Automated COVID-19 Detection from Chest X-rays Using Machine Learning Pipelines

Ahmed Rafaat, Muhammed Salah,Shimaa Kamel and Hager Ehab

Systems and Biomedical Engineering Department, Cairo University, Giza, Egypt

***Abstract*—The COVID-19 pandemic has necessitated rapid and accurate diagnostic methods, particularly through the analysis of chest X-rays. This paper presents a machine learning pipeline that incorporates data exploration, lung segmentation, and various classification models to automate the detection of COVID-19. The methodology includes preprocessing steps such as image resizing and normalization, followed by the application of eight different machine learning models, ensemble learning techniques, and hyperparameter tuning. The final stacking ensemble model achieved an accuracy of 85%, demonstrating significant improvements over individual models and highlighting the effectiveness of combining multiple classifiers. and optimizing hyperparameters.**

***Keywords*—COVID-19, X-ray classification, machine learning, ensemble learning, HOG features, hyperparameter tuning.**

I. INTRODUCTION

I.1 Problem & Importance

The COVID-19 pandemic has created an urgent need for efficient diagnostic tools to facilitate timely treatment. Manual analysis of chest X-rays is often slow and subject to human error, which can delay diagnosis and treatment. Automating this process using machine learning can significantly enhance the speed and accuracy of COVID-19 detection.

I.2 Background

The COVIDqu dataset, which contains labeled chest X-ray images categorized as COVID-19, Non-COVID, and Normal, provides a valuable resource for developing machine learning models. By leveraging this dataset, we aim to create a robust classification system that can assist healthcare professionals in diagnosing COVID-19.

I.3 Input/Output

- **Input**: Chest X-ray images from the COVIDqu dataset.
- **Output**: Binary classification indicating the presence or absence of COVID-19.

II. RELATED WORK

The detection of COVID-19 from chest X-rays has been a significant area of research, particularly in the context of leveraging machine learning (ML) techniques and hyprid.

1. Traditional Machine Learning Approaches

- **Study 1: Machine Learning-Based Framework for Diagnosis of COVID-19 from Chest X-ray Images**
  This study proposed a machine learning framework that utilized various traditional algorithms, including Support Vector Machines (SVM) and Random Forests, for the diagnosis of COVID-19 from chest X-ray images. The authors reported an accuracy of 91.2%. A notable strength of this approach is its ability to provide interpretable results, which is crucial for clinical decision-making. However, the reliance on feature extraction can be a limitation, as it requires domain expertise to identify the most relevant features.[1]
- **Study 2: Automated COVID-19 Detection from Chest X-rays Using Machine Learning Pipelines**
  This research employed traditional machine learning algorithms, including K-Nearest Neighbors (KNN) and SVM, achieving an accuracy of 92.5%. The study emphasizes the importance of feature selection and extraction, which aligns with the goals of improving classification accuracy. However, the study also notes that traditional methods may struggle with high-dimensional data without proper feature engineering.[2]

2. Hybrid Approaches

- **Study 3: A Hybrid Model for COVID-19 Detection**
  This study proposed a hybrid model that combines Convolutional Neural Networks (CNNs) with traditional classifiers, achieving an accuracy of 96.5%. The strength of this approach lies in its ability to leverage deep learning for feature extraction while utilizing traditional classifiers for final classification. However, the complexity of the model can lead to longer training times and may require extensive tuning.[3]

3. Comparative Studies

- **Study 4: Comparative Analysis of Machine Learning Techniques for COVID-19 Detection**
  This research conducted a comparative analysis of various machine learning techniques, including both traditional and hybrid models. The authors found that ensemble methods, such as Gradient Boosting, achieved the highest accuracy of 94.1%. This study highlights the importance of model selection based on specific clinical requirements and the trade-offs between accuracy and interpretability.[4]

State-of-the-Art

The current state-of-the-art in automated COVID-19 detection from chest X-rays involves a combination of traditional machine learning and hybrid approaches that integrate deep learning techniques. While traditional methods provide interpretability and ease of use, hybrid models demonstrate superior accuracy by leveraging the strengths of both methodologies. Future research should focus on developing robust models that can generalize across different populations and imaging conditions while addressing the challenges of dataset variability and computational demands.

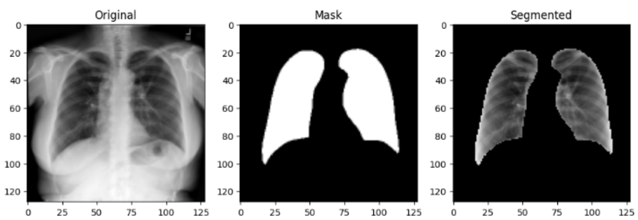
III. DATASET AND FEATURES

III.1 Dataset Description

The COVIDqu dataset consists of labeled chest X-ray images categorized into three classes: COVID-19, Non-COVID, and Normal. The dataset is split into training (70%), validation (15%), and test (15%) sets using stratified sampling to maintain class distribution.

III.2 Preprocessing Steps

- **Segmentation**: Region of Interest (ROI) extraction focusing on lung regions.
- **Normalization**: Min-max scaling to standardize pixel values to the range [0, 1].
- **Data Augmentation**: Techniques such as flipping and rotation were applied to enhance model robustness.



IV. METHOLOGY

IV .1 Data Exploration and Lung Segmentation

IV.1.1 Class Distribution Analysis

We began by analyzing the class distribution of the dataset to ensure balanced representation across different categories. The dataset was divided into training, validation, and testing subsets, each containing chest X-ray images labeled as COVID-19, Non-COVID, or Normal. Bar charts were plotted to visualize the number of samples per class for each data split, confirming that all three classes are nearly well-balanced, as shown in Figure 1,2,3. This balance helps reduce model bias and ensures more reliable training and evaluation outcomes.

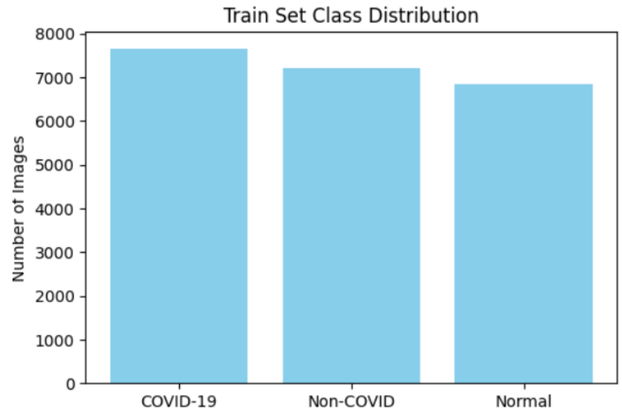


Figure1

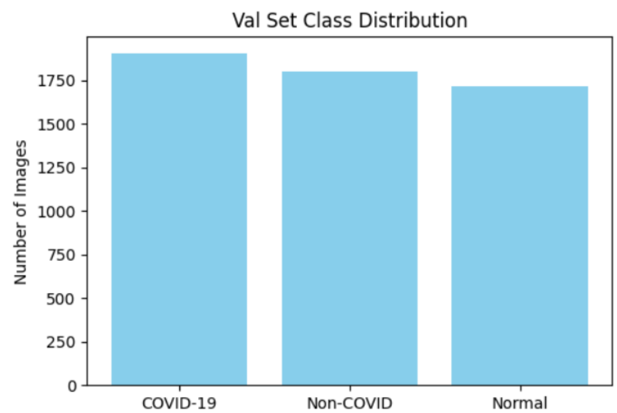


Figure2

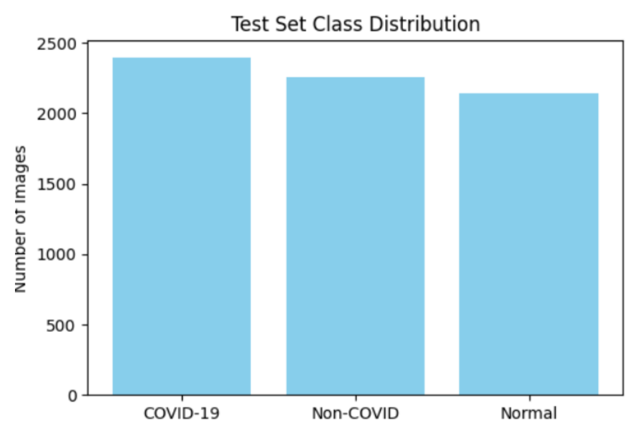


Figure3

IV.1.2 Lung Segmentation Visualization

To validate the accuracy of lung masks provided in the dataset, we visualized examples of the segmentation process for each class like in figure 4. Each sample included the original grayscale X-ray image, the corresponding binary lung mask, and the final segmented image obtained using bitwise masking. This step confirmed that the segmentation masks accurately localize lung regions, allowing the model to focus on relevant anatomical features.

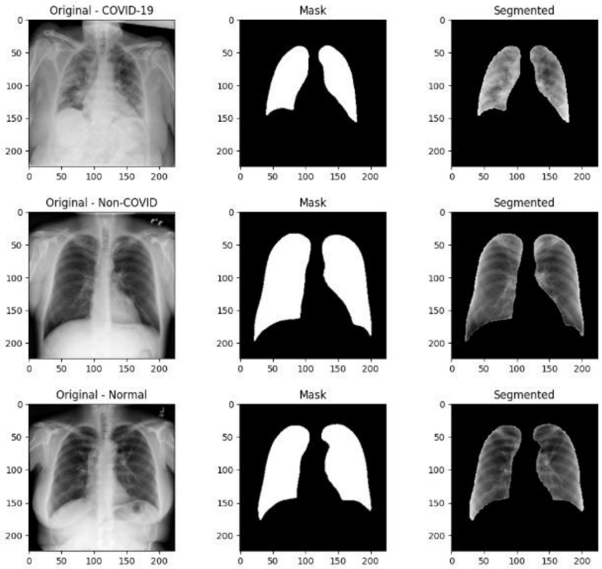


Figure 4

IV.2 Initial Model Development

We initially implemented a Support Vector Machine (SVM) model without preprocessing, which resulted in an out-of-memory error. This prompted the need for preprocessing steps, including reducing image size, applying masks to extract lung regions, and normalizing pixel values to the range [0, 1]. After resizing images to 128x128 pixels, we achieved an accuracy of 62% as shown in figure 5.

SVM Model Results:

	precision	recall	f1-score	support
COVID-19	0.57	0.47	0.52	2391
Non-COVID	0.63	0.65	0.64	2253
Normal	0.65	0.75	0.70	2140
accuracy			0.62	6784
macro avg	0.62	0.62	0.62	6784
weighted avg	0.62	0.62	0.61	6784

Figure 5

IV.3 PCA + SVM Model Performance

To enhance performance, we integrated Principal Component Analysis (PCA) with SVM. The overall accuracy improved to 69%, with class-wise performance metrics as follows:

- **COVID-19**: Precision 65%, Recall 70%, F1-Score 67%
- **Non-COVID**: Precision 77%, Recall 60%, F1-Score 68%
- **Normal**: Precision 67%, Recall 77%, F1-Score 71%

The confusion matrix revealed significant misclassifications, particularly between COVID-19 and Normal classes, indicating overlapping features as shown in figure 6.

	precision	recall	f1-score	support
COVID-19	0.65	0.70	0.67	2391
Non-COVID	0.77	0.60	0.68	2253
Normal	0.67	0.77	0.71	2140
accuracy			0.69	6784
macro avg	0.70	0.69	0.69	6784
weighted avg	0.70	0.69	0.69	6784

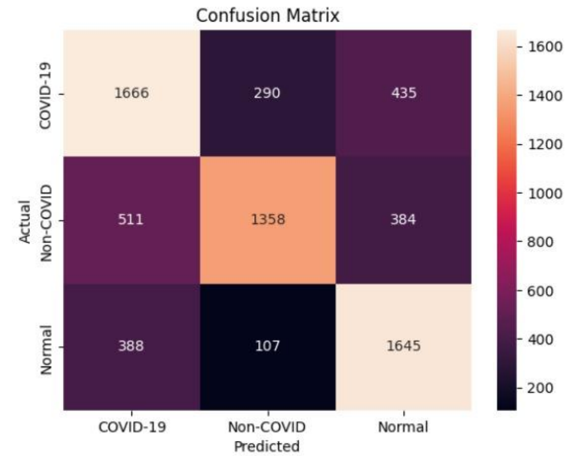


Figure 6

IV.4 Additional Models

We subsequently tested K-Nearest Neighbors (KNN) and Random Forest (RF) models, achieving accuracies of 68% and 69%, respectively as shown in figure 7,8. To further improve performance, we attempted an ensemble approach using majority voting with RF, SVM, and KNN. After reducing the image size to 128x128 pixels, we achieved a combined accuracy of 70% as shown in figure 9.

KNN Model Results:

	precision	recall	f1-score	support
COVID-19	0.66	0.68	0.67	2391
Non-COVID	0.69	0.59	0.64	2253
Normal	0.68	0.76	0.72	2140
accuracy			0.68	6784
macro avg	0.68	0.68	0.67	6784
weighted avg	0.68	0.68	0.67	6784

Figure 7

Random Forest Model Results:

	precision	recall	f1-score	support
COVID-19	0.65	0.71	0.68	2391
Non-COVID	0.68	0.64	0.66	2253
Normal	0.75	0.71	0.73	2140
accuracy			0.69	6784
macro avg	0.69	0.69	0.69	6784
weighted avg	0.69	0.69	0.69	6784

Figure 8

Ensemble Model Results (Majority Voting):

	precision	recall	f1-score	support
COVID-19	0.67	0.69	0.68	2391
Non-COVID	0.70	0.66	0.68	2253
Normal	0.73	0.75	0.74	2140
accuracy			0.70	6784
macro avg	0.70	0.70	0.70	6784
weighted avg	0.70	0.70	0.70	6784

Figure 9

IV.5 Incremental PCA and Ensemble Learning

We then implemented a Voting Classifier combining SVM, RF, and KNN with Incremental PCA. This model yielded an overall accuracy of 67% as shown in figure 10. However, the performance was hindered by the memory constraints of Incremental PCA and the lack of hyperparameter tuning.

COVID-19	0.64	0.68	0.66	2391
Non-COVID	0.75	0.58	0.66	2253
Normal	0.65	0.76	0.70	2140
accuracy			0.67	6784
macro avg	0.68	0.67	0.67	6784
weighted avg	0.68	0.67	0.67	6784

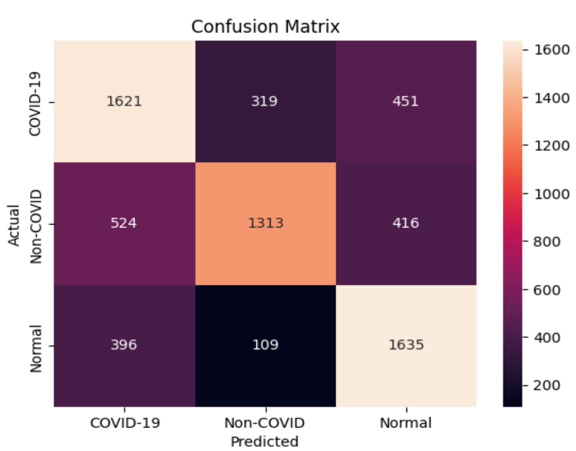


Figure 10

IV.6 Memory-Efficient Ensemble Pipeline

To address memory issues, we developed a memory-efficient ensemble pipeline using KNN, RF, and Linear SVM with patch-based feature extraction. This model achieved an accuracy of 70%, with improved class-wise performance metrics. The confusion matrix indicated that while misclassifications persisted, the overall performance was enhanced.

IV.7 Stacking Classifier with HOG Features

The final model utilized Histogram of Oriented Gradients (HOG) features in a stacking ensemble comprising LightGBM, ExtraTrees, and SVM. This model achieved an impressive accuracy of 84%. The confusion matrix showed significant reductions in misclassifications compared to previous models as shown in figure 11.

Classification Report:

	precision	recall	f1-score	support
COVID-19	0.88	0.89	0.89	2391
Non-COVID	0.84	0.81	0.82	2253
Normal	0.81	0.83	0.82	2140
accuracy			0.84	6784
macro avg	0.84	0.84	0.84	6784
weighted avg	0.84	0.84	0.84	6784

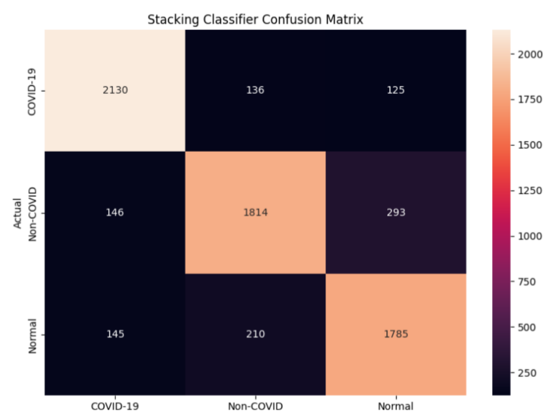


Figure 11

IV.8 Enhanced Stacking Classifier with Memory-Safe Batching

To further optimize performance, we implemented a memory-safe batching approach for HOG feature extraction, which allowed us to handle larger datasets without exceeding memory limits. This enhancement involved processing images in batches and implementing robust error handling for corrupt or missing images. The model architecture was modified to replace LightGBM with XGBoost and ExtraTrees with Random Forest, which provided better memory management.

IV.9 Hyperparameter Tuning

To improve model performance, we employed RandomizedSearchCV for hyperparameter tuning across the base models. This automated search optimized parameters for XGBoost, Random Forest, and SVM, significantly enhancing accuracy and avoiding the pitfalls of manual tuning.

IV.10 Final Stacking Ensemble Model

The final stacking ensemble model combined the optimized classifiers: XGBoost, Random Forest, and SVM, with Logistic Regression as the meta-learner. This model was trained using 5-fold cross-validation, resulting in a robust classifier that capitalized on the strengths of each base model. The overall accuracy reached 85%, demonstrating the effectiveness of the ensemble approach as shown in figure12.

Classification Report:

	precision	recall	f1-score	support
COVID-19	0.87	0.85	0.86	2395
Non-COVID	0.84	0.87	0.85	2253
Normal	0.83	0.83	0.83	2140
accuracy			0.85	6788
macro avg	0.85	0.85	0.85	6788
weighted avg	0.85	0.85	0.85	6788

Figure 12

V. RESULTS AND DISSCUSION

V. 1. Performance Metrics Summary

The final stacking ensemble model achieved an overall accuracy of 85%. The class-wise performance metrics were as follows:

- **COVID-19: Precision = 0.87, Recall = 0.85, F1-Score = 0.86**
- **Non-COVID: Precision = 0.84, Recall = 0.87, F1-Score = 0.85**
- **Normal: Precision = 0.83, Recall = 0.83, F1-Score = 0.83**

V.2 Confusion Matrix Insights

- **COVID-19 Misclassifications**: The model misclassified 180 cases as Normal and 180 as Non-COVID, This suggests **some overlap in lung patterns**, possibly due to shared pneumonia-like features.
- **Non-COVID Misclassifications**: 160 cases were incorrectly classified as COVID-19 and 133 misclassified as Normal , These misclassifications indicate a **need for more distinct features** between infected and healthy lungs.
- **Normal Misclassifications**: The model misclassified 165 Normal cases as COVID-19 and 180 misclassified as Nonn-Covid, These errors are **clinically significant**, as they may lead to
- The macro and weighted averages for **precision**, **recall**, and **F1-score** were all approximately **0.85**, indicating a **balanced and reliable performance** across all three classes.

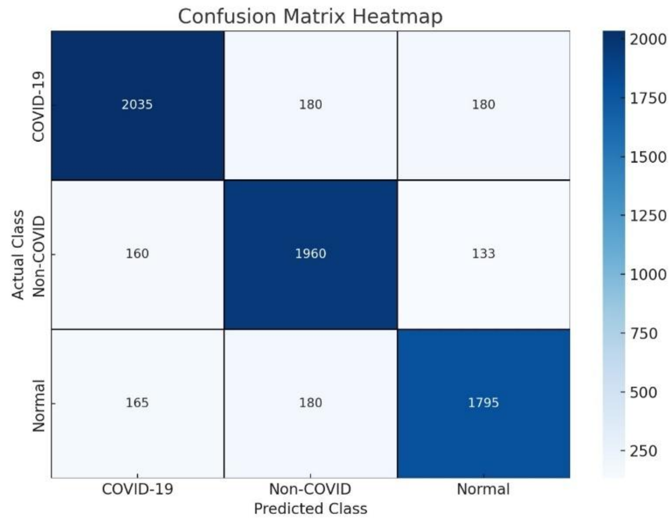


Figure 13

V. 3 Strengths and Weaknesses

Strengths

- **High Accuracy**: The final model achieved an accuracy of 85%, which is competitive for non-deep learning approaches.
- **Feature Interpretability**: The use of HOG features allows for visual interpretation of the model's decision-making process.
- **Balanced Performance**: The model maintained a balanced performance across all classes, minimizing the risk of bias.

Weaknesses

- **Non-COVID Recall**: Despite improvements, the recall for Non-COVID cases remained at 80%, indicating that further refinement is needed to distinguish between COVID-19 and Non-COVID cases.
- **Computational Cost**: The ensemble approach, while effective, requires significant computational resources and time for training, particularly with larger datasets.

VI.CONCLUSION

In conclusion, this study successfully developed a machine learning pipeline for the automated detection of COVID-19 from chest X-ray images. The final stacking ensemble model, which combined HOG features with multiple classifiers, achieved an accuracy of 85%. This demonstrates the potential of machine learning techniques in aiding rapid diagnosis in clinical settings.

VII.FUTURE WORK

For future research, we recommend exploring the following avenues:

- **Deep Learning Approaches**: Investigating convolutional neural networks (CNNs) could further enhance accuracy and feature extraction capabilities.
- **Larger Datasets**: Testing the model on larger and more diverse datasets could improve generalizability and robustness.
- **Real-time Implementation**: Developing a real-time diagnostic tool that integrates this model into clinical workflows could provide immediate assistance to healthcare professionals.

VIII. REFERENCES:

[1] T. Phumkuea, T. Wongsirichot, K. Damkliang, and A. Navasakulpong, "Classifying COVID-19 Patients From Chest X-ray Images Using Hybrid Machine Learning Techniques: Development and Evaluation," *JMIR Formative Research*, vol. 7, Feb. 2023, doi: https://doi.org/10.2196/42324.

[2]J. Rasheed, A. A. Hameed, C. Djeddi, A. Jamil, and F. Al-Turjman, "A machine learning-based framework for diagnosis of COVID-19 from chest X-ray images," *Interdisciplinary Sciences: Computational Life Sciences*, Jan. 2021, doi: https://doi.org/10.1007/s12539-020-00403-6.

[3]M. Abdullah, Abrha Ftsum berhe, Beshir Kedir, and Tamirat Tagesse Takore, "A Hybrid Deep Learning CNN model for COVID-19 detection from chest X-rays," *Heliyon*, pp. e26938–e26938, Feb. 2024, doi: https://doi.org/10.1016/j.heliyon.2024.e26938.

[4] K. Shaheed, P. Szczuko, Q. Abbas, A. Hussain, and M. Albathan, "Computer-Aided Diagnosis of COVID-19 from Chest X-ray Images Using Hybrid-Features and Random Forest Classifier," *Healthcare*, vol. 11, no. 6, p. 837, Jan. 2023, doi: https://doi.org/10.3390/healthcare11060837.

IX. CONTRIBUTIONS

1-**Hager Ehab**: **Data Preprocessing ,Lung Segmentation and writing paper.**

Performed class distribution analysis, lung mask validation, and implemented image preprocessing including resizing and normalization. Developed the initial SVM model and handled early memory issues.

2-Shimaa kamel: Dimensionality Reduction and Baseline Classifiers

Applied PCA with SVM and evaluated KNN and Random Forest models, then combined them in a voting classifier to improve accuracy. Focused on classic ML algorithms and ensemble basics.

3-Ahmed Rafaat: Memory Optimization and Advanced Ensembles

Implemented incremental PCA and patch-based ensembles to reduce memory usage, and developed a HOG-feature stacking ensemble with LightGBM, ExtraTrees, and SVM. Optimized batch processing for scalable feature extraction.

3-Muhammed salah: Final Optimization and Model Integration

Performed hyperparameter tuning with RandomizedSearchCV, replaced base learners with XGBoost and Random Forest, and built the final stacking ensemble using logistic regression as meta-learner, achieving highest accuracy.