iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** VII **Month of publication:** July 2025

DOI: https://doi.org/10.22214/ijraset.2025.73133

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue VII July 2025- Available at www.ijraset.com

An Approach for Identifying COVID-19 from CT-SCANS using Machine Learning Algorithms

Joshna Gorantala, G. Narasimham
Information Technology, Jawaharlal Nehru Technological University Hyderabad, UCESTH

Abstract: ***This project introduces a machine learning approach for rapid COVID-19 detection using chest CT scans, addressing delays in RT-PCR testing. After enhancing and reducing the dimensionality of CT images, models like KNN, Decision Tree, SVM, and Random Forest were evaluated. The best—Random Forest, KNN, and SVM—were combined into a soft voting ensemble for improved accuracy. A Streamlit web app was developed to allow real-time image upload and instant COVID vs non-COVID predictions, providing a scalable diagnostic aid.***
Keywords: COVID-19, Machine Learning, Soft Voting Ensemble, Streamlit, Automated Diagnosis

I. INTRODUCTION

In December 2019, the outbreak of a novel coronavirus led to the global COVID-19 pandemic, significantly impacting public health systems worldwide. The rapid transmission of the virus made it essential to quickly identify infected individuals and isolate them to prevent further spread. The standard diagnostic method for COVID-19 has been the Reverse Transcription Polymerase Chain Reaction (RT-PCR) test. However, despite its widespread use, RT-PCR has limitations—its sensitivity is approximately 70%, and the test often involves lengthy processing times due to procedural complexities and environmental constraints.

As an alternative, medical imaging techniques such as chest X-rays (CXR) and Computed Tomography (CT) scans have proven useful in identifying COVID-19-related lung abnormalities. CT scans, in particular, offer higher sensitivity and can reveal infection patterns even in asymptomatic patients. However, analysing CT scans requires trained radiologists, which adds strain to already overwhelmed healthcare systems. To address this challenge, this project proposes an automated system that leverages machine learning algorithms to classify CT scan images as COVID or Non-COVID. The approach aims to support medical professionals by providing a fast, reliable, and scalable diagnostic tool that can assist in early detection and treatment planning.

A. Objective

The objective of this project is to automate the classification of chest CT scans into COVID and Non-COVID using machine learning. It enhances images, reduces feature dimensions with PCA, combines top models in a soft voting ensemble, and deploys the system through a Streamlit web app for real-time diagnosis.

II. LITERATURE SURVEY

Machine learning has shown significant promise in medical image analysis, especially for COVID-19 detection using chest X-rays and CT scans. Various studies have used traditional and deep learning methods to enhance diagnostic speed and accuracy. This chapter reviews existing literature, highlighting key methods and gaps that inspire the current work.

The paper by Lakshmi Sravani Videla, presents a CNN-based approach to detect COVID-19 from chest X-rays, beginning with exploratory image analysis using techniques like HSV conversion, Gaussian blur, erosion, dilation, and edge detection. It observes fewer edges in COVID-19 X-rays, which aids in classification. A CNN with four convolutional layers and three fully connected layers is proposed and trained on an augmented dataset, achieving 75% accuracy on a 20-image test set. The model's performance is compared with VGG16, supported by accuracy and loss curves. The study attributes lower accuracy to outlier images and recommends dataset refinement and expansion for better results. [1].

The paper by Safynaz Abdel-Fattah Sayed presents a machine learning framework for predicting COVID-19 severity and mortality risk using chest X-rays. It involves four stages: preprocessing, feature extraction (handcrafted and CheXNet-based), feature selection (PCA and RFE), and classification using models like KNN, SVM, Random Forest, and XGBoost. The combination of PCA and RFE improved handcrafted feature performance, with XGBoost and SVM achieving up to 97% accuracy and 100% ROC-AUC. CheXNet features with RFE further boosted performance to 99.6%, especially with Extra Tree and SVM. The study highlights the crucial role of feature selection in model effectiveness and its potential for clinical use [2].

The paper by Shanjiang Tang Yang Zhang proposes EDL-COVID, an ensemble deep learning model for detecting COVID-19 from chest X-ray images using a weighted average ensembling (WAE) technique. Built on the COVID-Net architecture, it employs snapshot ensemble learning to boost performance. Tested on the COVIDx dataset, EDL-COVID achieved 95% accuracy, 96% sensitivity, and 94.1% positive predictive value. The results show that the ensemble model outperforms individual deep learning models in COVID-19 detection [3].

The reseach by Amin Ul Haq describes a deep learning-based method for the precise and prompt identification of COVID-19 from chest X-ray images using a Convolutional Neural Network (CNN) model. The proposed CNN model demonstrates strong performance, achieving 96% accuracy, 97% precision, 96% recall, and a 97% F1-score in classifying COVID-19, normal, and pneumonia cases. Based on these results, the research recommends the proposed CNN model as an effective tool for COVID-19 detection. The paper also outlines plans for future work, which include conducting further experiments with additional deep learning models to enhance and validate the approach [4].

The study by Emrah Irmak introduces a custom deep CNN model for COVID-19 detection from chest X-rays, featuring 12 weighted layers including two convolutional layers and one fully connected layer. The model uses specific kernel configurations, takes 227×227×3 input images, and is trained with SGDM optimization. It achieves 99.20% accuracy and an AUC of 0.9998, outperforming several existing models. Evaluation via confusion matrix confirms its high classification performance [5].

III.EXISTING SYSTEM VS PROPOSED SYSTEM

A. Existing System

Recent studies on COVID-19 detection using chest imaging have employed both traditional machine learning and deep learning methods, with CNNs and ensemble models like EDL-COVID achieving high accuracy and AUC scores. While traditional approaches rely on structured pipelines and feature selection, deep models excel at learning directly from images. However, limitations such as small datasets, lack of standardization, poor generalizability, and limited clinical deployment remain. These gaps emphasize the need for more robust, interpretable, and clinically viable diagnostic systems.

B. Proposed System

To overcome limitations in existing COVID-19 diagnostic systems, this project proposes an automated ensemble-based approach using CT scan images for improved accuracy. The pipeline includes image enhancement, PCA-based dimensionality reduction, and training multiple models—Random Forest, KNN, and SVM—combined via a Soft Voting Ensemble. A Streamlit online application provides a useful and scalable diagnostic tool by allowing real-time CT image submission and rapid COVID/Non-COVID categorization.

IV.METHODOLOGY

A. System Architecture

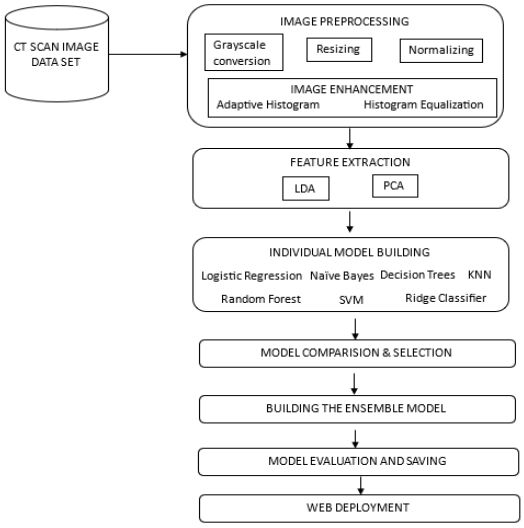


Fig. 1 System Architecture of the Approach for Identifying Covid-19 From Ct scans Using Machine Learning Algorithms System Architecture

The Fig. 1 illustrates the system architecture of the project. The system architecture for this project outlines a structured workflow that begins with a CT scan image dataset, followed by a series of preprocessing steps including grayscale conversion, resizing, and normalization to standardize input data. Feature extraction is then performed using dimensionality reduction techniques like PCA and LDA. Multiple machine learning algorithms—including Logistic Regression, Naïve Bayes, Decision Tree, KNN, Random Forest, SVM, and Ridge Classifier—are trained individually and evaluated. Based on performance comparison, the top three models are selected and combined into a soft voting ensemble. The final ensemble model is evaluated, saved, and deployed through a Streamlit web interface, enabling real-time COVID-19 classification from CT images.

B. Methodology

The system employs a structured, modular approach to classify CT scan images as COVID or Non-COVID. Each module in the pipeline is responsible for a specific task, contributing to a robust, flexible, and easy-to-deploy diagnostic model.

1) Image Preprocessing Module: Raw CT images are converted to grayscale, resized for uniformity, and normalized to stabilize model training. Contrast enhancement is applied using Histogram Equalization and Adaptive Histogram Equalization to highlight critical lung features.

2) Feature Extraction and Dimensionality Reduction Module: Preprocessed images are flattened into vectors and standardized. Principal Component Analysis (PCA) is then used to reduce dimensionality while retaining 90% of the variance, improving model efficiency. LDA was tested but PCA was chosen for final use.

3) Model Training and Evaluation Module: Several classifiers—including Logistic Regression, Naïve Bayes, Decision Tree, KNN, Random Forest, SVM, and Ridge Classifier—are trained individually. Their performance is evaluated using metrics like accuracy, precision, recall, and F1-score to identify top models.

4) Ensemble Construction Module: The best-performing models—Random Forest, KNN, and SVM—are combined using a soft voting strategy. This method averages prediction probabilities, improving classification accuracy and reducing model bias.

5) Model Evaluation and Saving Module: The ensemble model is validated on test data using a confusion matrix and performance metrics. Once validated, the model is saved using joblib or pickle for reuse without retraining during deployment.

6) Web Deployment Module: The saved model is deployed via a Streamlit web app. Users can upload a CT scan image, which is preprocessed and passed through the ensemble model to instantly predict COVID or Non-COVID status, enabling real-time, user-friendly diagnosis.

V. IMPLEMENTATION

The implementation phase involves converting the proposed methodology into a functional system that can classify CT scan images as COVID or Non-COVID. The pipeline integrates image preprocessing, feature transformation, model training, ensemble learning, and deployment into a real-time web-based interface. Python and its machine learning libraries such as Scikit-learn, NumPy, OpenCV, and Streamlit are used throughout the development process.

A. Machine Learning Algorithm Development

CT scan images are first enhanced using Adaptive Histogram Equalization and Histogram Equalization to improve contrast and highlight important features. The enhanced images are flattened and standardized, followed by dimensionality reduction using Principal Component Analysis (PCA) to retain 90% variance. Various classifiers including K-Nearest Neighbors (KNN), Random Forest, Support Vector Machine (SVM), Logistic Regression, and others are trained and evaluated using metrics like accuracy, precision, recall, and F1-score. The top three models—Random Forest, KNN, and SVM—are combined using a soft voting ensemble strategy to improve prediction robustness and accuracy. The final ensemble model is saved using joblib for deployment.

B. Soft Voting Ensemble Algorithm

Soft Voting is an ensemble learning technique used in classification tasks, where multiple base classifiers are combined to improve prediction performance. Unlike hard voting, where each classifier simply votes for a class label and the majority wins, soft voting takes a more nuanced approach by considering the predicted probabilities from each model. In soft voting, each classifier outputs a probability distribution across all possible classes. These probabilities are then averaged (optionally weighted) for each class. The class that has the highest average likelihood is chosen as the final outcome. This method allows the ensemble to factor in not just the predicted class but also the confidence of each individual model in its prediction.

C. Streamlit Web Application

To make the model accessible and user-friendly, a web application is developed using the Streamlit framework. The app allows users to upload a CT scan image, which is then processed using the same enhancement and transformation steps applied during training. The preprocessed image is passed through the saved ensemble model to generate a real-time prediction (COVID or Non-COVID). The result is displayed instantly, providing a practical diagnostic tool for clinicians, researchers, or general users.

VI. RESULTS

A. Results

The implementation of the proposed COVID-19 prediction system is evaluated through a user-friendly web application developed using Streamlit. The interface allows users to upload CT scan images and instantly receive predictions.

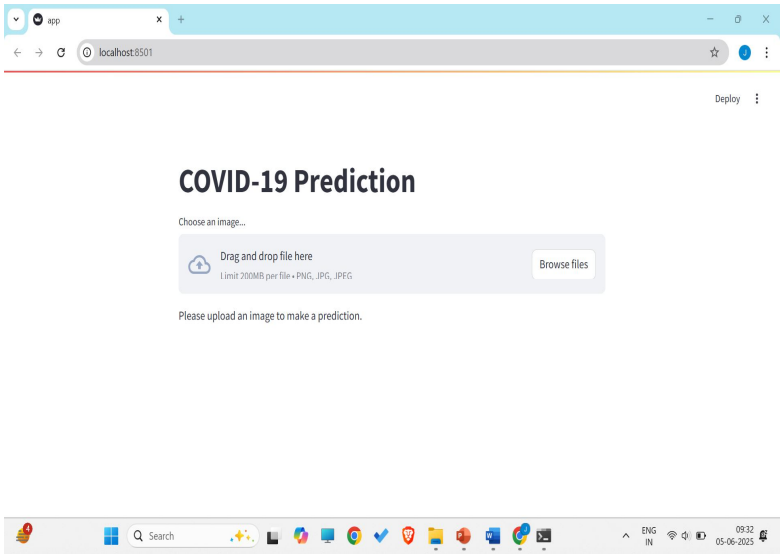


Fig. 2 Main Screen of Covid-19 Prediction System

Fig. 2 shows the main screen of the web application, where users can upload CT images via the "Browse files" button.

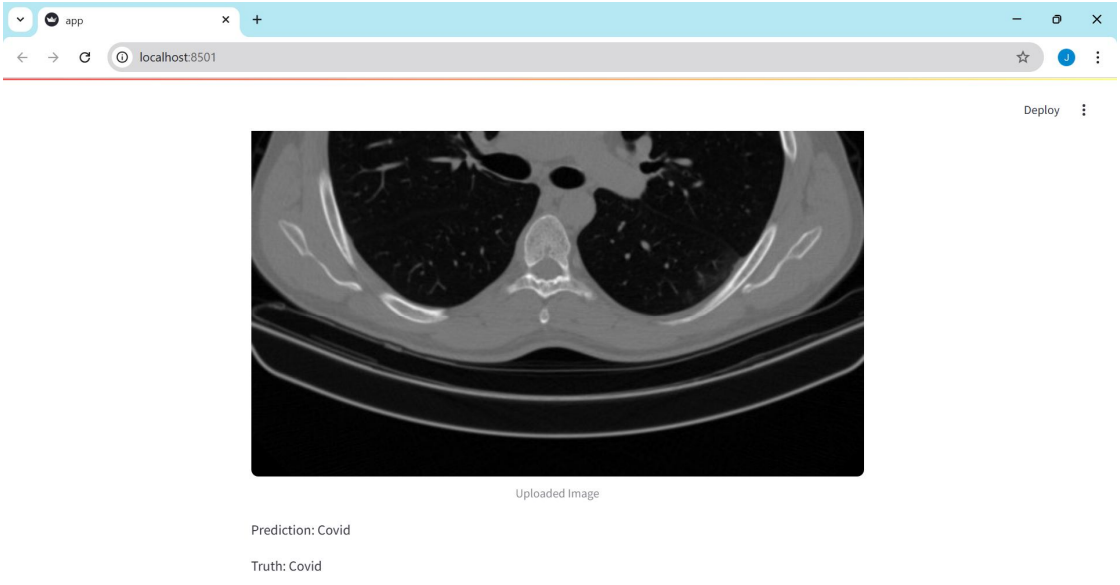


Fig. 3 Covid image being predicted correctly as Covid

Fig. 3 illustrates a successful prediction where a COVID-positive CT image was correctly classified as COVID by the system.

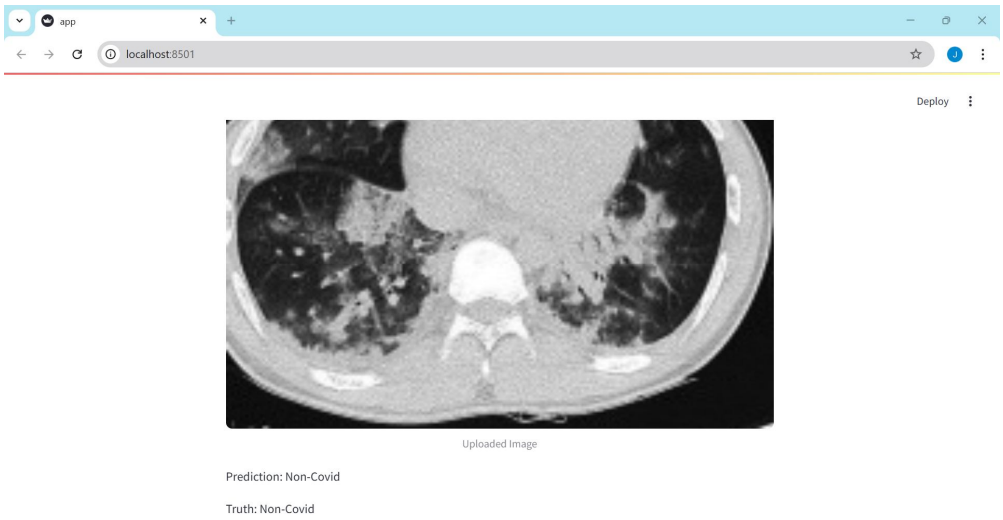


Fig. 4 non-Covid image being predicted as non – covid correctly

Fig. 4 demonstrates that a Non-COVID image was accurately classified by the system as Non-COVID.

B. Analysis

The results of this project are analyzed based on the performance of various machine learning models applied to CT scan images after preprocessing and dimensionality reduction. The evaluation focused on key classification metrics including accuracy, precision, recall, F1-score, and confusion matrix values for both individual models and the final ensemble classifier.

Initially, individual models such as Logistic Regression, Naive Bayes, K-Nearest Neighbors (KNN), Decision Tree, Random Forest, Support Vector Machine (SVM), and Ridge Classifier were trained on features extracted from preprocessed CT images. The belowe are the results observed on individual models.

S.No	ALGORITHM	ATTRIBUTE SELECTOR	CONFUSION MATRIX	ACCURACY	PRECISSION	RECALL	F1 SCORE
1	LOGISTIC REGRESSION	LDA	[1301, 292], [262, 1334]	0.82627783	0.826390299	0.82627783	0.826260917
		PCA	[1201, 392], [463, 1133]	0.731890875	0.732362784	0.731890875	0.731763534
2	NAÏVE BAYES	LDA	[1324, 269], [295, 1301]	0.823142051	0.823230704	0.823142051	0.823131651
		PCA	[1300, 293], [945, 651]	0.61179053	0.634393617	0.61179053	0.594939761
3	KNN	LDA	[1363, 230], [375, 1221]	0.810285356	0.812888621	0.810285356	0.809900489
		PCA	[1591, 2], [5, 1591]	0.997804955	0.997806721	0.997804955	0.997804955
4	DECISION TREE	LDA	[1229, 364], [377, 1219]	0.767638758	0.767658332	0.767638758	0.767635788
		PCA	[1505, 88], [88, 1508]	0.944810285	0.944810285	0.944810285	0.944810285
5	RANDOM FOREST	LDA	[1229, 364], [377, 1219]	0.767638758	0.767658332	0.767638758	0.767635788
		PCA	[1586, 7], [5, 1591]	0.996237065	0.996237841	0.996237065	0.996237061
6	SVM	LDA	[1302, 291], [266, 1330]	0.825337096	0.825414518	0.825337096	0.825325073
		PCA	[1554, 39], [43, 1553]	0.97428661	0.974289656	0.97428661	0.9742866
7	RIDGE CLASSIFIER	LDA	[1324, 269], [296, 1300]	0.822828473	0.822923885	0.822828473	0.822817183
		PCA	[1203, 390], [492, 1104]	0.723424271	0.724358913	0.723424271	0.723149381

Fig. 5 Results of individual models on Adaptive Histogram data

S.No	ALGORITHM	ATTRIBUTE SELECTOR	CONFUSION MATRIX	ACCURACY	PRECISSION	RECALL	F1 SCORE
1	LOGISTIC REGRESSION	LDA	[1313, 280], [272, 1324]	0.826904986	0.826912398	0.826904986	0.826903488
		PCA	[1193, 400], [477, 1119]	0.724992161	0.725530547	0.724992161	0.724837993
2	NAÏVE BAYES	LDA	[1338, 255], [305, 1291]	0.824396362	0.824720825	0.824396362	0.824355776
		PCA	[1312, 281], [958, 638]	0.611476952	0.6361579	0.611476952	0.593229482
3	KNN	LDA	[1337, 256], [376, 1220]	0.801818752	0.803551818	0.801818752	0.80154478
		PCA	[1593, 0], [5, 1591]	0.99843211	0.998437016	0.99843211	0.998432109
4	DECISION TREE	LDA	[1251, 342], [366, 1230]	0.77798683	0.778052966	0.77798683	0.777975827
		PCA	[1482, 111], [104, 1492]	0.932580746	0.932588805	0.932580746	0.932580282
5	RANDOM FOREST	LDA	[1251, 342], [366, 1230]	0.77798683	0.778052966	0.77798683	0.777975827
		PCA	[1593, 0], [1, 1595]	0.999686422	0.999686619	0.999686422	0.999686422
6	SVM	LDA	[1302, 291], [263, 1333]	0.82627783	0.826375607	0.82627783	0.826263001
		PCA	[1576, 17], [35, 1561]	0.983693948	0.983755768	0.983693948	0.983693515
7	RIDGE CLASSIFIER	LDA	[1338, 255], [305, 1291]	0.824396362	0.824720825	0.824396362	0.824355776
		PCA	[1217, 376], [512, 1084]	0.721542803	0.723188565	0.721542803	0.721046673

Fig 6 Results of individual models on Histogram Equalized data

Key analysis observed are:

- Histogram Equalization (HE) consistently outperforms Adaptive Histogram Equalization (CLAHE) in terms of accuracy, precision, and F1-score across most models.
- K-Nearest Neighbors (KNN) with PCA-transformed features emerges as the top-performing model, showing high classification performance on both HE and CLAHE images.
- Support Vector Machine (SVM) and Random Forest models significantly improve when combined with HE + PCA, making them strong candidates for inclusion in the ensemble model.
- Naïve Bayes performs poorly across all configurations and is not recommended for ensemble use unless interpretability or fast computation is required.
- Principal Component Analysis (PCA) is clearly more effective than LDA for this dataset, offering better accuracy and model consistency.

An ensemble model was constructed using the findings of the various models, and the ensemble model's results are shown below.

- Accuracy: 99.84%
- Confusion Matrix:

	Predicted covid	Predicted non -covid
Actual covid	1566	3
Actual non – covid	2	1618

Out of 3189 test samples, only 5 misclassifications occurred, resulting in extremely low false positive and false negative rates. This is critical in medical applications where false negatives—failing to detect actual COVID cases—can have serious consequences. Overall, the model exhibits excellent sensitivity (recall) and specificity, confirming its reliability and effectiveness in real-time COVID-19 diagnosis using chest CT scans.

VII. CONCLUSION

This project presents an effective machine learning-based system for detecting COVID-19 from chest CT scan images, achieving a high accuracy of 99.84%. It combines image preprocessing, histogram equalization, PCA-based dimensionality reduction, and multiple classifiers—KNN, SVM, and Random Forest—into a soft voting ensemble for robust performance. The system is deployed through a Streamlit web interface, enabling real-time image classification. Evaluation results, including a near-perfect confusion matrix, demonstrate the model's reliability. This approach offers a fast, automated alternative to RT-PCR testing and holds promise for future integration with larger datasets and advanced deep learning techniques.

VIII. FUTURE SCOPE

Future improvements to this system include integrating deep learning models like CNNs or ResNet for enhanced accuracy, expanding the dataset for better generalization, and extending the model to detect multiple lung diseases. Cloud or mobile deployment can increase accessibility, while a radiologist feedback loop can enable continuous learning. These enhancements will make the system more scalable, accurate, and clinically useful.

IX. ACKNOWLEDGMENT

I would like to express my sincere gratitude to Sri G. Narasimham sir for his constant guidance, encouragement, and invaluable support throughout the course of this project. I also extend my heartfelt thanks to the authors of the research papers referenced in this study, whose work provided a strong foundation and direction for this research.

REFERENCES

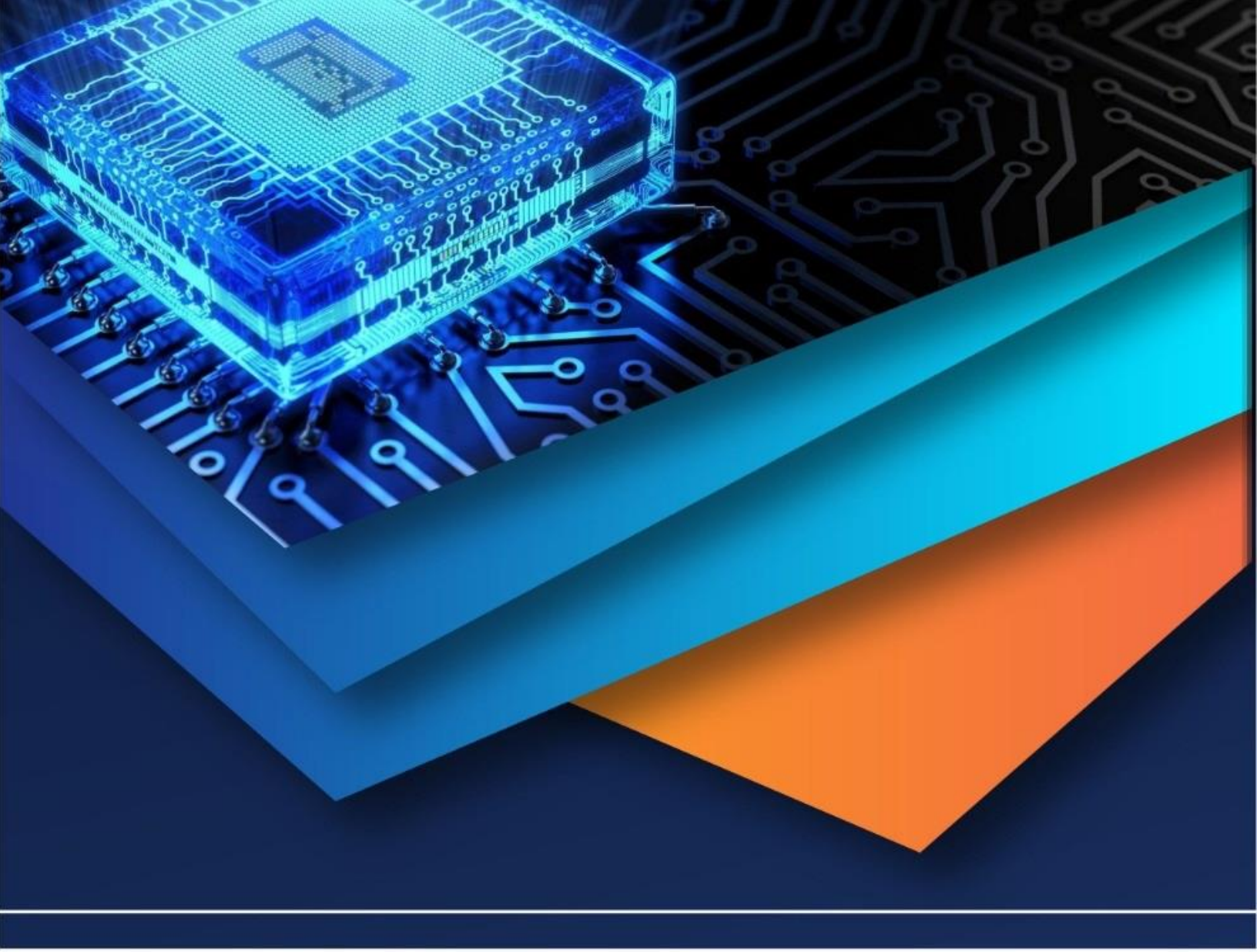
[1] Videla, Lakshmi Sarvani, et al. "Convolution Neural Networks based COVID-19 Detection using X-ray Images of Human Chest." *2022 8th International Conference on Smart Structures and Systems (ICSSS)*. IEEE, 2022.

[2] Sayed, Safynaz Abdel-Fattah, Abeer Mohamed Elkorany, and Sabah Sayed Mohammad. "Applying different machine learning techniques for prediction of COVID-19 severity." *Ieee Access* 9 (2021): 135697-135707.

[3] Tang, Shanjiang, et al. "EDL-COVID: Ensemble deep learning for COVID-19 case detection from chest X-ray images." *IEEE Transactions on Industrial Informatics* 17.9 (2021): 6539-6549.

[4] Haq, Amin Ul, et al. "Deep Learning Approach for COVID-19 Identification." *2021 18th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)*. IEEE, 2021.

[5] Irmak, Emrah. "A novel deep convolutional neural network model for COVID-19 disease detection." *2020 Medical Technologies Congress (TIPTEKNO)*. IEEE, 2020.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY