An Effective Learning Ensemble based Prediction Model for Detecting COVID-19 from Human Chest X-ray Images

S.Deepankumar[1], R.MaruthaVeni[2]

deepankumarmca3@gmail.com[1], dhanuvene1@gmail.com[2]

Research Scholar, Department of Computer Science, Dr. SNS RajaLakshmi College of Arts and science, Coimbatore[1], Assistant Professor, Department of Computer Science, Dr. SNS RajaLakshmi College of Arts and Science, Coimbatore[2]

Abstract. The emergence of the Coronavirus Disease-19 (COVID-19) has presented a formidable global challenge. Manual diagnosis has been hindered by limited access to radiologists and experts, as well as their varying proficiency in interpreting the intricate features within chest X-ray (CXR) images associated with the disease. In response to this, this study introduces an innovative automated screening ensemble model designed to differentiate between normal patients, those infected with COVID-19, and individuals with suspected COVID-19, all through the utilization of radiomic texture descriptors extracted from CXR images. The ensemble based prediciton model leverages a selection based classifier that combines and ensembles the strength of two well-established supervised classification algorithms. Remarkably, this ensemble model achieves a substantial enhancement in prediction accuracy, boasting an impressive 94% accuracy rate. Furthermore, it excels in the precise identification of COVID-19 cases, achieving an accuracy and precision of 93%. This advancement in automated screening not only streamlines the diagnostic process but also provides a reliable tool in the ongoing battle against the global COVID-19 pandemic.

Keywords: Ensemble Machine Learning, Novel Coronavirus Disease 2019 (Covid – 19), Intelligent Prediction, Image andComputer Vision.

1 Introduction

The emergence of Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) and the subsequent disease it causes, known as COVID-19, represents a profound global challenge.[1] This contagious disease was initially identified in Wuhan City, Hubei Province, China, within the seafood wholesale market, in December 2019. Rapidly, it spread across international borders, prompting the World Health Organization (WHO) to declare it a pandemic on March

IACIDS 2023, November 23-25, Lavasa, India

DOI 10.4108/eai.23-11-2023.2343172

11, 2020. By September 20, 2020, COVID-19 had already afflicted millions, with approximately 30,675,675 confirmed cases and nearly 954,417 fatalities reported[2].

To curb the dissemination of this deadly contagion, nations have instituted a range of measures.These include advocating for social distancing, emphasizing the importance of hygiene practices, enhancing screening procedures through versatile testing, and launching extensive vaccination campaigns. The “Reverse Transcriptase-Polymerase Chain Reaction“ (RT-PCR) [3] method has been the primary diagnostic tool, but it possesses limitations, notably delayed diagnosis of suspected cases and the issue of false-negative results, hindering efforts to control and prevent the disease.

To overcome these testing challenges, researchers have been actively working on the development of a more efficient and rapid diagnostic approach for COVID-19. Notably, recommendations from “Wuhan University‘s Zhongnan Hospital“ and the “WHO“ have emphasized the importance of chest imaging in addition to clinical symptoms as a diagnostic tool. Rubin and colleagues have provided valuable guidelines for utilizing radiography and CT scan in disease prediction and assessment. While CT scans offer high sensitivity, they are associated with drawbacks, including high costs and radiation exposure, particularly concerning vulnerable populations. However, diagnosis based on chest X-ray has emerged as the preferred method for COVID-19 detection and management. Studies conducted by Ng and others have demonstrated that pulmonary infections related to COVID-19 can be accurately identified through chest X-ray images.

Furthermore, healthcare professionals are recognizing the importance of chest X-rays, in conjunction with Artificial Intelligence (AI) systems, for detecting COVID-19 symptoms, particularly opaque lung patterns [4]. Additionally, the adoption of ensembling approaches that combine various classification and learning algorithms into a single model has shown promise in achieving superior predictive performance. Such ensemble techniques not only reduce error rates but also offer speed and efficiency advantages compared to individual models. In conclusion, the battle against COVID-19 has seen advancements in diagnostics and the utilization of chest imaging, AI, and ensemble machine learning methods to improve disease detection and management while addressing the limitations of existing testing techniques.

The remaining part of the paper is organized as follows,section1 provides Introduction and background of corona disease and diagnostic challenges.Section 2 briefly discusses the background of the problem and literature survey. Detailed methodology, including data collection, preprocessing, and ensemble learning are discussed in section 3.Section3 discusses Experimental results in detail and section 5 concludes with future research directions.

2 Background Study

Amid the pressing need for swift COVID-19 identification, CNN-based AI systems have gained immense popularity [4]. Their application is primarily driven by the potential to expedite the analysis of medical images, particularly chest X-rays, a powerful tool for COVID-19 detection due to shared symptoms with pneumonia. Several studies have explored the use of CNNs and diverse feature extraction techniques to detect COVID-19, playing a crucial role in easing the burden on overwhelmed healthcare systems. Notable surveys have delved into CNN technology,

evaluating its application for the detection of COVID-19 and automated lung segmentation, often focusing on Computed Tomography (CT) and X-ray imaging. Researchers have also modified and tested pre-trained CNNs like AlexNet, Inception, and ResNet, achieving high accuracy rates, some nearing 98% [5].

Furthermore, novel CNN frameworks like COVID-Net and advanced CNN architectures, including MobileNetV2 [6], VGG19 [7], and Xception, trained through transfer learning, have shown promise in radiological feature extraction for COVID-19 detection. Ensembles of heterogeneous or homogeneous models have been proposed, with many studies highlighting the potential of ensemble methods in reducing prediction errors. Although many methods rely on single-model predictions, the ensemble approach has emerged as a versatile and effective means to enhance accuracy. The study reveals the extensive research and development in the domain of COVID-19 detection through CNN-based AI systems and underscores the potential of ensemble models to improve diagnostic outcomes.

3 Proposed Research Methodology

The methodology comprises four levels: data collection, pre-processing, feature selection, and an ensemble machine learning model. Data was obtained from two databases, images pre-processed, features selected, and two classifiers (KNN and RF) fine-tuned. These models' predictions were combined through soft voting, ensuring robust COVID-19 detection.

3.1 Methodological Approach for COVID-19 Detection Model

The methodological approach employed in this research encompasses several key levels, each contributing to thecorona detection based on the the development of an effective ensemble model based classification.

The research initiated with data collection from two prominent public databases: the corona diseased patients Chest X-Ray Dataset Initiative and the corona Radiography Database. These comprehensive datasets encompassed a diverse range of chest X-ray images, including normal, corona positive, and viral pneumonia cases. The COVID-19 Radiography Database featured 1341 normal X-ray images, 1143 abnormal positive images, and 1345 viral pneumonia images. Additionally, the corona Chest X-ray Dataset Initiative contributed an extra 48 corona abnormal positive images.

Following data acquisition, a crucial data pre-processing phase was executed to prepare the chest X-ray images for classification. This entailed converting the images into arrays and numerical arrays, followed by normalization to ensure feature values fell within the 0 to 1 range.

Subsequently, the pivotal feature selection process aimed to enhance the machine learning model's performance by reducing data redundancy and eliminating noise. This research employed new methods for extracting features from X-ray images which is „vector-based and color histogram features“ and for feature reduction.. From an initial 90,000 features, 65,536 were meticulously chosen.

The heart of the proposed methodology lies in the ensemble machine learning paradigm, which amalgamates diverse classification algorithms into a single, robust model. The two most accurate classifiers, the K-Nearest Neighbor (KNN) improved with initial mean evaluation as

medoids algorithm [9] and the Random Forest (RF) [9] algorithm, underwent parameter tuning for optimization. RF parameters, 'n tree' – number of trees and 'mtry,' - (number offeatures ineach split were fine-tuned, and the KNN algorithm's 'k' parameter was meticulously determined through a bootstrap procedure.This methodological framework offers a comprehensive and refined approach to effectively prepare, select features, and employ ensemble learning for COVID-19 detection in chest X-ray images.

The key achievement was combining these two algorithms' predictions through soft voting, where probabilities under the curve plots were calculated to ensure accuracy.

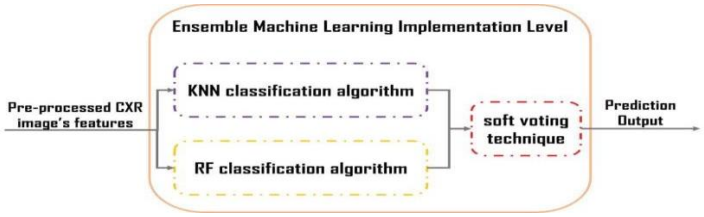


Fig. 1. Overall Architecture of the Ensemble Model

The architecture of the proposed ensemble machine learning model, as shown in Figure 1, summarizes the comprehensive methodological approach. The final model underwent empirical evaluation using state-of-the-art methods to assess its effectiveness in COVID-19 detection. Every step of implementation, from data collection to ensemble machine learning, was meticulously executed to build a robust model. This methodological approach forms the foundation of the research, enabling the development of an ensemble machine learning model designed for accurate COVID-19 detection through the analysis of chest X-ray images.

4 Experimental Results and Discussions

In the evaluation of the model, essential metrics are utilized to assess performance. Metrics comprise Precision, Recall, Accuracy, F1 Measure, ROC Curve, and AUC value. Accuracy, which measures overall correctness, is computed as (True Positives + True Negatives) divided by the total predictions. Precision assesses positive prediction accuracy, calculated as True Positives divided by (True Positives + False Positives). Recall quantifies the model's ability to identify relevant instances, computed as True Positives divided by (True Positives + False Negatives). The F1 Score balances Precision and Recall with F1 = (2 * Precision * Recall) divided by (Precision + Recall). These metrics are vital for evaluating the model by comparing predictions to actual outcomes. The Confusion Matrix, AUC Score, and ROC Curve provide additional insights into predictive power and the ability to handle various scenarios.

Table1. Accuracy obtained for the various classifiers

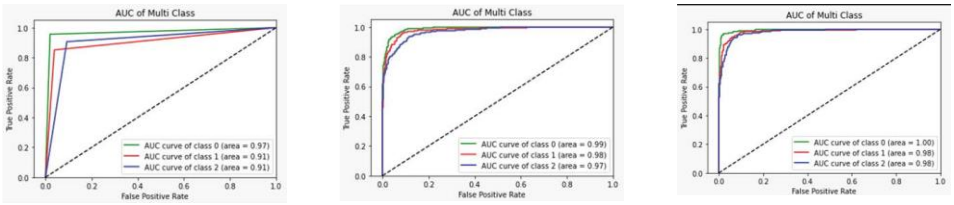
Classification	**Prediction Accuracy**
KNNMethod	0.91
RF Method	0.91
Proposed EL Method	0.95

Table 1 displays the accuracy ratings of several machine learning techniques: K-Nearest Neighbors (KNN), Random Forest, and an Ensemble Learning approach. Accuracy, a pivotal performance metric in machine learning, gauges the model's prediction accuracy. Specifically, the KNN Algorithm attained a 90% accuracy rate, signifying its correctness in approximately 90% of predictions. Similarly, the Random Forest Algorithm achieved a 90% accuracy rate, mirroring its precision in around 90% of cases. Notably, the Ensemble Learning Algorithm excelled with a 94% accuracy rate, indicating the effectiveness of combining classification algorithms, possibly including KNN and Random Forest, to enhance prediction accuracy. This ensemble approach proves particularly adept at prediction tasks. A higher accuracy score generally implies a more reliable and effective model, suggesting that the Ensemble Learning Algorithm excels in making correct predictions, while KNN and Random Forest perform well but with slightly lower accuracy rates. However, the choice of the best algorithm also depends on other factors such as the specific problem, computational resources, and the trade-off between precision and recall.Taking into account individual class predictions, the performance of all three algorithms under consideration is improved. Table 2 displays the Precision, Recall, and F1-score metrics for the proposed model and standard KNN, standard RF medthods.

Table2. Class-wiseclassificationresultsofProposed Ensemblemodel

	Class	**Precision**	**Recall**	**F1-score**
KNN	0	0.95	0.96	**0.95**
	1	0.92	0.85	**0.88**
	2	0.85	0.91	**0.88**
RF	0	0.93	0.92	**0.92**
	1	0.91	0.90	**0.90**
	2	0.86	0.88	**0.87**
Ensemble	0	0.96	0.96	**0.96**
	1	0.94	0.90	**0.92**
	2	0.89	0.93	**0.91**

When considering Precision(%), Recall(%), and F1-score and Accuracy (%), it can be challenging to gauge the model's performance fully. The AUC scores achieved were 1.00 for abnormal, 0.98 for normal cases, and 0.98 for viral pneumonia in the ROC curves. Significantly, the AUC for abnormal outperforms the other classes.



(a) (b) (c)

Fig. 2. Comparison of the ROC curve

In Figure 2, the ROC curve for all the models under evaluation are presented. It's evident that out of the 340 test images, 14 images were misclassified. When it comes to classifying COVID-19 images, both the KNN and ensemble models demonstrate similar performance. However, in the case of other classes, the performance of KNN and RF declines, while they exhibit a slight improvement in the ensemble model. This behavior can be attributed to the unique nature of corona infection, which shares some common features with other types of viral pneumonia, potentially causing confusion among the models.

The results indicate that K-Nearest Neighbors (KNN) models exhibit commendable recall when categorizing viral pneumonia and COVID-19 cases. Conversely, they exhibit diminishing performance in predicting normal cases. On the other hand, Random Forest (RF) shows proficiency in identifying normal chest X-ray images but experiences a decrease in performance when detecting viral pneumonia compared to other models. These findings confirm the exceptional efficacy of the introduced ensemble model in detecting corona diasese using X-ray images of human chest. As a result, hypothesized that the model's emphasis on identifying unique features that enable differentiation from other categories, like abnormal and regular or normal cases, contributes to its success. Table 3 offers a comparative evaluation of the proposed model against other models for corona detection using human cehst X-Rays, highlighting the model's superior performance.

Table3 Performance of the proposed ensemble model

Method	**Input Classes**	**Totalchest X-ray**	**Precision(%)**	**Recall(%)**	**Accuracy (%)**
COVIDNET	Corona Affected	573	99.0	95.0	94.5
	Normal Xrays	8,066			
	Pneumonia Affected	5559			
Flat Classification with EfficientNet-B3 [10]	Corona Affected	183	100	96.8	93.9
	Normal Xrays	8,066			
	Pneumonia Affected	5,521			
HierarchicalClassification with	Corona Affected	183	100	80.6	93.5
	Normal Xrays	8,066			

EfficientNet-B3[10]	Pneumonia Affected	5,521			
CovXNet	Corona Affected	305	88.5	90.3	89.6
	Pneumonia Affected	305			
	Pneumonia Affected	305			
	Corona Affected	305	90.8	89.9	90.3
	Pneumonia Affected	305			
	Pneumonia Affected	305			
	Normal	+305			
Proposed	Corona Affected	1198	96	96	94.0
	Normal Xrays	1341			
	Pneumonia Affected	1345			

Previously, models like COVID-Net and EfficientNet-B3used ImageNet weights with the COVIDx dataset. In contrast, CovXNet employed an ensemble and transfer learning strategy with separate non-COVID (normal) human chest X-ray images. COVID-Net yielded similar results but was complex, requiring substantial parameter training and computational resources. In a comprehensive analysis, our ensemble model efficiently classified corona disease while minimizing resource use. Some confusion emerged, with seven viral pneumonia images misclassified label as affected due to overlapping features. Additionally, distinguishing viral and normal pneumonia posed challenges due to varying radiological manifestations, including early-stage viral pneumonia intricacies

5 Conclusion and Furture Work

This research paper, introduced an ensemble based predcition model designed to detect coronadisease using human chest X-ray images. Proposed ensemble based prediction model has demonstrated its effectiveness in capturing corona disease specific features, outperforming established methods in terms of classification accuracy. This work, useddatasets whihc has more number of corona images, which significantly contributed to the success of our ensemble precition model. The comprehensive evaluation indicated its superiority over other models, achieving both precision and recall rates of 93%. Additionally, the model displayed robust performance across various metrics, including the weighted averages of precision, recall, F1 scores, and overall accuracy.

In the further research, exploration of the integration of diverse corona datasets(COVID) to bolster the model's resilience and adaptability is planned.Additionally, the model will expand its analysis to encompass short-term historical chest X-ray patterns, providing insights into the potential threat posed by the infection to a patient's well-being. Given the evolving nature of research on radiological markers, upcoming studies will delve deeper into visualization

techniques to gain a more profound understanding of the virus's unique characteristics and attributes.

References

[1] Zafar, T. (2022). The emergence of severe acute respiratory syndrome-coronavirus 2 epidemic and pandemic. In Advanced Biosensors for Virus Detection (pp. 1-18). Academic Press.

[2] Padmanabhan, N., Natarajan, I., Gunston, R., Raseta, M., &Roffe, C. (2021). Impact of COVID-19 on stroke admissions, treatments, and outcomes at a comprehensive stroke centre in the United Kingdom. Neurological Sciences, 42, 15-20.

[3] Yin, J. L., Shackel, N. A., Zekry, A., McGuinness, P. H., Richards, C., Van Der Putten, K., ... & Bishop, G. A. (2001). Real-time reverse transcriptase–polymerase chain reaction (RT–PCR) for measurement of cytokine and growth factor mRNA expression with fluorogenic probes or SYBR Green I. Immunology and cell biology, 79(3), 213-221.

[4] Khemasuwan, D., Sorensen, J. S., & Colt, H. G. (2020). Artificial intelligence in pulmonary medicine: computer vision, predictive model and COVID-19. European respiratory review, 29(157).

[5] Heidari, A., Navimipour, N. J., Unal, M., &Toumaj, S. (2022). The COVID-19 epidemic analysis and diagnosis using deep learning: A systematic literature review and future directions. Computers in biology and medicine, 141, 105141.

[6] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L. C. (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4510-4520).

[7] Rajinikanth, V., Joseph Raj, A. N., Thanaraj, K. P., & Naik, G. R. (2020). A customized VGG19 network with concatenation of deep and handcrafted features for brain tumor detection. Applied Sciences, 10(10), 3429.

[8] Karam, A. F., Helmy, A., & Mohammed, A. (2022, May). An approach to enhance KNN based on data clustering using K-medoid. In 2022 2nd International Mobile, Intelligent, and Ubiquitous Computing Conference (MIUCC) (pp. 1-7). IEEE.

[9] Liu, Y., Wang, Y., & Zhang, J. (2012). New machine learning algorithm: Random forest. In Information Computing and Applications: Third International Conference, ICICA 2012, Chengde, China, September 14-16, 2012. Proceedings 3 (pp. 246-252). Springer Berlin Heidelberg.

[10] Ozturk, Tulin, et al. “Automated detection of COVID-19 cases using deep neural networks with X-ray images.” Computers in Biology and Medicine, Vol. 121, 2020, p. 103792.