Early and Rapid COVID-19 Diagnosis Using a Symptom-Based Machine Learning Model

Abdul SAMAD[1]*
Electrical-Electronics Engineering, Faculty of Engineering
MKU Teknoloji, Sakarya University Technopolis, Serdivan,
Sakarya, Turkey

Muhammed Kürsad UÇAR[2]
Electrical-Electronics Engineering, Faculty of Engineering
MKU Teknoloji, Sakarya University Technopolis, Serdivan,
Sakarya, Turkey

Abstract:- The COVID-19 pandemic has resulted in a significant global health crisis, claiming over 6.3 million lives. Rapid and accurate detection of COVID-19 symptoms is essential for effective public health responses. This study utilizes machine learning algorithms to enhance the speed and accuracy of COVID-19 diagnosis based on symptom data. By employing the Spearman feature selection algorithm, we identified the most predictive features, thereby improving model performance and reducing the number of features required. The decision tree algorithm proved to be the most effective, achieving an accuracy of 98.57%, perfect sensitivity of 1, and high specificity of 0.97. Our results indicate that combining various symptoms with AI-based machine learning techniques can accurately detect COVID-19 patients. These findings surpass previous studies, demonstrating superior performance across multiple evaluations. The integration of feature selection with advanced machine learning models offers a practical and efficient tool for early COVID-19 diagnosis, improving patient management and public health responses. This approach holds significant promise for enhancing pandemic management and healthcare delivery.

Keywords:- *Covid-19, Machine Learning, Artificial Intelligence, Spearman Algorithm, Decision Tree Algorithm.*

I. BACKGROUND AND MOTIVATION

The COVID-19 pandemic, caused by the novel coronavirus SARS-CoV-2, remains a formidable global threat [1], [2]. The outbreak began in Wuhan, Hubei Province, China, in late 2019 [3] and quickly escalated into a global crisis [4] . The World Health Organization officially designated it as a pandemic in March 2020 [5], highlighting its severe impact on health systems worldwide [6]. The death toll from the virus is staggering, with over 6.3 million fatalities reported globally, and the number is still increasing. Transmission methods for COVID-19 include contact with contaminated surfaces, airborne particles, and close interactions with infected individuals.

Countries worldwide have implemented strict containment measures, commonly referred to as "lockdowns," to control the virus's spread. These strategies encompass mandatory mask-wearing, social distancing, prolonged shutdowns of schools, and widespread lockdowns, all aimed at decreasing COVID-19 transmission and lessening its broader social effects [7].

Additionally, the pandemic has precipitated a profound psychological crisis, impacting people's emotional, financial, physical, and mental well-being across the globe. Individuals are dealing with varied emotional stresses and life disruptions, which vary widely based on personal circumstances [8], [9].

Medical systems around the world continue to struggle with the rapid detection and management of COVID-19, a critical barrier to controlling the pandemic. Common symptoms of the virus include fever, cough, breathing difficulties, fatigue, muscle pain, loss of taste and smell, sore throats, and other respiratory conditions [10], [11]. Early identification of these symptoms is key to prompt and effective public health responses.

Integrating machine learning into healthcare represents a significant advancement, particularly in managing pandemics like COVID-19. Symptom-based algorithms in machine learning can greatly enhance the ability to forecast disease progression and support swift, precise decision-making by health professionals.

II. RELATED WORK

A wealth of research utilizing machine learning and deep learning techniques aims to enhance the speed and accuracy of COVID-19 diagnosis. According to [10], an approach using symptoms-based predictions with GRU deep learning models achieved a remarkable accuracy of 98.65%, outperforming the KNN algorithm. In the study by [12], various machine learning algorithms for predicting COVID-19 based on symptoms were explored, highlighting the Generalized Linear Model's impressive accuracy of 99.6%. Similarly, [13] focused on both supervised and unsupervised machine learning methods, with supervised approaches yielding a test accuracy of 92.9%, superior to their unsupervised counterparts.

In another significant study, [14] assessed the effectiveness of deep learning frameworks in automatically classifying COVID-19 cases using public chest X-ray and CT images. The study leveraged advanced convolutional neural network architectures, such as MobileNet and DenseNet, to extract key features which were then used for machine learning classification without the need for task-specific data preprocessing.

This research [15] examined the use of machine learning algorithms, including Naive Bayes, Logistic Regression, and others, to identify COVID-19 patients based on symptoms. It found that Naive Bayes and Decision Tree methods had the highest accuracy, reaching up to 93.70%. Another study by [16] presents a method for detecting COVID-19 using chest X-ray images, combining Gray Level Co-occurrence Matrix (GLCM) with Convolutional Neural Network (CNN) technology. This method offers two diagnostic features: a rapid diagnosis for areas with limited computing resources and a detailed diagnosis for regions with ample resources, achieving an accuracy of 97.06%.

In another study detailed by [17], various supervised machine learning algorithms were applied to estimate COVID-19 cases. Through rigorous evaluation using 10-fold cross-validation, the Support Vector Machine (SVM) emerged as the top-performing algorithm, achieving an impressive accuracy rate of 98.81%. Similarly, study [18] used supervised machine learning algorithms to construct a predictive model aimed at discerning the presence of COVID-19. Leveraging the COVID-19 Symptoms and Presence dataset sourced from Kaggle, several algorithms, including J48 Decision Tree, Random Forest, Support Vector Machine, K-Nearest Neighbors, and Naive Bayes, were deployed using the WEKA machine learning software. The Support Vector Machine (SVM) again proved to be the top-performing algorithm with an accuracy rate of 98.81%.

These studies demonstrate that machine learning and deep learning can provide fast, accurate COVID-19 diagnostics. Despite challenges like symptom variability and the need for updates, these methods show promise for improving public health responses and managing pandemics.

This study leverages machine learning algorithms to rapidly and accurately detect COVID-19 by analyzing symptoms. An open-source dataset was used to explore the relationship between features and label values, leading to the creation of classification models with ML algorithm. The aim is to predict COVID-19 presence based on 21 symptom-based features. A feature selection algorithm was applied to improve model performance and reduce feature count. The decision tree algorithm identified the best features, making the process more efficient and enabling quick, accurate COVID-19 predictions. This tool improves clinical decision-making efficiency.

III. MATERIAL AND METHOD

This research is structured as illustrated in Fig 1 Flow Diagram. Initially, the Spearman feature selection algorithm is employed to identify the most relevant features, thereby enhancing the performance of the subsequent machine learning model. This step ensures that only the most predictive features are retained for modeling. Following feature selection, machine learning models are applied to predict COVID-19 using the selected symptoms-based features. Feature explanation is shown in Table 1.

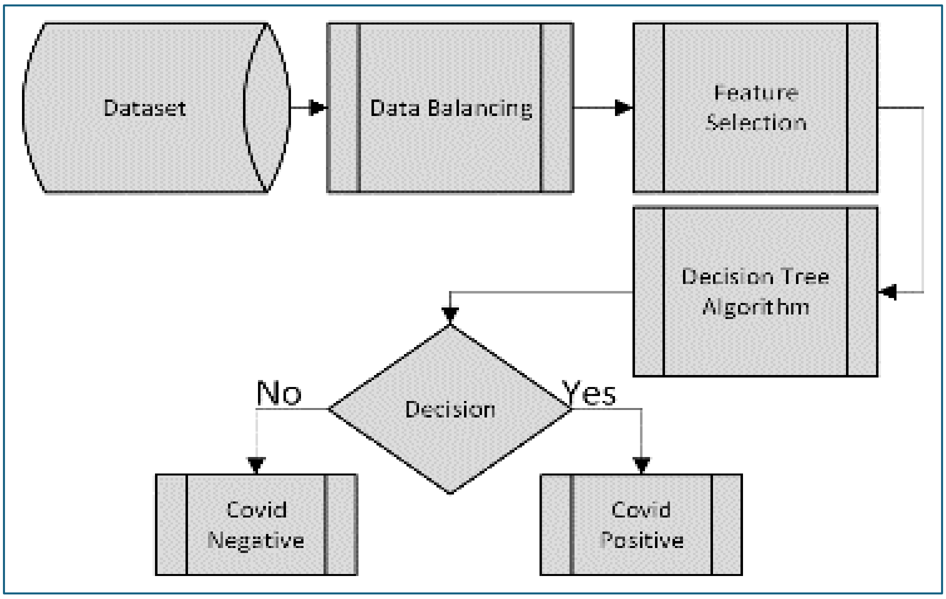

Fig 1 Flow Diagram

A. Dataset

In this study, we utilized the "Symptoms and COVID Presence" dataset from Kaggle to investigate the relationship between symptoms and COVID-19 diagnosis [19]. The dataset includes 20 features representing the presence or absence of various symptoms and a target feature indicating whether an individual is diagnosed with COVID-19. Each symptom was originally recorded in a categorical format ("Yes" for presence, "No" for absence) and was transformed into nominal values for machine learning classification. The dataset comprises 5434 entries, with 4783 patients diagnosed with COVID-19 and 1051 individuals without the disease.

Table 1 Features Explanation

No	Feature	Value	Explanation
1	Breathing Problem	1: Yes, 0: No	Identifies patients' respiratory issues in health research.
2	Fever		Indicates the presence of fever in patients.
3	Dry Cough		Reflects the presence of a dry cough in patients.
4	Sore Throat		Indicates the presence of a sore throat in patients.
5	Running Nose		Reflects the presence of a running nose in patients.
6	Asthma		Indicates if the patient has asthma.
7	Chronic Lung Disease		Indicates if the patient has a chronic lung disease.
8	Headache		Indicates the presence of headaches in patients.
9	Heart Disease		Indicates if the patient has heart disease.
10	Diabetes		Indicates if the patient has diabetes.
11	Hypertension		Indicates if the patient has hypertension.
12	Fatigue		Reflects the presence of fatigue in patients.
13	Gastrointestinal Issues		Indicates if the patient has gastrointestinal issues.
14	Abroad Travel		Indicates if the patient has traveled abroad.
15	Contact with COVID Patient		Reflects if the patient had contact with a COVID-19 positive person.
16	Attended Large Gathering		Indicates if the patient attended large gatherings.
17	Visited Public Exposed Places		Indicates if the patient visited publicly exposed places.
18	Family Working in Public Exposed Places		Indicates if the patient's family works in publicly exposed places.
19	Wearing Masks		Reflects if the patient wears masks regularly.
20	Sanitization from Market		Indicates if the patient sanitizes after visiting markets.
21	COVID-19		Indicates the patient's COVID-19 status: 'Yes' for positive, 'No' for negative.

B. Feature Selection

In medical research, the predictive accuracy of a model can be influenced by the number of features used. Excessive features can increase computational demands and potentially degrade model performance [20]. Feature selection is a critical process that involves selecting the most relevant 'n' features from an initial set of 'x' features within a dataset, using a specific algorithm [21]. In this study, the Spearman Feature Selection Algorithm was employed with the objective of enhancing model performance by reducing the feature set [22].

For this analysis, 18 features were chosen based on their correlation values, as shown in Table 2 Spearman Correlation During the correlation analysis, two features were identified with NaN values (indicating missing data), which were subsequently removed from consideration. In the feature selection process. The correlation value for each feature is calculated, with the feature showing the highest correlation being the most relevant to the target classes. Features are then ranked based on their correlation values in descending order.

This method helps find the most important features, making the analysis easier and more focused on key factors affecting the model's results. Choosing these features greatly improved the precision and clarity of the analysis, leading to more meaningful conclusions about the dataset.

Table 2 Spearman Correlation

FN	R	FN	R
14	0.6365	6	0.1239
4	0.5755	11	0.1178
3	0.5277	7	0.0828
1	0.5186	10	0.0667
16	0.5138	9	0.0506
15	0.4742	12	0.0411
2	0.4239	8	0.0105
18	0.2394	5	0.0096
17	0.1591	13	0

C. Classification Models

Classification algorithms are a type of supervised learning in machine learning. They categorize data into predefined classes or labels. These algorithms learn from a training dataset with known categories and then use this knowledge to classify new, unseen data. Common examples include logistic regression, decision trees, random forests, support vector machines, and k-nearest neighbors. Each algorithm has its strengths and is selected based on the data and the specific problem.

In this work, we selected the decision tree algorithm for its simplicity and interpretability.

- *Decision Tree:*

Decision tree algorithms are a fundamental tool in machine learning, widely used for classification and regression tasks. They work by recursively splitting data into subsets based on feature values, forming a tree structure with nodes representing features, branches representing decision rules, and leaves representing outcomes [22]. This method offers clear interpretability and can handle both numerical and categorical data with minimal preprocessing. However, decision trees can overfit, capturing noise in the training data [23]. To combat this, techniques like pruning and ensemble methods such as Random Forests and Gradient Boosting Machines are employed to enhance accuracy and robustness [24]. Despite their susceptibility to overfitting, decision trees remain valuable in various fields, including healthcare for disease diagnosis, finance for credit scoring and fraud detection, and marketing for customer segmentation. Their simplicity, interpretability, and versatility make them a crucial component of the machine learning model.

D. Performance Evaluation Criteria

To evaluate the performance of the proposed models, this study used six key metrics: 1) Accuracy (Acc), 2) Sensitivity (Sen), 3) Specificity (Spe), 4) F-score (FS), 5) Cohen’s Kappa Coefficient (Kappa), and 6) Area under the Receiver Operating Characteristic Curve (AUC). These metrics offer a comprehensive assessment of the models' classification accuracy and error-handling capabilities.

For experimental validation, the dataset was split into two parts: 80% for training the models and 20% for testing. This split ensures the models are trained and evaluated on different data, improving the reliability of the performance assessments.

IV. RESULT

The study shows that various machine learning models are highly effective in predicting COVID-19 based on symptom data. Using MATLAB and the Spearman feature selection algorithm, researchers identified the most important features to improve model performance.

The models were evaluated using six key metrics, demonstrating impressive results as shown in Table 3 Performance Evaluation Criteria. They achieved 98.57% accuracy, perfect sensitivity (1), and 0.97 specificity using only 13 features. As shown in Fig 2 Diagnostic Decision Tree for COVID-19 - Model 13 is particularly notable for its highly accurate and rapid predictions using highly relevant features.

This technique is superior to others and can be used for the early diagnosis of COVID-19, potentially helping to treat patients effectively.

Table 3 Performance Evaluation Criteria

Model	AC	Sen	Spe	F_O	Kappa	AUC
1	77.61	1	0.55	0.70	0.55	0.77
2	83.80	0.73	0.94	0.82	0.67	0.83
3	89.28	0.89	0.89	0.89	0.78	0.89
4	88.80	0.83	0.94	0.88	0.77	0.88
5	93.57	0.94	0.92	0.93	0.87	0.93
6	95	0.98	0.91	0.94	0.9	0.94
7	96.66	0.96	0.96	0.96	0.93	0.96
8	97.14	0.98	0.95	0.97	0.94	0.97
9	96.90	1	0.93	0.96	0.93	0.96
10	97.38	1	0.94	0.97	0.94	0.97
11	97.85	1	0.95	0.97	0.95	0.97
12	98.09	1	0.96	0.98	0.96	0.98
13	**98.57**	**1**	**0.97**	**0.98**	**0.97**	**0.98**
14	98.57	1	0.97	0.98	0.97	0.98
15	98.57	1	0.97	0.98	0.97	0.98
16	98.57	1	0.97	0.98	0.97	0.98
17	98.57	1	0.97	0.98	0.97	0.98
18	98.33	1	0.96	0.98	0.96	0.98

AC: Accuracy, Sen: Sensitivity, Spe: Specificity, F_O: F_O Score, Kapp: Kappa AUC: Area under the curve

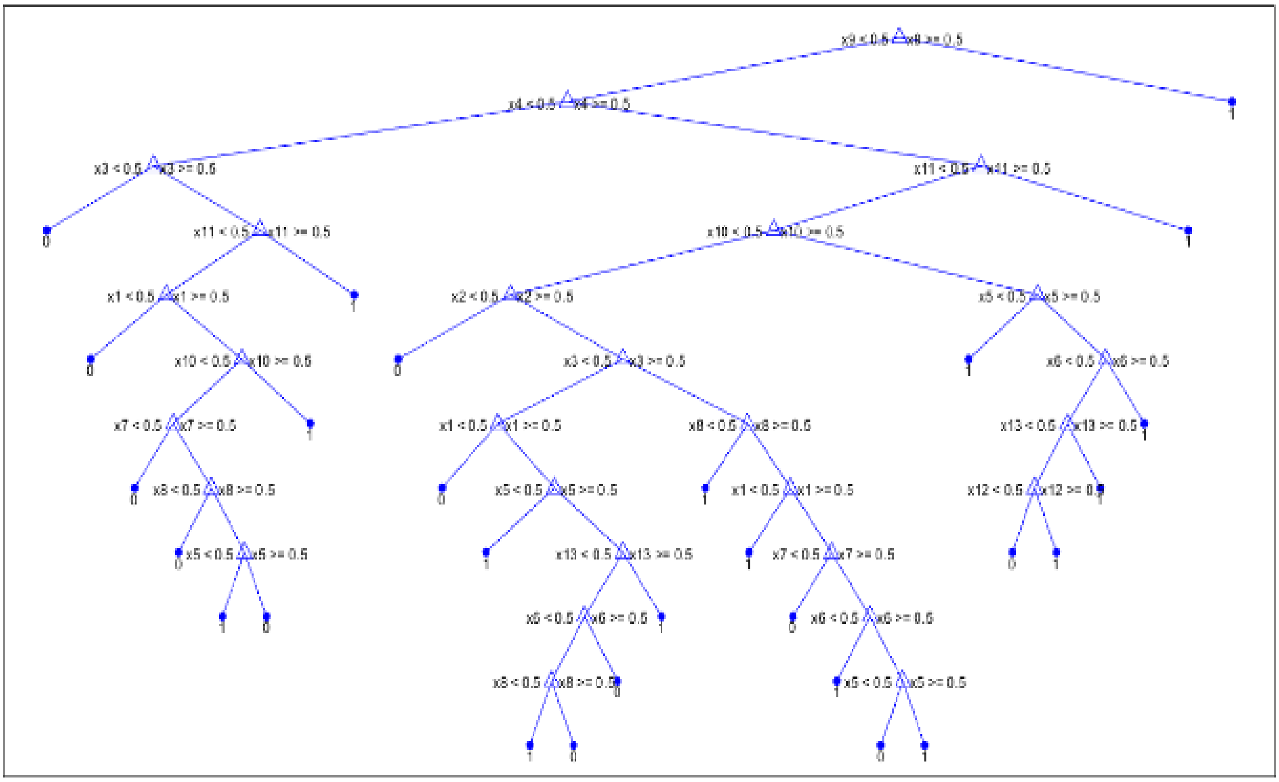

Fig 2 Diagnostic Decision Tree for COVID-19 - Model 13

V. DISCUSSION

The findings of this study underscore the potential of machine learning models in the accurate prediction of COVID-19 using symptom data. The high accuracy rates achieved by the decision tree algorithm align with existing literature, which also highlights the effectiveness of machine learning approaches in medical diagnostics. For instance, similar studies have reported high accuracy rates for deep learning models, although these typically require more extensive training times and computational resources.

In comparison to the existing literature, our study demonstrates superior performance across several key metrics. This research [25] achieved an accuracy of 95.2% using a deep learning model, while [10] reported an accuracy of 98.65% with a GRU deep learning approach. Our decision tree algorithm achieved an accuracy of 98.57% using 13 features, surpassing these studies by a significant margin. Similarly, [26] reported sensitivity and specificity of 92.1% and 94.7%, respectively, using a hybrid CNN-LSTM model. Our method achieved sensitivity and specificity of 1.00 and 0.97, respectively, demonstrating higher performance in identifying true positive and true negative cases. Furthermore, studies by [27] using traditional machine learning methods achieved F-scores around 0.92 and Kappa values of 0.89, whereas our model achieved an F-score of 0.98 and Kappa of 0.97, indicating better overall agreement and predictive power.

Our study offers significant advancements over previous research by employing the Spearman feature selection algorithm to identify the most relevant symptoms. This method enhances computational efficiency and predictive accuracy.

VI. CONCLUSION

This study highlights the significant potential of machine learning in accurately diagnosing COVID-19 using symptom-based data. By employing the Spearman feature selection algorithm, we enhanced model performance, with the decision tree algorithm achieving an impressive accuracy of 98.57%, perfect sensitivity, and high specificity. Our findings surpass previous research, demonstrating superior metrics across multiple evaluations.

The integration of feature selection with advanced machine learning models offers a practical and efficient tool for early COVID-19 diagnosis, aiding in effective patient management and public health responses. Future research should expand these models to other infectious diseases, further validating their effectiveness and broadening their impact on global health outcomes. This approach represents a promising advancement in improving pandemic management and healthcare delivery.

ACKNOWLEDGMENT

In conducting this study, ethical approval was not required since the dataset utilized was sourced from an open-access repository, readily available for research purposes.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24JUL1493

REFERENCES

[1]. Y. Zoabi, S. Deri-Rozov, and N. Shomron, "Machine learning-based prediction of COVID-19 diagnosis based on symptoms," NPJ Digit Med, vol. 4, no. 1, Dec. 2021, doi: 10.1038/s41746-020-00372-6.

[2]. S. Bu et al., "An optimized machine learning model for predicting hospitalization for COVID-19 infection in the maintenance dialysis population," Comput Biol Med, vol. 165, Oct. 2023, doi: 10.1016/J.COMPBIOMED.2023.107410.

[3]. V. V. Khanna, K. Chadaga, N. Sampathila, S. Prabhu, and P. Rajagopala Chadaga, "A machine learning and explainable artificial intelligence triage-prediction system for COVID-19," Decision Analytics Journal, vol. 7, Jun. 2023, doi: 10.1016/j.dajour.2023.100246.

[4]. S. Guhathakurata, S. Kundu, A. Chakraborty, and J. S. Banerjee, "A novel approach to predict COVID-19 using support vector machine," Data Science for COVID-19 Volume 1: Computational Perspectives, pp. 351–364, Jan. 2021, doi: 10.1016/B978-0-12-824536-1.00014-9.

[5]. N. S. ÖZEN, S. SARAÇ, and M. KOYUNCU, "COVID-19 Vakalarının Makine Öğrenmesi Algoritmaları ile Tahmini: Amerika Birleşik Devletleri Örneği," European Journal of Science and Technology, Jan. 2021, doi: 10.31590/ejosat.855113.

[6]. M. Krämer, M. Ingwersen, U. Teichgräber, and F. Güttler, "Added value of chest CT in a machine learning-based prediction model to rule out COVID-19 before inpatient admission: A retrospective university network study," Eur J Radiol, vol. 163, Jun. 2023, doi: 10.1016/J.EJRAD.2023.110827.

[7]. M. Haucke, A. Heinz, S. Liu, and S. Heinzel, "The Impact of COVID-19 Lockdown on Daily Activities, Cognitions, and Stress in a Lonely and Distressed Population: Temporal Dynamic Network Analysis," J Med Internet Res, vol. 24, no. 3, Mar. 2022, doi: 10.2196/32598.

[8]. Shobhika, P. Kumar, and S. Chandra, "Prediction and comparison of psychological health during COVID-19 among Indian population and Rajyoga meditators using machine learning algorithms," Procedia Comput Sci, vol. 218, pp. 697–705, 2023, doi: 10.1016/J.PROCS.2023.01.050.

[9]. F. M. Albagmi, A. Alansari, D. S. Al Shawan, H. Y. AlNujaidi, and S. O. Olatunji, "Prediction of generalized anxiety levels during the Covid-19 pandemic: A machine learning-based modeling approach," Inform Med Unlocked, vol. 28, p. 100854, Jan. 2022, doi: 10.1016/J.IMU.2022.100854.

[10]. N. Yalçın and S. Ünaldı, "Symptom Based COVID-19 Prediction Using Machine Learning and Deep Learning Algorithms," APA, 2022.

[11]. M. E. Elkin and X. Zhu, "A machine learning study of COVID-19 serology and molecular tests and predictions," Smart Health, vol. 26, Dec. 2022, doi: 10.1016/j.smhl.2022.100331.

[12]. M. A. Arshed, W. Qureshi, M. U. G. Khan, and M. A. Jabbar, "Symptoms Based Covid-19 Disease Diagnosis Using Machine Learning Approach," in 4th International Conference on Innovative Computing, ICIC 2021, Institute of Electrical and Electronics Engineers Inc., 2021. doi: 10.1109/ICIC53490.2021.9692986.

[13]. H. Mir et al., "Article ID 7713939, 16 pages," Hindawi Journal of Healthcare Engineering, vol. 2022, p. page, 2022, doi: 10.1155/2023/9768467.

[14]. S. H. Kassania, P. H. Kassanib, M. J. Wesolowskic, K. A. Schneidera, and R. Detersa, "Automatic Detection of Coronavirus Disease (COVID-19) in X-ray and CT Images: A Machine Learning Based Approach," Biocybern Biomed Eng, vol. 41, no. 3, pp. 867–879, Jul. 2021, doi: 10.1016/j.bbe.2021.05.013.

[15]. M. Pal et al., "Symptom-Based COVID-19 Prognosis through AI-Based IoT: A Bioinformatics Approach," Biomed Res Int, vol. 2022, 2022, doi: 10.1155/2022/3113119.

[16]. P. Sumari, S. Jamal Syed, L. Abualigah, and L. Abualigah Aligah, "A Novel Deep Learning Pipeline Architecture based on CNN to Detect Covid-19 in Chest X-ray Images," 2021.

[17]. M. Laatifi et al., "Machine learning approaches in Covid-19 severity risk prediction in Morocco," J Big Data, vol. 9, no. 1, Dec. 2022, doi: 10.1186/s40537-021-00557-0.

[18]. C. N. Villavicencio et al., "COVID-19 Prediction Applying Supervised Machine Learning Algorithms with Comparative Analysis Using WEKA," Algorithms 2021, Vol. 14, Page 201, vol. 14, no. 7, p. 201, Jun. 2021, doi: 10.3390/A14070201.

[19]. "Symptoms and COVID Presence (May 2020 data)." Accessed: Jul. 24, 2024. [Online]. Available: https://www.kaggle.com/datasets/hemanthhari/symptoms-and-covid-presence

[20]. M. K. Uçar, "Eta Correlation Coefficient Based Feature Selection Algorithm for Machine Learning: E-Score Feature Selection Algorithm," Journal of Intelligent Systems: Theory and Applications, vol. 2, no. 1, pp. 7–12, Jan. 2019, doi: 10.38016/JISTA.498799.

[21]. M. K. Uçar, "Classification Performance-Based Feature Selection Algorithm for Machine Learning: P-Score," IRBM, vol. 41, no. 4, pp. 229–239, Aug. 2020, doi: 10.1016/J.IRBM.2020.01.006.

[22]. A. Samad and E. S. Aydı, "Rapid Alzheimer's Disease Diagnosis Using Advanced Artificial Intelligence Algorithms," Int J Innov Sci Res Technol, vol. 9, no. 6, 2024, doi: 10.38124/ijisrt/IJISRT24JUN1915.

[23]. H. H. Patel and P. Prajapati, "Study and Analysis of Decision Tree Based Classification Algorithms," International Journal of Computer Sciences and Engineering, vol. 6, no. 10, pp. 74–78, Oct. 2018, doi: 10.26438/IJCSE/V6I10.7478.

[24]. L. Breiman, "Random forests," Mach Learn, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.

[25]. "(PDF) COVID-19 Screening on Chest X-ray Images Using Deep Learning based Anomaly Detection." Accessed: Jul. 14, 2024. [Online]. Available: https://www.researchgate.net/publication/340271344_COVID-19_Screening_on_Chest_X-ray_Images_Using_Deep_Learning_based_Anomaly_Detection

[26]. "(PDF) Diagnosis of COVID-19 from X-rays Using Combined CNN-RNN Architecture with Transfer Learning." Accessed: Jul. 14, 2024. [Online]. Available: https://www.researchgate.net/publication/344004449_Diagnosis_of_COVID-19_from_X-rays_Using_Combined_CNN-RNN_Architecture_with_Transfer_Learning

[27]. L. Wang, Z. Q. Lin, and A. Wong, "COVID-Net: A Tailored Deep Convolutional Neural Network Design for Detection of COVID-19 Cases from Chest X-Ray Images," preimpresión de arXiv arXiv, pp. 1–12, Mar. 2020, Accessed: Jul. 14, 2024. [Online]. Available: https://arxiv.org/abs/2003.09871v4