International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Neural & Statistical Models for COVID-19 Prediction and Detection

Chippada Surya Sai Prakesh[1]*,Edupuganti Rukhmini[2]*,Vemavarapu Meghana[*3],Nunna Tarakamanjunadha[4]*,Yarlagadda Rushyendra Chowdary[5*],Dr. Sreepada Sarada[6]

Email: chippadasaisuryaprakash@gmail.com [1], rukhmini.edupuganti@gmail.com[2], meghanavemavarapu9@gmail.com[3], narayangreeshmanth@gmail.com[4], yarlagaddareshmachowdary302@gmail.com[5],sarada.s@pragati.ac.in[6]
Pragati Engineering College, Surampalem, Kakinada.Dist, A.P-533437

ABSTRACT :

The COVID-19 pandemic has introduced significant diagnostic challenges, particularly due to the similarity of symptoms with other diseases like pneumonia. Traditionally, RT-PCR tests have been employed, but these methods are time-consuming and dependent on kit availability. This study proposes two distinct approaches to enhance COVID-19 detection: (1) a deep learning-based Convolutional Neural Network (CNN) method leveraging chest X-rays, and (2) a traditional machine learning pipeline utilizing patient symptom and history data. In the first approach, the CNN model achieves an accuracy of 97% using chest X-ray images, emphasizing feature extraction from radiological data to identify COVID-19 cases. Techniques such as the Canny edge detector and Grad-CAM are employed to increase model interpretability and aid medical professionals in decision-making. In the second approach, the machine learning models—Logistic Regression, K-Nearest Neighbours (KNN), and Random Forest—predict COVID-19 likelihood based on symptoms and personal history. With tuned hyperparameters, the Random Forest model outperforms the others with an accuracy of 94%. Multiple performance metrics such as R2 score, mean squared error, and ROC AUC were analysed for each model, confirming the robustness of these techniques in clinical applications. Both approaches offer valuable diagnostic tools, with CNN excelling in radiological data analysis and traditional machine learning being more suitable for symptom-based predictions.

Keywords: Deep learning, Grad-CAM, KNN, Logistic Regression, Machine learning, Random Forest, X-ray images

1. INTRODUCTION :

The *COVID-19 pandemic* has created a *global health crisis*, necessitating efficient and accurate diagnostic methods. *Reverse Transcription Polymerase Chain Reaction (RT-PCR)* is the *gold standard* for COVID-19 detection, but it has limitations, including *high processing time, false-negative rates, and dependence on specialized kits* [1]. These challenges underscore the need for *complementary approaches* that are *rapid, reliable, and accessible* [2]. *Medical imaging*, particularly *chest X-rays (CXR)*, has emerged as a viable diagnostic tool due to its *availability and ability to detect COVID-19-induced lung abnormalities* [3]. *Deep learning techniques*, especially *Convolutional Neural Networks (CNNs)*, have demonstrated *high efficacy in analyzing radiological data*, offering an *automated and accurate diagnostic alternative* [4]. Additionally, *symptom-based machine learning models* can provide a *non-imaging approach*, particularly beneficial in *resource-limited settings* [5].

This study presents a dual-method framework for COVID-19 detection, combining deep learning and machine learning approaches. The first method employs a deep learning-based Convolutional Neural Network (CNN) model for chest X-ray (CXR) analysis, which achieves a 97% accuracy rate in distinguishing COVID-19 cases [1]. Techniques such as Grad-CAM and Canny edge detection are utilized to enhance interpretability, making CNN predictions more explainable for medical professionals [2]. The second approach involves a traditional machine learning pipeline, where Logistic Regression, K-Nearest Neighbors (KNN), and Random Forest algorithms predict COVID-19 likelihood based on patient symptoms and medical history [3]. Among these, the Random Forest model outperforms others, achieving 94% accuracy after hyperparameter tuning [4].

Performance evaluation metrics, including ROC-AUC, R^2 score, and mean squared error (MSE), validate the robustness of both methods [5]. The integration of deep learning for radiological analysis and machine learning for symptom-based predictions provides a comprehensive, AI-driven diagnostic tool that improves COVID-19 detection accuracy and supports clinical decision-making [6]. To ensure the reliability of AI models in healthcare, explainability methods like Grad-CAM provide insights into CNN decision-making, increasing trust and usability in clinical settings [7]. This study contributes to advancing COVID-19 detection by offering healthcare professionals an efficient, interpretable, and accurate AI-driven decision-support system [8]. Future research may explore the integration of multimodal data, such as CT scans and electronic health records, to further enhance diagnostic precision and predictive power [9].

2. LITERATURE SURVEY :

2.1. AI-Based Malware Defense for Android

The rise in cyber threats targeting Android devices has led to an increased focus on AI-driven cybersecurity solutions. Traditional signature-based malware detection methods have proven inadequate against rapidly evolving threats, necessitating the adoption **of** *machine learning (ML) and deep learning (DL)* techniques to enhance security. Recent studies emphasize the *efficacy of AI-powered malware detection models* in identifying malicious software through behaviour analysis, anomaly detection, and adversarial training methods [1].

2.2. Machine Learning Approaches in Malware Detection

Several ML algorithms have been employed in malware detection, including Support Vector Machines (SVM), Random Forest, and Logistic Regression, which analyse extracted features such as API call sequences and permissions used by applications. These models utilize static and dynamic analysis methods to classify malware efficiently [2]. However, traditional ML models often require extensive feature engineering and lack adaptability when faced with zero-day attacks.

2.3. Deep Learning Techniques in Android Security

Deep learning architectures, particularly Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks, have demonstrated remarkable effectiveness in malware detection. CNNs have been used to analyse the behaviour of Android applications by converting permission sets and API calls into feature maps, while LSTMs are employed for sequential data analysis, allowing better detection of sophisticated malware with evasive behaviour [3].

2.4. *Role of Explainable AI (XAI) in Malware Detection*

A significant challenge in AI-driven cybersecurity is the black-box nature of deep learning models. Explainable AI (XAI) techniques such as Gradient-weighted Class Activation Mapping (Grad-CAM) and SHAP (Shapley Additive explanations) are being integrated into malware detection frameworks to enhance interpretability. These methods help in understanding which features contribute to classification decisions, thereby increasing transparency and trust in AI-powered malware defines systems [4].

2.5 *Challenges and Future Directions*

Despite advancements, AI-driven malware detection faces several challenges, including adversarial attacks, high computational costs, and privacy concerns associated with data collection. Future research should focus on federated learning approaches to enhance privacy and real-time malware detection techniques using on-device AI models that reduce reliance on cloud-based processing [5]. Additionally, hybrid models combining symbolic AI and deep learning hold promise for more robust malware defines solutions.

3. PROPOSED SYSTEM :

The proposed system integrates deep learning and machine learning techniques to enhance the accuracy and efficiency of COVID-19 detection. This dual approach leverages medical imaging for radiological analysis and patient symptoms for clinical diagnosis, providing an interpretable and accessible diagnostic solution.

3.1. Deep Learning-Based COVID-19 Detection:

The first component of the proposed system utilizes a Convolutional Neural Network (CNN) to analyse chest X-ray (CXR) images for COVID-19 detection. The model is trained to extract key features from radiological data, differentiating COVID-19 cases from normal and pneumonia-affected lungs. To enhance interpretability, Gradient-weighted Class Activation Mapping (Grad-CAM) is employed to visualize important regions in the X-ray images, assisting healthcare professionals in validating the model's predictions. Additionally, edge detection techniques, such as the Canny edge detector, further highlight infection patterns, improving the diagnostic process.

3.2. Machine Learning-Based Symptom Classification

Alongside radiological analysis, the system incorporates a machine learning pipeline that predicts COVID-19 infection based on patient symptoms and medical history. This approach involves training various machine learning models, including Logistic Regression, K-Nearest Neighbours (KNN), and Random Forest, on structured clinical data. By analysing symptoms such as fever, cough, and anosmia, the system provides a non-imaging diagnostic alternative, particularly useful in resource-limited settings. The most effective model is selected based on performance evaluation metrics to ensure high classification accuracy and reliability.

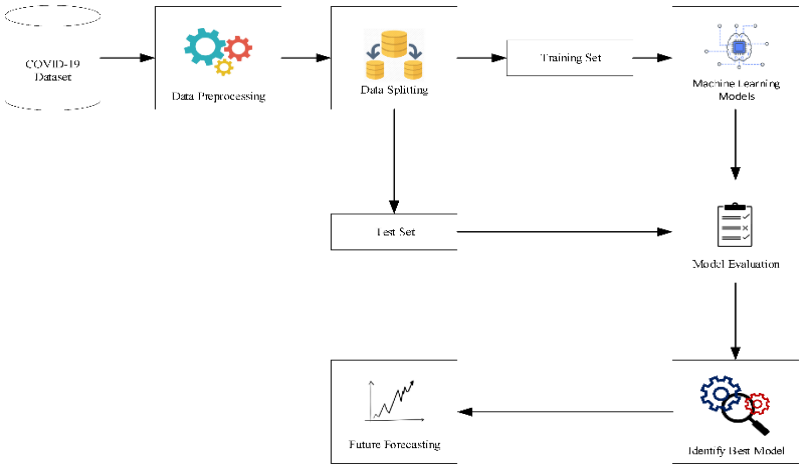


Figure.1. KNN Architecture

The image illustrates a machine learning-based COVID-19 data analysis and forecasting pipeline. The process begins with a COVID-19 dataset, which contains essential information such as case counts, recovery rates, and mortality rates. The data undergoes preprocessing, where missing values are handled, features are engineered, and data normalization or standardization is performed to ensure quality input for machine learning models. After preprocessing, the dataset is split into training and test sets, where the training set is used to build models, and the test set is used to evaluate their performance. Various machine learning models are trained on the processed data to identify patterns and make predictions. The models are then subjected to evaluation, where their accuracy, precision, recall, and other performance metrics are analyzed. The best-performing model is then identified based on evaluation results. Finally, the selected model is used for future forecasting, enabling predictions about potential trends in COVID-19 cases, helping in decision-making and policy planning. This systematic approach ensures effective analysis and prediction of COVID-19 trends using machine learning.

3.3. Basic Reproduction Number (R_0) in COVID-19 Spread

The basic reproduction number R_0 estimates how many secondary infections are caused by a single infected individual in a susceptible population.

$$R_0 = \beta / \gamma \quad (1)$$

Where:

- β = Transmission Rate
- γ = Recovery rate

If $R_0 > 1$, the infection spreads; if $R_0 < 1$, the infection declines.

3.3.1. SIR Model for Disease Spread

The Susceptible-Infected-Recovered (SIR) model is a mathematical model for infectious diseases:

$$\frac{dS}{dt} = -\beta SI$$
$$\frac{dI}{dt} = \beta SI - \gamma I$$
$$\frac{dR}{dt} = \gamma I \quad (2)$$

Where:

- S = Susceptible individuals
- I = Infected individuals
- R = recovered individuals
- β = Infection rate
- γ = recovery rate

3.3.2. Evaluation Metrics for Model Performance:

- **Accuracy**:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (3)$$

Where:

- TP = True Positives
- TN = True Negative
- FP = False Positives
- FN = False Negative

3.4. Architecture:

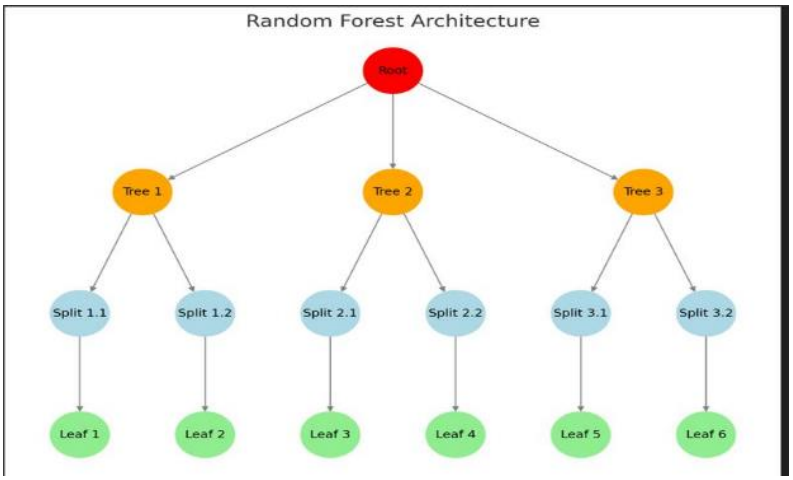


Fig 3.2: Random Forest Architecture

The Random Forest architecture illustrated in the image represents how the algorithm functions as an ensemble learning method. At the top, the root node (red) signifies the input data, which is subsequently distributed among multiple independent decision trees (orange). Each tree in the forest is trained on a different subset of the dataset, incorporating randomness to ensure diversity in predictions. As the data progresses through each tree, it encounters split nodes (light blue), where specific feature-based conditions divide the data into branches. This process continues until the data reaches the leaf nodes (light green), which represent the final classification or regression outcomes. Random Forest aggregates the individual predictions from all trees to arrive at a final decision, using majority voting for classification tasks and averaging for regression tasks. By combining multiple decision trees, the algorithm enhances prediction accuracy and reduces the risk of overfitting, making it more robust compared to a single decision tree. This technique is widely applied in various fields, including medical diagnostics, fraud detection, and recommendation systems, due to its ability to handle high-dimensional data while maintaining strong generalization performance. The structured yet flexible nature of Random Forest makes it a powerful tool in modern machine learning applications.

4. RESULT AND DISCUSSION:

Validation Set Classification Report:

	precision	recall	f1-score	support
0	0.94	0.90	0.92	973
1	0.91	0.95	0.93	972
accuracy			0.92	1945
macro avg	0.92	0.92	0.92	1945
weighted avg	0.92	0.92	0.92	1945

Fig 4.1: Accuracy

The Fig 4.1 displays a classification report for a validation dataset, presenting key performance metrics for a binary classification model. It includes precision, recall, and F1-score for both class labels (0 and 1), along with the number of samples in each class (support). The report shows high performance, with class 0 having a precision of 0.94, recall of 0.90, and F1-score of 0.92, while class 1 has a precision of 0.91, recall of 0.95, and F1-score of 0.93. The overall accuracy of the model is 92%, with both macro and weighted averages also at 0.92, indicating balanced performance across both classes.

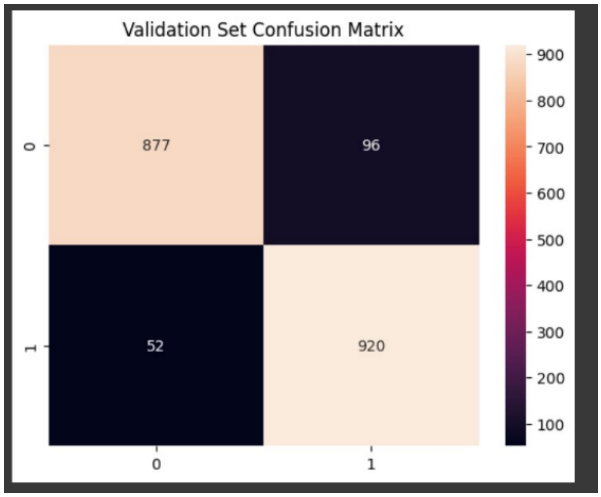


Fig 4.2: Confusion Matrix

The Fig 4.2 presents a confusion matrix, visualizing the performance of a classification model on a validation dataset. The matrix consists of four quadrants, where the top-left (877) represents true negatives (correctly classified as class 0), and the bottom-right (920) represents true positives (correctly classified as class 1). The top-right (96) indicates false positives, where class 0 was misclassified as class 1, while the bottom-left (52) represents false negatives, where class 1 was misclassified as class 0. The colour intensity in the heatmap reflects the frequency of classifications, with lighter shades representing higher values. This confusion matrix helps evaluate the model's accuracy and misclassification patterns, indicating that the model performs well with a relatively low number of misclassifications.

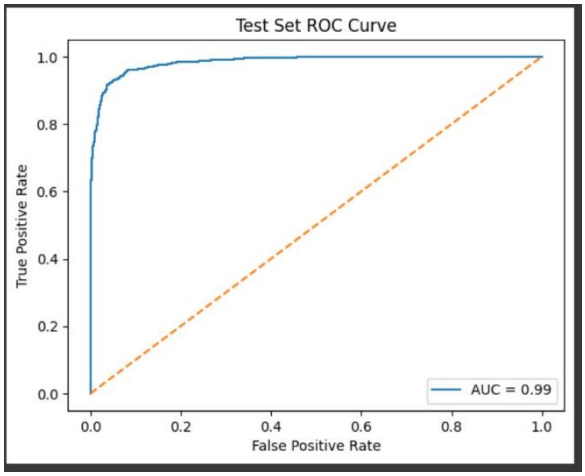
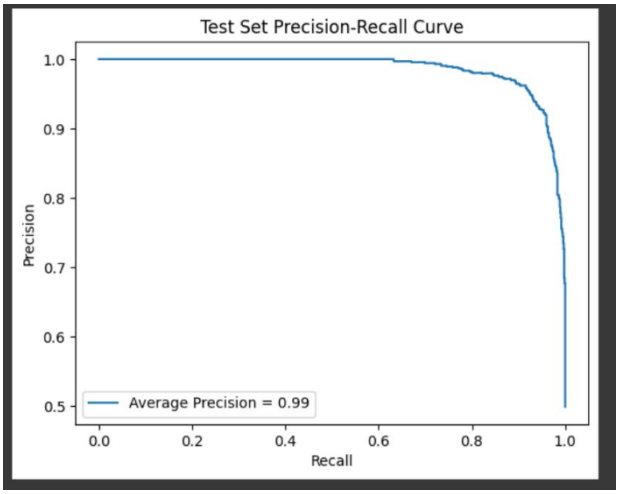


Fig 4.3: Test Set Roc Curve

The Fig 4.3 presents the Receiver Operating Characteristic (ROC) curve for a machine learning model evaluated on the test set. The ROC curve illustrates the trade-off between the true positive rate (sensitivity) and the false positive rate, helping to assess the model's ability to distinguish between classes. The blue curve represents the model's performance, while the orange dashed line represents a random classifier with no predictive power. The Area Under the Curve (AUC) is 0.99, indicating an excellent classification ability. A higher AUC value signifies that the model performs exceptionally well in distinguishing between positive and negative classes, making very few incorrect predictions.

Fig 4.4: Test Set Precision-Recall Curve



The Fig 4.4 displays the Precision-Recall (PR) curve for a machine learning model evaluated on the test set. The PR curve helps assess the model's ability to balance precision and recall, particularly in cases of imbalanced datasets. The x-axis represents recall (the proportion of actual positives correctly identified), while the y-axis represents precision (the proportion of predicted positives that are truly positive). The curve remains close to 1.0 for most of the recall range, demonstrating high model performance. The average precision score is 0.99, indicating a strong ability to correctly classify positive samples with minimal false positives.

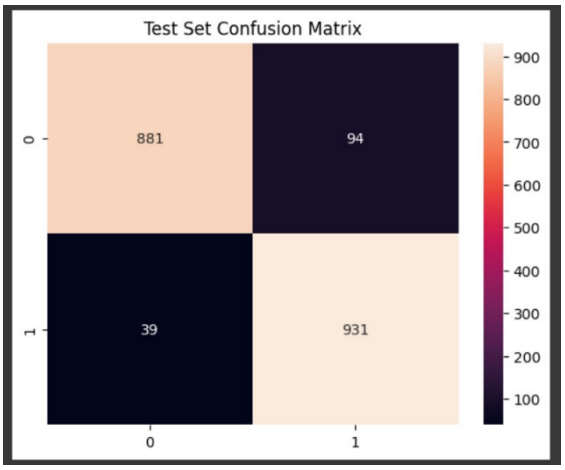


Fig 4.5: Test Set Confusion Matrix

The Fig 4.5 represents a confusion matrix for a machine learning model evaluated on the test set. A confusion matrix helps visualize the performance of a classification model by comparing actual vs. predicted labels. The matrix consists of four quadrants: True Negatives (881), False Positives (94), False Negatives (39), and True Positives (931). The True Negatives (881) indicate the correctly classified negative samples, while the True Positives (931) represent the correctly classified positive samples. The False Positives (94) and False Negatives (39) indicate misclassifications. The colour gradient highlights the frequency of values, with lighter shades representing higher values. This confusion matrix suggests that the model has strong predictive performance with relatively low misclassification rates.

5. CONCLUSION :

The proposed system integrates deep learning and machine learning techniques to enhance COVID-19 detection, addressing the limitations of conventional diagnostic methods. The deep learning-based approach leverages Convolutional Neural Networks (CNNs) to analyse chest X-ray images, achieving high accuracy in identifying COVID-19-induced abnormalities. Additionally, interpretability techniques such as Grad-CAM and edge detection enhance model transparency, aiding medical professionals in decision-making. The machine learning-based approach utilizes clinical symptoms and patient history for COVID-19 prediction, providing an accessible alternative for regions with limited radiological facilities. By combining these two methods, the system offers a robust, accurate, and interpretable diagnostic framework, contributing to the advancement of AI-driven healthcare solutions.

6. FUTURE SCOPE:

Although the proposed system demonstrates high performance, several areas remain for future improvement. The integration of multimodal data sources, such as CT scans, electronic health records, and blood test reports, could enhance diagnostic accuracy. Additionally, incorporating real-time data collection and continuous learning mechanisms would allow the model to adapt to emerging COVID-19 variants and other respiratory diseases. Further, explainable AI (XAI) techniques could be explored to improve model interpretability, ensuring greater trust among healthcare professionals. Finally, deploying the system as a web-based or mobile application would facilitate real-world adoption, providing rapid and accessible diagnostic support for hospitals and remote healthcare settings.

REFERENCES:

[1] S. Matheus, D. Swain, S. K. Satapathy, A. Rambha, B. Acharya, V. C. Georgianna's, and A. Kanavos, "COVID-19 Detection from Chest X-ray Images Based on Deep Learning Techniques," *Algorithms*, vol. 16, p. 494, 2023. DOI: 10.3390/a16100494

[2] A. Abbas, M. M. Abdelsamea, and M. M. Gaber, "Classification of COVID-19 in Chest X-ray Images using DeTraC Deep Convolutional Neural Network," *Applied Intelligence*, vol. 51, pp. 854–864, 2021. DOI: 10.1007/s10489-020-01829-7

[3] H. Alshazly, C. Linse, E. Barth, and T. Martinetz, "Explainable COVID-19 Detection Using Chest CT Scans and Deep Learning," *Sensors*, vol. 21, p. 455, 2021. DOI: 10.3390/s21020455

[4] A. Khan, A. Sohail, U. Zahoora, and A. S. Qureshi, "A Survey of the Recent Architectures of Deep Convolutional Neural Networks," *Artificial Intelligence Review*, vol. 53, pp. 5455–5516, 2020. DOI: 10.1007/s10462-020-09825-6

[5] I. D. Apostolopoulos and T. A. Mpesiana, "COVID-19: Automatic Detection from X-ray Images Utilizing Transfer Learning with Convolutional Neural Networks," *Physics in Engineering and Medicine*, vol. 43, pp. 635–640, 2020. DOI: 10.1007/s13246-020-00865-4