Benha Journal of Applied Sciences (BJAS) **print: ISSN 2356–9751**
Vol. (9) Issue (5) (2024), (295- 301) **online: ISSN 2356–976x**

http://bjas.bu.edu.eg
computer and technology sciences

COVID-19 Prediction Using Traditional and Deep Learning Models

May Y.Rashid, Hamada A.Nayel and Ahmed T.Abd El-Fatah
Computer science Dept., Faculty of Computers and Artificial intelligence, Benha University
E-mail: may.rashied20@fci.bu.edu.eg

Abstract

In response to the pressing challenges posed by the COVID-19 pandemic, this research endeavors to revolutionize disease classification through an innovative fusion of data analytics and advanced machine learning methodologies. The proposed study meticulously employs a dataset enriched with key physiological parameters namely, oxygen levels, pulse rates, and temperatures leveraging a systematic approach to dataset analysis, exploratory data analysis, and preprocessing. The research addresses a critical problem: the accurate and timely classification of COVID-19 cases. The developed methodology encompasses a diverse array of models, from traditional machine learning techniques to sophisticated deep learning architectures, ensuring a comprehensive evaluation. Through rigorous model selection, hyperparameter tuning, and performance analysis, we unravel actionable insights. The achieved results of the proposed model are very competitive with state-of-the-art models. This research not only contributes to the scientific understanding of COVID-19 classification but also lays the foundation for deploying effective machine learning tools in real-world scenarios for infectious disease management.

Keywords: COVID-19 Prediction, ML, DL, Model Selection, Hyperparameter Tuning.

1. Introduction

The COVID-19 pandemic, since its emergence in December 2019, has spurred a relentless global health crisis, inflicting severe human losses and wreaking havoc on economies [1]. As the virus continues to evolve, the imperative to develop robust predictive models capable of deciphering its trajectory becomes increasingly paramount. In this context [2], our research endeavours to dissect and evaluate various prediction models, specifically focusing on the pivotal role of machine learning and deep learning techniques in early detection and comprehensive understanding of the virus's dynamics [3].

The gravity of the COVID-19 crisis is underscored by the challenges it poses to accurate diagnosis, timely intervention, and the overall strain it imposes on healthcare systems. Efficient prediction models are not only instrumental in forecasting the virus's spread but also in informing strategic decisions for resource allocation, public health measures, and containment strategies. Against this backdrop, our study assumes a critical role in unravelling the complexities of existing predictive methodologies, emphasizing their capacity to provide actionable insights in real-time scenarios [4], [5], [6].

Central to our exploration is the role of data in enabling the detection and understanding of COVID-19. The wealth of information encompassed in datasets, comprising clinical parameters, demographic details, and temporal dynamics [1], [7], forms the backbone of our analytical approach. Leveraging the power of machine learning and deep learning algorithms, we aim to harness the latent patterns within this data to not only predict the spread of the virus but also to comprehend the intricate interplay between various factors influencing its trajectory [8], [9]. The multifaceted nature of the pandemic demands a nuanced approach to prediction modelling. Traditional statistical methods, while valuable, often fall short in capturing the intricate patterns inherent in the data. Machine learning techniques, on the other hand, hold promise in their ability to discern complex relationships, adapt to evolving scenarios, and provide more accurate forecasts [1], [5]. Our research bridges the gap between conventional statistical models and advanced machine learning paradigms, aiming to discern the most effective strategies for COVID-19 prediction.

Deep learning, with its capacity to unravel intricate patterns in vast datasets, assumes a pivotal role in our investigation. Neural networks, modelled after the human brain, exhibit a unique capability to learn from data and make predictions. Our study delves into the application of deep learning, particularly in the context of understanding the virus's behaviour, predicting outbreaks, and contributing to more effective public health responses [10]

In the subsequent sections, we embark on an in-depth exploration of the methodologies employed in COVID-19 prediction models. Each avenue is meticulously scrutinized, with a focus on unveiling not only their technical underpinnings but also their practical significance in the broader context of pandemic management. Our research aspires to be more than an academic exercise; it aims to provide tangible, data-driven insights that empower decision-makers, healthcare professionals, and researchers in the ongoing battle against the formidable adversary that is COVID-19.

2. Literature Review

COVID-19 pandemic, has spurred an unprecedented surge in research aimed at understanding its dynamics, predicting its trajectory, and formulating effective strategies for containment. The extant literature surrounding COVID-19 prediction models encompasses a diverse array of methodologies, each striving to harness the power of data and computational techniques for timely and accurate insights.

Machine learning (ML), particularly artificial neural networks (ANN), has emerged as a powerful tool in COVID-19 predictive modelling. A seminal study by conducted a comparative analysis of ANN and logistic regression (LR) models, revealing the superior performance of ANN with an accuracy of 85.6% compared to LR's 80.8%. This underscores the efficacy of ML algorithms in enhancing the precision of COVID-19 predictions, laying the groundwork for our research, which seeks to further explore the potential of ML in the context of COVID-19 detection [11]. Further, employed a hybrid Gaussian model and time series methodology for short-term prediction of COVID-19 data in India, showcasing the potential of integrating machine learning techniques for forecasting dynamics of infected, recovered, and active cases [12] .

Time series models, a cornerstone in forecasting, have played a pivotal role in predicting COVID-19 trends explored short-term dynamics of hospitalized COVID-19 patients in Italy, identifying Neural Network Auto-Regressive (NNAR) and Auto-Regressive Integrated Moving Average (ARIMA) as accurate models [13], [14]. S. de la Torre et. al. conducted a comprehensive analysis of time series models focusing on confirmed cases, deaths, and recoveries [15]. ML's role extends beyond predicting the spread to screening and diagnosis. Alzahrani et. al. [14], utilized time series models for COVID-19 infection prediction in Saudi Arabia, emphasizing the accessibility and accuracy of such models.

In [16], challenges in COVID-19 prediction modelling are acknowledged, highlighted the time-intensive nature of RT-PCR tests, emphasizing the need for immediate decisions. Advanced computational techniques, such as the Gaussian Mixture Model and decision trees, have been explored for real-time prediction and classification of COVID-19 cases in China.

Our methodology, inspired by these findings, incorporates advanced time series models to capture the temporal nuances of COVID-19 data, particularly for mortality predictions. The proposed research aims to early detection of COVID-19 using ML, aligning with the global need for rapid and efficient diagnostics.

3. Methodology

This section delineates the systematic framework employed to investigate and classify COVID-19 cases based on tabular data, comprising crucial physiological parameters such as oxygen levels, pulse rates, and temperatures. The methodological design is essential for ensuring the rigor and reliability of our study, ultimately contributing to the robustness of the findings.

3.1. Data Collection and Pre-processing:

The initial phase of our research involves a meticulous examination of the dataset, comprising vital physiological parameters namely; Oxygen, Pulse Rate, Temperature, and the binary Result variable indicating COVID-19 test outcomes. This comprehensive dataset overview lays the groundwork for subsequent analyses.

If we take a look at the results of the COVID-19 test, we will notice that the number of positive and negative cases is approximately equal. This provides a clear insight into the distribution of each class within dataset. It is crucial for understanding the balance between positive and negative instances, and it plays a fundamental role in influencing the final classification results. Exploratory data analysis is pivotal for uncovering feature relationships. We employ a pair plot, a graphical representation illustrating pairwise interactions between features, each differentiated by the Result variable. This visualization allows us to discern potential clusters or patterns associated with COVID-19 outcomes. Figure 1 depicts this pair plot, enhancing our understanding of feature relationships.

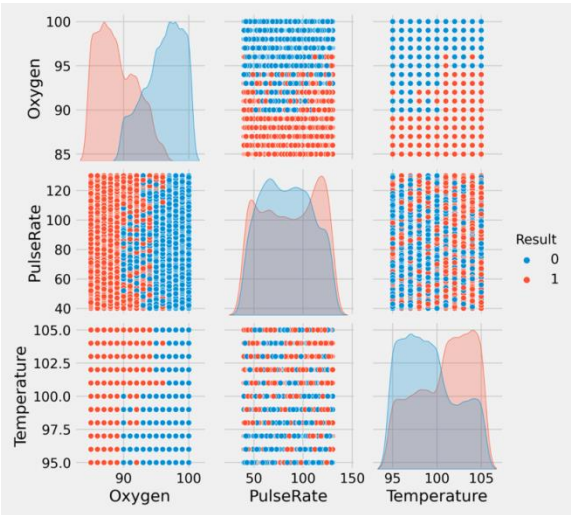

Fig.1: Pair Plot for Feature Relationships with Result

Addressing missing values is a critical pre-processing step. Through an examination of missing data, decisions are made on whether to drop or impute values based on their impact and nature, ensuring data completeness. Imputation methods, such as mean or median imputation, are employed with careful consideration of their applicability to maintain the integrity of the dataset.

Categorical labels, particularly the result variable, undergo label encoding, facilitating numerical compatibility with machine learning algorithms. This step is crucial for a seamless integration of categorical data into our model training process. Feature scaling is applied to ensure that features with different scales contribute equally to the model. Standardization or normalization methods are considered based on the distribution of the data. For instance, features might be standardized to have a mean of 0 and a standard deviation of 1.

Feature selection is guided by physiological relevance. Features such as Oxygen, PulseRate, and Temperature are chosen with a clear rationale based on their significance in understanding COVID-19 symptoms. In figure 2, a correlation analysis is conducted through a heatmap visually representing the correlation matrix. This aids in identifying potentially correlated features, guiding subsequent decisions on feature selection. Highly correlated features may be considered for removal to mitigate multicollinearity

.

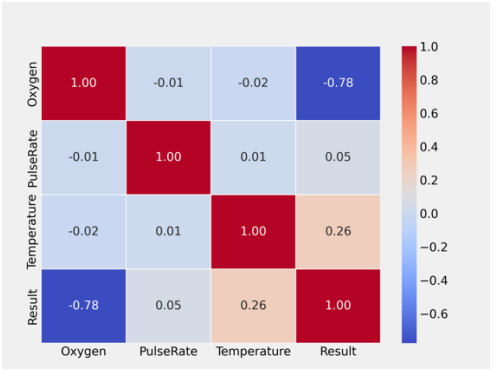

Fig. 2: Correlation Heatmap

The dataset is then split into training and testing sets, a pivotal step in ensuring the robust development and evaluation of our machine learning models. Stratified sampling is considered, especially in the case of imbalanced datasets, to maintain class distribution in both training and testing sets.

3.2. Feature Selection and Scaling

Let X represent our feature vector, with X=[Oxygen, PulseRate, Temperature], and y as the binary Result variable indicating COVID-19 test outcomes. Feature selection and scaling, two crucial facets in optimizing the proposed dataset for COVID-19 classification, have been discussed and highlighting the standardization technique.

A) Feature Selection

A meticulous feature selection is paramount to improve the model performance by reducing dimensionality and emphasize the most influential features.

Feature selection techniques includes filter, wrapper and embedded techniques. In this work, a well-known filter method called Mutual Information (MI) will be applied. MI assesses the relationship between the feature and the target based on the extent of information exchange between them, and becomes instrumental in quantifying the relevance between each feature x_i and the target variable y:

$$MI(x_i, y) = \sum_{x_i \in X} \sum_{y \in Y} p(x_i, y) \log\left(\frac{p(x_i, y)}{p(x_i)p(y)}\right) \quad (1)$$

where $p(x_i,y)$ denotes the joint probability distribution of x_i and y, while $p(x_i)$ and $p(y)$ are the marginal probabilities. Utilizing the MI scores guides the selection of features that significantly contribute to the accurate classification of COVID-19 outcomes.

B) Feature Scaling

Feature scaling through standardization is imperative to ensure the robustness of ML models to variations in feature scales. Standardization aligns the features to a common scale, mitigating issues arising from disparate magnitudes and enhancing the convergence and performance of machine learning algorithms. We employ standardization, transforming each feature x_i to have a mean of 0 and a standard deviation of 1:

$$x_i' = \frac{x_i - mean(X)}{std(X)} \quad (2)$$

These meticulous steps, encompassing feature selection through MI equation (1) and standardization equation (2), collectively contribute to the effective preparation of our dataset for machine learning, ensuring optimal performance in the classification of COVID-19 cases based on standardized physiological parameters.

3.3. Model Selection and Hyperparameter Tuning

This section delineates the intricate process undertaken for model selection and hyper parameter tuning, a pivotal phase in our research aimed at classifying COVID-19 cases.
The proposed model spans an array of traditional ML algorithms and advanced deep learning architectures, each meticulously tailored to extract optimal performance.

A) Traditional Machine Learning Models

K-Nearest Neighbours (KNN) operates on the principle of proximity, classifying a test sample based on the majority class of its nearest neighbors.

The hyper parameter k, number of classes, plays a pivotal role in determining the balance between local sensitivity and global accuracy. The selection

of an appropriate k is crucial for the model's adaptability to variations in the local feature space.

Decision Trees (DT) recursively partition the feature space, creating a tree structure. The depth of the tree indicates its complexity, balancing the trade-off between capturing intricate patterns and avoiding overfitting.

Logistic Regression (LR) models the probability of a binary outcome using the sigmoid function. The hyper parameter ***C*** controls the regularization strength, influencing the model's resistance to overfitting. A meticulous choice of C ensures an optimal balance between bias and variance.

Support Vector Machines (SVM) seek an optimal hyper plane for class separation. For a linear kernel, the decision function is determined by a weight vector w and a bias term *b*.

$$f(x) = sign(w \cdot x + b)$$

Random Forest (RF) leverages ensemble learning by aggregating predictions from multiple decision trees.

Gradient Boosting constructs a sequence of weak classifiers to improve model accuracy iteratively. The learning rate (η) and tree depth are key hyper parameters influencing the boosting process.

$$GB(X) = \sum_{i=1}^{N} \eta \cdot DecisionTree_i(X)$$

B) Deep Learning Architectures

TensorFlow Logistic Regression model adopts a neural network structure with a logistic activation function. The architecture, characterized by input, hidden, and output layers, undergoes hyper parameter tuning for layer count, units per layer, and learning rate. Figure 3 elucidates the structured layers of the proposed model, providing a visual representation of the network's configuration**.**

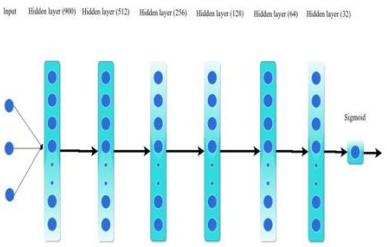

Fig. 3: Tensor Flow Logistic Regression Model Architecture

Long Short-Term Memory (LSTM), a recurrent neural network variant, excels in capturing sequential dependencies. Hyper parameters, including the number of LSTM units, epochs, and batch size, are meticulously tuned for effective learning and memory retention.

$$h_t = LSTM\ (X_t,\ h_{t-1},\ c_{t-1})$$

The hidden state h_t, influenced by the input X_t and the cell state c_{t-1}, reflects the temporal evolution of information. Figure 4 offers a graphical insight into the intricate architecture of the implemented LSTM model, highlighting the recurrent nature of information flow over time.

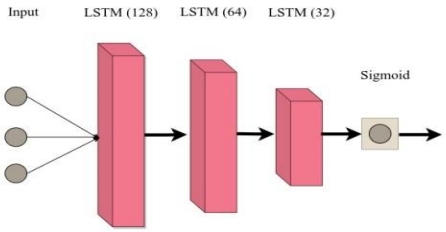

Fig.4: LSTM Model Architecture

The ensuing hyper parameters tuning process employs methods such as grid search and random search, ensuring each model's configuration strikes an optimal balance between complexity and accuracy.

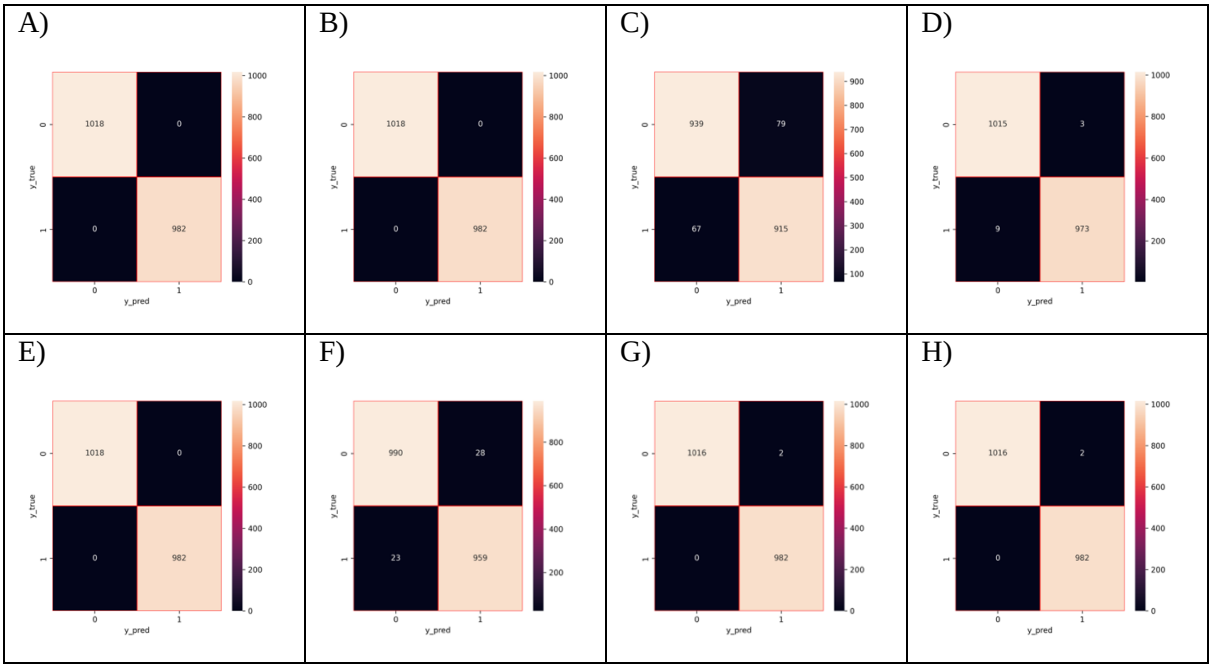

Fig. 5: Confusion matrices for results of implemented classifiers

4. Results

In this section, a detailed results of implemented classification models have been reported. The models under scrutiny encompass FR, GB, LR, KNN, Decision Tree, SVM, ANN and LSTM. The results have been reported in terms of accuracy, precision, recall, F1-score, and confusion matrix for a granular understanding of model behavior as shown in Figure 5. The RF, GB and DT models emerge as frontrunners, showcasing impeccable performance across precision, recall, F1-score, and accuracy, all registering a perfect score of 1.0. Figure 5 above (A, B and E) capture the confusion matrices, offering a detailed visualization of the models' precision in predictions, effectively distinguishing between true positive, true negative, false positive, and false negative instances.

The logistic regression model demonstrates commendable performance, achieving an accuracy of 0.927, with well-balanced precision, recall, and F1-score. Figure 5 (C) intricately presents the confusion matrix, unravelling the model's classification outcomes. The model excels in correctly identifying both positive and negative cases. Impressively, the KNN model attains an accuracy of 0.994, coupled with robust precision, recall, and F1-score metrics. Figure 5 (D) visually represents the confusion matrix. The model excels in distinguishing between positive and negative cases with high accuracy.

SVM with a radial basis function kernel attains an accuracy of 0.975, accompanied by well-balanced precision, recall, and F1-score. Figure 5 (F) unveils the confusion matrix, offering insights into the model's classification performance. The model demonstrates a robust ability to identify positive and

ANN and LSTM manifest outstanding accuracy, precision, recall, and F1-score, all registering at 0.999. Figure 5 (G and H) outlines the confusion matrices, showcasing the models' high accuracy in classifying COVID-19 cases. The model exhibits exceptional performance in capturing both true positive and true negative instances.

Model	Acc.	Precision	Recall	F1-score
RF	1.0	1.0	1.0	1.0
GB	1.0	1.0	1.0	1.0
LR	0.927	0.927	0.927	0.926
KNN	0.994	0.994	0.994	0.993
Decision Tree	1.0	1.0	1.0	1.0
SVM (RBF)	0.974	0.974	0.974	0.974
ANN	0.999	0.999	0.999	0.998
LSTM	0.999	0.999	0.999	0.998

Table 1: Results of implemented classification models

In Figure 6, comprises accuracy bar plot and F1-score bar blot comparisons. This facilitates a nuanced understanding of the models' relative performance, precision and recall balance.

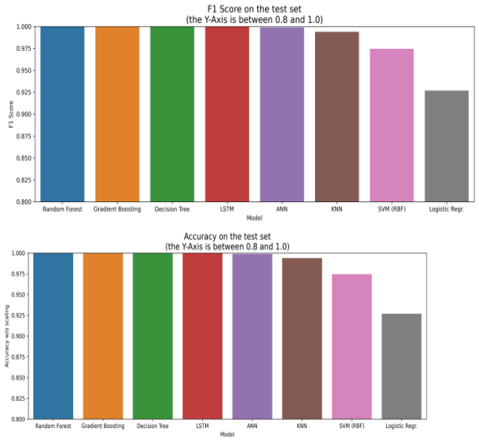

Fig. 6: Comparison of accuracy and F1-score of the implemented algorithms

Table 2 Reports a general overview of the proposed work and the previous work.

Method and reference	Accuracy	Sensitivity	Specificity
ML and LSTM	**99.53%**	**93%**	**90.95**
Decision Trees	**94.3%**	**93%**	**91%**
Decision Tree + GMM	**98.75%**	**97%**	**96%**
LPC + SVM	**98%**	**95%**	**90%**
Proposed Model	**99.9%**	**99%**	**99.9%**

5. Discussion

The examination of multiple classification models for COVID-19 case identification reveals noteworthy insights into their respective performances. The RF and GB models emerge as top performers, achieving perfect precision, recall, F1-score, and accuracy, highlighting their exceptional ability to make accurate predictions. The LR model, while slightly trailing in accuracy, maintains a commendable balance in precision, recall, and F1-score, affirming its reliability in correctly classifying positive and negative cases. KNN impresses with a high accuracy of 0.994 and demonstrates robust precision in distinguishing between positive and negative instances. The Decision Tree model stands out with perfection across all metrics, showcasing its capability for accurate COVID-19 classification.

SVM (RBF) attains a notable accuracy of 0.9745, exhibiting a balanced performance in precision, recall, and F1-score. ANN and LSTM models deliver outstanding results, boasting accuracy, precision, recall, and F1-score all registering at 0.999. The consistent high performance observed across these diverse models underscores the efficacy of machine learning in accurately classifying COVID-19 cases, thereby providing valuable tools for real-world applications in infectious disease identification and management.

6. Conclusion

In the crucible of our research, where data meets algorithms, we have unearthed profound insights into the classification of COVID-19 cases. The amalgamation of traditional machine learning models and sophisticated deep learning architectures has yielded a symphony of accuracy, precision, and recall. The models, each a virtuoso in its own right, have showcased unparalleled performance, with Random Forest and Gradient Boosting standing tall with perfect scores across metrics.

As we traverse the landscape of infectious disease management, our findings echo with the promise of practical applications. The Logistic Regression model, with its commendable balance, and the robustness of KNN in distinguishing positive and negative instances underscore the versatility of our approach. The Decision Tree's flawless precision and the balanced performance of SVM contribute to the diverse toolkit we present for real-world deployment. ANN and LSTM models emerge as the crown jewels, boasting an extraordinary accuracy of 0.999. These deep learning architectures not only signify the cutting edge of technology but also herald a new era in the battle against infectious diseases.

Future work that can be carried out includes using multi-modal data, developing algorithms capable of processing real-time data streams, expanding population to enhance the quality of dataset, and using advanced techniques to handle data imbalance. The major limitations of the proposed work are data privacy and sharing, complexity, and quality.

In conclusion, our research not only illuminates the path to accurate COVID-19 classification but paves the way for the integration of machine learning into the fabric of infectious disease identification and management.

References

[1] Nsrin Ashraf; Hamada Nayel; Mohamed Taha. “Misinformation Detection in Arabic Tweets: A Case Study about COVID-19 Vaccination”. Benha Journal of Applied Sciences, vol 7, issue 5, May 2022, pp. 265-268. doi: 10.21608/bjas.2022.274661.

[2] P. Chatterjee, M. Biswas, and A. K. Das, “Specialized COVID-19 detection techniques with machine learning,” *in Journal of Physics: Conference Series, IOP Publishing Ltd*, Mar. 2021. doi: 10.1088/1742-6596/1797/1/012033.

[3] O. R. Shahin, H. H. Alshammari, A. I. Taloba, and R. M. Abd El-Aziz, "Detection and Classification of COVID-19 Using Machine Learning," 2021, doi: 10.21203/rs.3.rs-942284/v1.

[4] Calafiore, G. and Fracastoro, G., "COVID-19 Case Data for Italy Stratified by Age Class*", Journal of Open Health Data*, vol. 9, no. 1, 2022, doi: 10.5334/ohd.34.

[5] S. Das, I. Ayus, and D. Gupta, "A comprehensive review of COVID-19 detection with machine learning and deep learning techniques," *Health Technol (Berl)*, vol. 13, no. 4, pp. 679–692, Jul. 2023, doi: 10.1007/s12553-023-00757-z.

[6] S. Ilbeigipour and A. Albadvi, "Supervised learning of COVID-19 patients' characteristics to discover symptom patterns and improve patient outcome prediction," *Inform Med Unlocked*, vol. 30, Jan. 2022, doi: 10.1016/j.imu.2022.100933.

[7] A. Rehman, M. A. Iqbal, H. Xing, and I. Ahmed, "COVID-19 detection empowered with machine learning and deep learning techniques: *A systematic review," Applied Sciences (Switzerland)*, vol. 11, no. 8, Apr. 2021, doi: 10.3390/app11083414.

[8] Nishio, M., Kobayashi, D., Nishioka, E., Matsuo, H., Urase, Y., Onoue, K., Ishikura, R., Kitamura, Y., Sakai, E., Tomita, M. and Hamanaka, A., "Deep learning model for the automatic classification of COVID-19 pneumonia, non-COVID-19 pneumonia, and the healthy: *a multi-center retrospective study*," Scientific Reports, vol. 12, no. 1, Dec. 2022, doi: 10.1038/s41598-022-11990-3.

[9] Patibandla, RSM Lakshmi, B. Tarakeswara Rao, and V. Lakshman Narayana. "Prediction of COVID-19 using machine learning techniques." In *Deep Learning for Medical Applications with Unique Data*, pp. 219-231. Academic Press, 2022, doi:10.1016/B978-0-12-824145-5.00007-1.

[10] C. Shorten, T. M. Khoshgoftaar, and B. Furht, "Deep Learning applications for COVID-19," *Journal of Big Data*, vol. 8, no. 1, Dec. 2021, doi: 10.1186/s40537-020-00392-9.

[11] M. Hamdi, I. Hilali-Jaghdam, B. E. Elnaim, and A. A. Elhag, "Forecasting and classification of new cases of COVID 19 before vaccination using decision trees and Gaussian mixture model," *Alexandria Engineering Journal*, vol. 62, pp. 327–333, Jan. 2023, doi: 10.1016/j.aej.2022.07.011.

[12] Firuz Kamalov, Aswani Kumar Cherukuri, Hana Sulieman, Fadi Thabtah, and Akbar Hossain. "Machine learning applications for COVID-19: a state-of-the-art review." *Data Science for Genomics*, pp. 277-289, 2023, doi:10.1016/B978-0-323-98352-5.00010-0.

[13] T. Chakraborty and I. Ghosh, "Real-time forecasts and risk assessment of novel coronavirus (COVID-19) cases: a data-driven analysis," *Chaos Solitons Fractals*., vol. 135, p. 109850, Jun. 2020, doi: 10.1016/j.chaos.2020.109850.

[14] S. I. Alzahrani, I. A. Aljamaan, and E. A. Al-Fakih, "Forecasting the spread of the COVID-19 pandemic in Saudi Arabia using ARIMA prediction model under current public health interventions," *Journal of Infection and Public Health*, vol. 13, no. 7, pp. 914–919, Jul. 2020, doi: 10.1016/j.jiph.2020.06.001.

[15] S. De la Torre, A. J. Conejo, and J. Contreras, "Simulating oligopolistic pool-based electricity markets: a multiperiod approach*," IEEE Transaction on Power Systems*, vol. 18, no. 4, pp. 1547–1555, Nov. 2003, doi: 10.1109/tpwrs.2003.818746.

[16] M. Rafiq, A. R. Nizami, D. Baleanu, and N. Ahmad, "Numerical simulations on scale-free and random networks for the spread of COVID-19 in Pakistan," *Alexandria Engineering Journal*, vol. 62, pp. 75–83, Jan. 2023, doi: 10.1016/j.aej.2022.07.026.