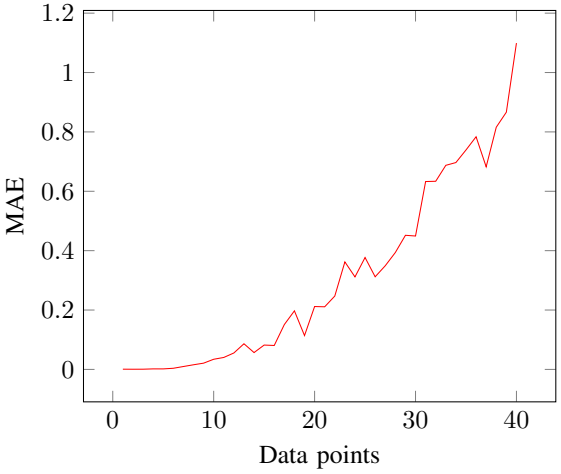

Fig. 4. Mean Absolute Error (MAE) for Renesansa ensemble model.

XGBoost is an outstanding model, with an MSE no higher than 1.8 and an MAE between 0.3 and 1. The coefficient of determination is between 0.82 and 0.98, and the correlation coefficient is between 0.91 and 0.99.

When it comes to important assessment measures such asS MSE (Fig. 3), MAE (Fig. 4), and R-squared, *Renesansa* performs admirably. Renesansa has an exceptionally low MSE (between 0.001 and 1.19), indicating small squared discrepancies between expected and actual values. This implies that the model offers extremely precise evaluations of the risk of COVID-19 in a nation, which is essential to inform public health initiatives and policy choices. In a similar vein, the model shows a low MAE, focusing on small changes between the predicted values and the actual values. This demonstrates how accurately Renesansa can estimate the risk variables for COVID-19, allowing policymakers and health authorities to make well-informed decisions. Renesansa also produces a high R-squared value (between 0.975 and 0.9997), suggesting that the model explains a substantial amount of the variability in

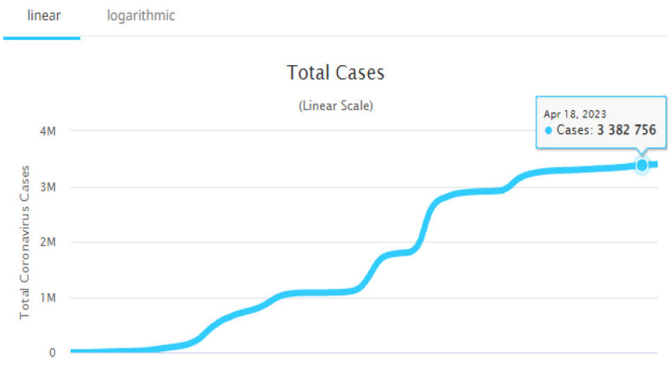

Fig. 5. Worldometer [27] value for Romania on 18th April 2023.

Fig. 6. Total cases value obtained by model.

COVID-19 risk variables.

This highlights the model's capacity to identify underlying trends and patterns in COVID-19 data, which is crucial for formulating workable plans to stop the virus's spread and handle public health emergencies. Renesansa's notable performance in evaluating the risk of COVID-19 for nations is highlighted by its remarkable results in MSE, MAE, and R-squared. While a high R-squared value indicates a great power of explanations, low MSE and MAE values indicate accurate predictions and minimum mistakes. These findings highlight the significance of sophisticated ensemble models such as Renesansa in directing evidence-based global responses to the COVID-19 epidemic.

The *Total Cases* value for Romania on 18 April 2023 on Worldometer [27] was 3,382,756. We can observe the result for Romania, obtained by model's prediction, which is very similar to the one from Worldometer (3,382,872). There is an extremely high performance of this model in the set of independent variable columns (compare the results of Fig. 5 and Fig. 6).

B. Binary Classification Considering the Proportion Between Total Cases and Population

The results obtained for the performance evaluation metrics of the models are strong (Fig. I). Performance evaluation metrics provide insight into the strengths and weaknesses of each model in the context of detecting critical geographic regions in the COVID-19 dataset. These metrics provide a comprehensive view of each model's performance. The best model for a specific use case might depend on the priorities: whether it is achieving high accuracy, minimizing false positives, maximizing sensitivity, or maintaining a balance between different metrics.

TABLE I
PERFORMANCE EVALUATION METRICS FOR MODELS IN TERMS OF COUNTRY'S COVID-19 RISK CLASSIFICATION.

	Sensitivity	**Specificity**	**Accuracy**	**Precision**
KNN	0.9286	0.9741	0.9606	0.9381
DT	0.8730	0.8427	0.8485	0.5670
SVM	0.9950	0.7614	0.7788	0.2474
CART	0.8846	0.8889	0.8879	0.7113
XGBoost	0.9970	0.9549	0.9667	0.8866
Renesansa	0.9975	0.9749	0.9818	0.9381

Important information is revealed by the evaluation results (see Table I) of several machine learning models used to group nations into COVID-19 risk categories. Countries, whose ratio between *Total Cases* and *Population* is above 0.003 (0.3%), are seen as areas at risk of infection. This value is calculated with Youden J Index technique. To determine a cutoff using the Youden's J Index, sensitivity and specificity of the diagnostic test are initially assessed. Subsequently, Youden's J Index is computed by summing the sensitivity and specificity, then subtracting one. Ultimately, the threshold value that yields the highest Youden's J Index is chosen as the cutoff, signifying the most favorable equilibrium between sensitivity and specificity in the diagnostic evaluation [28]. All metrics show that the ensemble learning model has the highest performance (99.75% sensitivity, 97.49% specificity, 98.18% accuracy, and 93.81% precision) demonstrating its reliability in differentiating between nations that are at risk and those that are not. This shows that the accuracy and dependability of the predictions can be improved by integrating various models. Notably, XGBoost also performs admirably, especially when it comes to sensitivity and specificity (99.7% and 95.49%), which are crucial for accurately identifying nations that are actually at danger while reducing false positives. SVM and decision tree models, on the other hand, show less accuracy (76.14% and 84.27%), suggesting a higher false alarm rate. In order to properly identify at-risk and non-at-risk nations, sensitivity and specificity are essential. These findings highlight how crucial trustworthy prediction models are in directing appropriate actions.

V. Conclusions

We find interesting paths to enhance classification accuracy and prediction performance by using ensemble learning approaches for COVID-19 prediction and classification tasks. Through the combination of a wide range of shallow models, such as DT, SVM, CART, KNN, and XGBoost, we have shown the ability to improve performance and mitigate the shortcomings of individual models on a variety of COVID-19-related datasets.

Our results highlight how crucial ensemble learning frameworks—like bagging, boosting, and stacking—are for efficiently combining predictions from several models to identify the intricate patterns and correlations present in COVID-19 data. We have demonstrated by thorough hyperparameter tuning, model validation, and interpretation analysis that ensemble learning models provide reliable solutions for this topic.

Our technique makes it easier to comprehend and evaluate model outputs by offering insights into the variables influencing COVID-19 predictions and classifications. We may improve real-time tracking, forecasting, and reaction efforts in the ongoing fight against the COVID-19 pandemic by integrating the ensemble model into decision support systems.

Acknowledgment

The author thanks the Computer Science Department of the West University of Timisoara for the support in terms of resources and some professors for the indicated suggestions.

References

[1] Ameer Sardar Kwekha-Rashid, Heamn N Abduljabbar, and Bilal Alhayani. Coronavirus disease (covid-19) cases analysis using machine-learning applications. *Applied Nanoscience*, 13(3), 2023. DOI: 10.1007/s13204-021-01868-7.

[2] Hafsa Bareen Syeda, Mahanazuddin Syed, Kevin Wayne Sexton, Shorabuddin Syed, Salma Begum, Farhanuddin Syed, Fred Prior, and Feliciano Yu Jr. Role of machine learning techniques to tackle the covid-19 crisis: systematic review. *JMIR medical informatics*, 9(1):e23811, 2021. DOI: 10.2196/23811.

[3] Sara Platto, Tongtong Xue, and Ernesto Carafoli. Covid19: an announced pandemic. *Cell Death & Disease*, 11(9):799, 2020. DOI: 10.1038/s41419-020-02995-9.

[4] Mustafa Hasöksüz, Selcuk Kilic, and Fahriye Saraç. Coronaviruses and sars-cov-2. *Turkish journal of medical sciences*, 50(9):549–556, 2020. DOI: 10.3906/sag-2004-127.

[5] World Health Organization et al. Coronavirus disease 2019 (covid-19): situation report, 116. 2020. DOI: 10.2139/ssrn.3566298.

[6] Marco Ciotti, Massimo Ciccozzi, Alessandro Terrinoni, Wen-Can Jiang, Cheng-Bin Wang, and Sergio Bernardini. The covid-19 pandemic. *Critical reviews in clinical laboratory sciences*, 57(6):365–388, 2020. DOI: 10.1080/10408363.2020.1783198.

[7] Rakesh Padhan and KP Prabheesh. The economics of covid-19 pandemic: A survey. *Economic analysis and policy*, 70:220–237, 2021. DOI: 10.1016/j.eap.2021.02.012.

[8] Walter Cullen, Gautam Gulati, and Brendan D Kelly. Mental health in the covid-19 pandemic. *QJM: An International Journal of Medicine*, 113(5):311–312, 2020. DOI: 10.1093/qjmed/hcaa110.

[9] Hua Ye, Peiliang Wu, Tianru Zhu, Zhongxiang Xiao, Xie Zhang, Long Zheng, Rongwei Zheng, Yangjie Sun, Weilong Zhou, Qinlei Fu, et al. Diagnosing coronavirus disease 2019 (covid-19): Efficient harris hawks-inspired fuzzy k-nearest neighbor prediction methods. *IEEE Access*, 9:17787–17802, 2021. DOI: 10.1109/access.2021.3052835.

[10] Ahmed Hamed, Ahmed Sobhy, and Hamed Nassar. Accurate classification of covid-19 based on incomplete heterogeneous data using a k nn variant algorithm. *Arabian Journal for Science and Engineering*, 46:8261–8272, 2021. DOI: 10.1007/s13369-020-05212-z.

[11] Mehmet Tahir Huyut and Hilal Üstündağ. Prediction of diagnosis and prognosis of covid-19 disease by blood gas parameters using decision trees machine learning model: a retrospective observational study. *Medical gas research*, 12(2):60–66, 2022. DOI: 10.4103/2045-9912.326002.

[12] Vijander Singh, Ramesh Chandra Poonia, Sandeep Kumar, Pranav Dass, Pankaj Agarwal, Vaibhav Bhatnagar, and Linesh Raja. Prediction of covid-19 corona virus pandemic based on time series data using support vector machine. *Journal of Discrete Mathematical Sciences and Cryptography*, 23(8):1583–1597, 2020. DOI: 10.1080/09720529.2020.1784535.

[13] Y Lebrini, A Boudhar, R Hadria, H Lionboui, L Elmansouri, R Arrach, P Ceccato, and T Benabdelouahab. Identifying agricultural systems using svm classification approach based on phenological metrics in a semi-arid region of morocco. *Earth Systems and Environment*, 3(2):277–288, 2019. DOI: 10.1007/s41748-019-00106-z.

[14] Sajja Tulasi Krishna and Hemantha Kumar Kalluri. Lung image classification to identify abnormal cells using radial basis kernel function of svm. In *Smart Technologies in Data Science and Communication: Proceedings of SMART-DSC 2019*, pages 279–285. Springer, 2020. DOI: 10.1007/978-981-15-2407-333.

[15] Leo Breiman, Jerome Friedman, Richard Olshen, and Charles Stone. Cart. *Classification and regression trees*, 1984. DOI: 10.1201/9781315139470-8.

[16] Richard K Zimmerman, Mary Patricia Nowalk, Todd Bear, Rachel Taber, Karen S Clarke, Theresa M Sax, Heather Eng, Lloyd G Clarke, and GK Balasubramani. Proposed clinical indicators for efficient screening and testing for covid-19 infection using classification and regression trees (cart) analysis. *Human Vaccines & Immunotherapeutics*, 17(4):1109–1112, 2021. DOI: 10.1080/21645515.2020.1822135.

[17] Sayato Fukui, Akihiro Inui, Takayuki Komatsu, Kanako Ogura, Yutaka Ozaki, Manabu Sugita, Mizue Saita, Daiki Kobayashi, and Toshio Naito. A predictive rule for covid-19 pneumonia among covid-19 patients: A classification and regression tree (cart) analysis model. *Cureus*, 15(9), 2023. DOI: 10.7759/cureus.45199.

[18] Tianqi Chen, Tong He, Michael Benesty, Vadim Khotilovich, Yuan Tang, Hyunsu Cho, Kailong Chen, Rory Mitchell, Ignacio Cano, Tianyi Zhou, et al. Xgboost: extreme gradient boosting. *R package version 0.4-2*, 1(4):1–4, 2015. DOI: 10.32614/cran.package.xgboost.

[19] Junling Luo, Zhongliang Zhang, Yao Fu, and Feng Rao. Time series prediction of covid-19 transmission in america using lstm and xgboost algorithms. *Results in Physics*, 27:104462, 2021. DOI: 10.1016/j.rinp.2021.104462.

[20] Edelson Damasceno Carvalho, Edson Damasceno Carvalho, Antonio Oseas de Carvalho Filho, Flávio Henrique Duarte de Araújo, and Ricardo de Andrade Lira Rabêlo. Diagnosis of covid-19 in ct image using cnn and xgboost. In *2020 IEEE Symposium on Computers and Communications (ISCC)*, pages 1–6. IEEE, 2020. DOI: 10.1109/iscc50000.2020.9219726.

[21] Zheng-gang Fang, Shu-qin Yang, Cai-xia Lv, Shu-yi An, and Wei Wu. Application of a data-driven xgboost model for the prediction of covid-19 in the usa: a time-series study. *BMJ open*, 12(7):e056685, 2022. DOI: 10.1136/bmjopen-2021-056685.

[22] Thomas G Dietterich et al. Ensemble learning. *The handbook of brain theory and neural networks*, 2(1):110–125, 2002. DOI: 10.7551/mitpress/3413.001.0001.

[23] Leo Breiman. Bagging predictors. *Machine learning*, 24:123–140, 1996. DOI: 10.1007/bf00058655.

[24] Robert E Schapire et al. A brief introduction to boosting. In *Ijcai*, volume 99, pages 1401–1406. Citeseer, 1999. DOI: 10.1007/3-540-49097-31.

[25] Kai Ming Ting and Ian H Witten. Stacking bagged and dagged models. 1997. DOI: 10.1109/icdm.2010.49.

[26] Covid-19 dataset. https://www.kaggle.com/datasets/selfishgene/covid19-worldometer-snapshots-since-april-18?resource=download, last accessed on 18 mar.

[27] Worldometer information about coronavirus. https://www.worldometers.info/coronavirus/, last accessed on 17 mar.

[28] Ronen Fluss, David Faraggi, and Benjamin Reiser. Estimation of the youden index and its associated cutoff point. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, 47(4):458–472, 2005. DOI: 10.1002/bimj.200410135.