Metadata of the chapter that will be visualized in SpringerLink

Book Title	Bioinformatics and Computational Biology	
Series Title		
Chapter Title	Explainable Convolutional Neural Network for COVID-19 Detection	
Copyright Year	2025	
Copyright HolderName	The Author(s), under exclusive license to Springer Nature Switzerland AG	
Author	Family Name	**Sam**
	Particle	
	Given Name	**Maxwell**
	Prefix	
	Suffix	
	Role	
	Division	Department of Computer Science
	Organization	North Carolina A&T State University
	Address	Greensboro, NC, USA
	Email	msam@aggies.ncat.edu
Author	Family Name	**Annan**
	Particle	
	Given Name	**Richard**
	Prefix	
	Suffix	
	Role	
	Division	Department of Computer Science
	Organization	North Carolina A&T State University
	Address	Greensboro, NC, USA
	Email	rkannan@aggies.ncat.edu
Author	Family Name	**Rhinehardt**
	Particle	
	Given Name	**Kristen**
	Prefix	
	Suffix	
	Role	
	Division	Department of Computational Data Science and Engineering
	Organization	North Carolina A&T State University
	Address	Greensboro, NC, USA
	Email	klrhineh@ncat.edu
Author	Family Name	**Roy**
	Particle	
	Given Name	**Kaushik**
	Prefix	
	Suffix	

	Role	
	Division	Department of Computer Science
	Organization	North Carolina A&T State University
	Address	Greensboro, NC, USA
	Email	kroy@ncat.edu
Author	Family Name	**Tang**
	Particle	
	Given Name	**Guoqing**
	Prefix	
	Suffix	
	Role	
	Division	Department of Mathematics and Statistics
	Organization	North Carolina A&T State University
	Address	Greensboro, NC, USA
	Email	tang@ncat.edu
Corresponding Author	Family Name	**Qingge**
	Particle	
	Given Name	**Letu**
	Prefix	
	Suffix	
	Role	
	Division	Department of Computer Science
	Organization	North Carolina A&T State University
	Address	Greensboro, NC, USA
	Email	lqingge@ncat.edu

Abstract	In the event of the COVID-19 outbreak, prompt and accurate diagnosis became critical for both public health interventions and efficient patient care. COVID-19 is a disease that affects the upper and lower respiratory tract and can have fatal consequences. Early diagnosis is crucial for effective treatment and containment. Studies have shown that COVID-19 manifests in the chest of infected patients, prompting the computer vision community to explore the use of CT scans and deep learning-based solutions for diagnosis. However, efforts to implement explainable artificial intelligence (AI) for interpreting deep learning models in COVID-19 recognition are still scarce. In this paper, we apply SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model- agnostic Explanations) techniques to enhance the interpretability of our developed CNN that used to detect COVID-19 from CT scan images. The dataset is consist of 4649 images, of which 2476 are from patients with COVID-19 and 2173 are from patients without COVID-19 was used in our implementation. We applied SHAP and LIME techniques to identify important features of COVID-19 images, which even improved the performance of our original model. The comparison results with other baseline models show the robustness of our proposed model and identified important features. We also find that the explainable ability of the SHAP and LIME techniques also depends on its model prediction accuracy.

Keywords (separated by '-')	SHapley Additive exPlanations - Local Interpretable Model- agnostic Explanations - COVID-19 Detection - Convolutional Neural Network

Explainable Convolutional Neural Network for COVID-19 Detection

Maxwell Sam[1], Richard Annan[1], Kristen Rhinehardt[2], Kaushik Roy[1], Guoqing Tang[3], and Letu Qingge[1(✉)]

[1] Department of Computer Science, North Carolina A&T State University, Greensboro, NC, USA
{msam,rkannan}@aggies.ncat.edu, {kroy,lqingge}@ncat.edu
[2] Department of Computational Data Science and Engineering, North Carolina A&T State University, Greensboro, NC, USA
klrhineh@ncat.edu
[3] Department of Mathematics and Statistics, North Carolina A&T State University, Greensboro, NC, USA
tang@ncat.edu

Abstract. In the event of the COVID-19 outbreak, prompt and accurate diagnosis became critical for both public health interventions and efficient patient care. COVID-19 is a disease that affects the upper and lower respiratory tract and can have fatal consequences. Early diagnosis is crucial for effective treatment and containment. Studies have shown that COVID-19 manifests in the chest of infected patients, prompting the computer vision community to explore the use of CT scans and deep learning-based solutions for diagnosis. However, efforts to implement explainable artificial intelligence (AI) for interpreting deep learning models in COVID-19 recognition are still scarce. In this paper, we apply SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model- agnostic Explanations) techniques to enhance the interpretability of our developed CNN that used to detect COVID-19 from CT scan images. The dataset is consist of 4649 images, of which 2476 are from patients with COVID-19 and 2173 are from patients without COVID-19 was used in our implementation. We applied SHAP and LIME techniques to identify important features of COVID-19 images, which even improved the performance of our original model. The comparison results with other baseline models show the robustness of our proposed model and identified important features. We also find that the explainable ability of the SHAP and LIME techniques also depends on its model prediction accuracy.

AQ1

Keywords: SHapley Additive exPlanations · Local Interpretable Model-agnostic Explanations · COVID-19 Detection · Convolutional Neural Network

This work is supported by the U.S. National Science Foundation under award 2434487 and U.S. National Institutes of Health U24 HG013013.

T. Aldwairi et al. (Eds.): BICOB 2025, CCIS 2535, pp. 1–13, 2025.
https://doi.org/10.1007/978-3-031-94039-2_17

1 Introduction

The COVID-19 pandemic has underscored the critical importance of accurate and timely medical diagnostics. Advanced machine learning techniques, particularly deep learning models, have shown immense potential in aiding the diagnosis of COVID-19 from medical imaging, such as chest X-rays and CT scans [1, 2]. However, the complexity of these models often renders them as "black boxes," where the decision-making process is not transparent to medical practitioners. This lack of interpretability can hinder clinical trust and adoption, as understanding the rationale behind a model's predictions is crucial in the medical domain [3].

To address this challenge, explainable artificial intelligence (XAI) techniques have been developed to provide insights into the decision-making processes of complex models. Among these techniques, SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) have gained prominence [3, 4]. SHAP leverages cooperative game theory to attribute the contribution of each feature towards the model's prediction, ensuring consistency and local accuracy. LIME, on the other hand, approximates the original model locally with an interpretable model, thereby providing insights into individual predictions.

The application of SHAP and LIME to COVID-19 imaging aims to eluci- date how deep learning models derive their conclusions, highlighting the critical features in medical images that drive diagnostic decisions. By doing so, these techniques not only enhance model transparency but also provide a valuable tool for medical professionals to validate and trust machine learning outputs.

The contributions and novelty of this paper are as follows. First, we built and fine-tuned our own developed COVID-CNN model to accurately detect COVID- 19 from CT scan images. Second, we applied XAI techniques, such as SHAP and LIME to identify important features that contributed to the model prediction. Third, we rerun our COVID-CNN model on identified COVID-19 images from XAI techniques and achieve higher prediction accuracy than the original CNN model. Fourth, we compared our CNN model performance with other baseline models, such as VGG-16, VGG-19, Deep-COVID, Deep-COVID DeteCT models. Fifth, we found out although SHAP and LIME identify important features from the input data for the model prediction, the explainable ability of those techniques also depend on model prediction accuracy. If model performance is lower on given dataset, even if we use XAI techniques to identify important features from the input data, that might not be useful and accurate for our prediction task.

The organization of our paper is as follows. In Sect. 2, we review previous related works using model-agnostic techniques to identify relevant features for COVID-19 predictions. Section 3 covers the dataset and provides details on our developed CNN model. In Sect. 4, we apply SHAP and LIME techniques to interpret the model. Section 5 presents the results and Sect. 6 concludes the paper.

2 Related Work

Recent advancements in machine learning (ML) and deep learning (DL) have significantly enhanced computer-aided medical diagnosis, particularly in the analysis of Chest-XR images and CT scans for detecting diseases like COVID-19, pneumonia, and tuberculosis (TB). This overview delves into the latest research findings, focusing on the utilization of medical image analysis, particularly Chest- XR images, to diagnose these critical illnesses.

According to [5], Chest X-ray (CXR) images are considered useful in diagnosing pulmonary disorders such as COVID-19, Pneumonia, and Tuberculosis (TB). A deep learning (DL) model was proposed to enhance disease recognition accuracy while maintaining effective feature extraction. Based on publicly available dataset comprising 7132 CXR images, their model achieved high average test accuracy of 94.31% with a margin of error of about 1.01% whiles a validation accuracy of 94.54% also with a margin of error of about 1.33% through 10-fold cross- validation. Interpretation of their results was facilitated by Gradient-weighted Class Activation Mapping (Grad-CAM), Local Interpretable Model-agnostic Explanation (LIME) and SHapley Additive exPlanation (SHAP). These techniques provided insights into the DL model's decision-making process. Inaddition, they as well made use of eXplainable Artificial Intelligence (XAI) techniques to consolidate and validate model-generated explanations, offering clinicians and medical professionals coherent insights into disease detection and categorization of COVID-19, Pneumonia and Tuberculosis. Authors in [6] also proposed and compared the LIME and SHAP techniques to enhance the interpretation of COVID diagnosis through X-ray scans. In their findings, they first applied SqueezeNet to recognise pneumonia, COVID-19 and normal lung image. Through SqueezeNet, an 84.34% recognition rate success in testing accuracy was obtained. Shapley Additive Explanation (SHAP) and Local Interpretable Model-Agnostic Explanations (LIME) were used to explain and interpret how Squeezenet achieved classification in order to gain a better understanding of what the network observes in relation to a particular task, namely image classification. Their results indicated LIME and SHAP area able to indicate the area of interest and as well help to improve the transparency and the interpretability of the Squeezenet model.

Ashan et al. [7] conducted to detect COVID 19 patients from CXR and CT images and implemented six deep CNN learning models, including VGG16, MobileNetV2, InceptionResNetV2, ResNet50, ResNet101 and VGG19 using 400 CXR and 400 CT images. Achieving an average accuracy of 82.94% on a dataset with CT images and 93.94% on a dataset with CXR images, MobileNetV2 outweigh NasNetMobile. The overall prediction was explained by applying the heatmap of class activation and analyzing the feature extraction implementing LIME. Manjurul et al. [8] achieved high performance on VGG16 model ($98.5 \pm 1.19\%$) among six different Deep CNN models: VGG16, MobileNetV2, InceptionResNetV2, ResNet50, ResNet101 and VGG19 with mixed dataset of CT and X-ray images to classify COVID-19 patients. Results were further explained with LIME. Regarding Sarp et al. [9], they proposed (XAI) technique to detect and interpret COVID-19 positive CXR images. Six deep learning models i.e. VGG16, VGG19, ResNet, InceptionV3, COVID-Net and CORODET were implemented using public dataset collected on GitHub. Their models achieved an accuracy of 96%, 96%, 98%, 91%, 99% and 98% respectively. LIME and SHAP

were applied of these models to give a better understanding of the prediction accuracies. To˘ga¸car et al. [10] approach involved three key steps. Firstly, a preprocessing technique to include Fourier Transform and Gradient-weighted Class Activation Mapping to the input images. Secondly, type-based activation sets were generated using three ResNet models before employing the Softmax method. Lastly, the most effective type-based activations are selected using the local interpretable model-agnostic explanations method and re-classified using Softmax. The proposed approach achieved an overall accuracy success of 99.15% across a dataset containing three types of image sets, and a remarkable 99.62% accuracy success specifically for COVID-19 findings. Authors in [11] proposed a comprehensive method for enhancing the interpretability of predictions generated by Convolutional Neural Networks (CNNs) in medical imaging, leveraging Explainable Artificial Intelligence (XAI) techniques. Their approach integrates various techniques, including LIME (Local Interpretable Model Agnostic Explanations), integrated gradients, Anchors and SHAP (Shapley Additive Explanations). Shapley values was used to explain the decisions of their model. Their proposed CNN model achieved a testing accuracy of 90%. XAI techniques was proposed on Deep Learning model for predicting brain tumour status using MRI images data [12]. Their findings presented an interpretable deep learning model for predicting brain tumor types (meningioma, glioma, pituitary) from MRI images. Using a dual-input CNN with Gaussian noise, their model achieved a 94.64% training accuracy and an overall testing accuracy of 85.37%. LIME and SHAP were employed for local and global interpretability to have a clear understanding of their model's prediction. The authors in [13] proposed an explainable framework for detecting spam images using Convolutional Neural Network (CNN) algorithms and Explainable Artificial Intelligence (XAI) algorithms. In their findings their proposed CNN model was used to classify image spam respectively whereas the post-hoc XAI techniques including Local Interpretable Model Agnostic Explanation (LIME) and Shapley Additive Explanations (SHAP) were deployed to provide explanations for the decisions that the black-box CNN models made about spam image detection. Their proposed CNN model achieved a training accuracy of 91% on 6636 image dataset including spam images and normal images collected publicly.

3 Dataset and Our CNN Model

In this study, we fine-tune our own developed convolutional neural network (CNN) model for COVID-19 prediction from CT scan images [14]. Then we utilize SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) for our CNN's model interpretability. SHAP values are computed for each pixel to determine their contribution to the model's output, offering a comprehensive global explanation to the model's predictions. On the other hand, LIME is being leveraged to generate local explanations by perturbing individual images and fitting an interpretable model to approximate the CNN's behavior around each prediction. This results in the identification of superpixels that significantly influence the model's decision.

3.1 The Dataset

The training dataset utilized in this paper was sourced from radiology centers at teaching hospitals in S˜ao Paulo, Brazil [15], and Tehran, Iran [16]. These datasets were combined to form a balanced dataset, designated as "Dataset Mod Dev" for model development. An additional dataset, "Dataset Mod Gen," was compiled from various countries including Russia, China, Italy, Turkey, and Iran [17] to evaluate the model's generalization ability. Dataset Mod Dev contains 4,649 images, with 2,476 from COVID-19 patients and 2,173 from non- COVID-19 patients. Dataset Mod Gen, a more diverse dataset previously used in [17], includes 14,486 images, comprising 7,593 COVID-19 cases and 6,893 non-COVID-19 cases (Fig. 1).

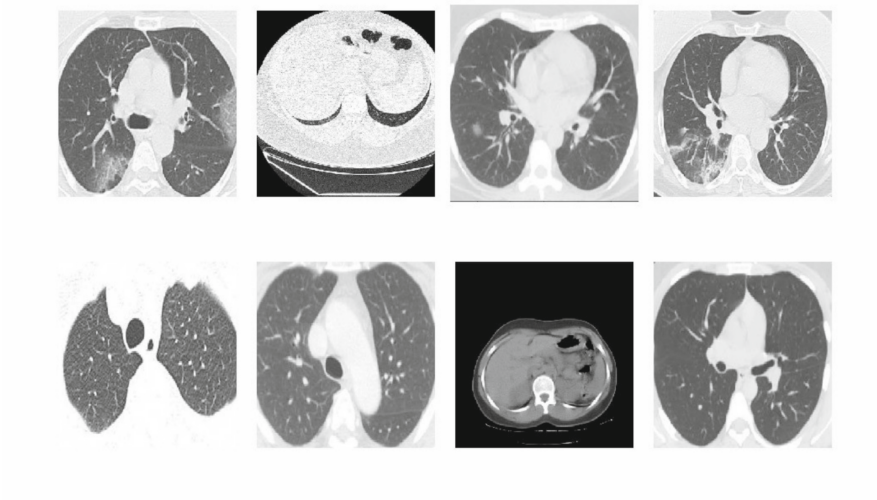


Fig. 1. Example CT-scan of patients with positive COVID-19 in the top row and negative COVID-19 cases in the bottom row.

3.2 The Proposed CNN Model

The COVID-CNN model as shown in Fig. 2 from [14] was used for the application of explainable AI for COVID-19 detection from CT scan images. The COVID-CNN model was specifically designed for grayscale images with dimensions of $300 \times 300 \times 1$. The model's processing begins with 116 feature maps (filters) in the first convolutional layer, which has a kernel size of 8×8 and a stride of 22. The output from this initial convolution has dimensions of $97 \times 97 \times 116$. Following the convolutional layer, a pooling layer with a stride of 2×2 was applied to downsample the feature maps, reducing the spatial dimensions while preserving key information from the previous layer. These steps were essential for capturing localized and hierarchical patterns in the images. Batch

normalization was also applied to the first convolutional layer to stabilize the training process. The second convolutional layer, which also includes batch normalization and max-pooling, mirrors the parameters of the first layer, using 116 filters, an 8×8 kernel, and a 2×2 stride. This results in an output with dimensions of $10 \times 10 \times 116$. This serves as the input to the fully connected layer. The fully connected layer consists of four layers with 362, 184, 78, and 12 neurons, respectively. The ReLU activation function was used in these layers, and dropout regularization applied to prevent overfitting. For the final layer of the COVID- CNN model, the output layer with a softmax activation function and a size of 2 was used for classifying the input image. The model's optimal hyperparameters included a learning rate of 0.001, the ADAM optimizer, and categorical cross-entropy as the loss function.

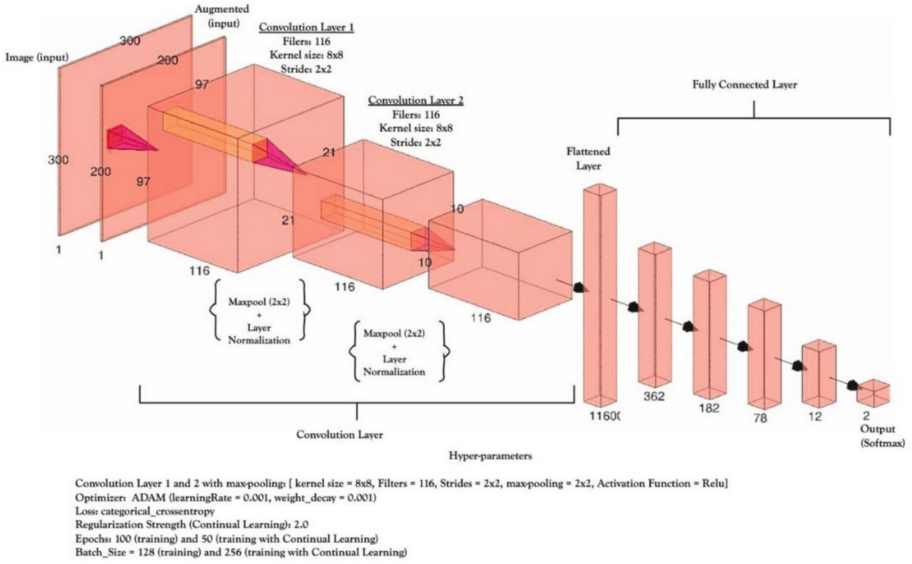


Fig. 2. COVID-CNN model architecture for COVID-19 detection [14].

4 Methodology

Interpretability of Convolutional Neural Network (CNN) on image data has grown to be a critical problem, particularly in high-stakes industries like healthcare, banking, and autonomous systems where decision-making transparency is essential. Notwithstanding their superior performance and accuracy, CNNs are frequently referred to as "black boxes". The term "black box" is mainly because of CNN's opaque and complicated internal workings that makes it challenging to determine which inputs causes which outputs. This lack of interpretability raises challenges in validating, trusting, and deploying CNN models in scenarios where understanding the rationale behind a prediction is as important as the prediction itself. Advanced model interpretability techniques like SHAP

(SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) are used to improve the transparency of CNNs. SHAP leverages game-theoretic principles to assign importance scores to individual regions of an image, offering a global perspective on how each region contributes to the model's overall prediction. This method quantifies the impact of each pixel or superpixel, thereby elucidating the model's decision-making process in a comprehensive manner. Conversely, LIME provides local interpretability by approximating the CNN's behavior with an interpretable surrogate model, specifically tailored to each input image. This approach highlights the critical regions that most significantly influenced the CNN's output, allowing for a more granular understanding of the model's reasoning on a per-instance basis.

4.1 SHAP

SHAP (SHapley Additive exPlanations) is a technique in machine learning for explaining the output of a model by attributing the contribution of each feature to the final prediction. SHAP values are based on cooperative game theory [4]. The primary objective of SHAP values is to provide interpretable and consistent explanations for complex models, thereby enhancing model transparency and trustworthiness. To be able to have much better understanding to which features are positively affecting our models' prediction, we apply the SHAP library to our image data. In this paper, SHAP is used to identify and visualize the contribution of individual pixels or region of our images to the model's prediction. Regions with higher SHAP values are considered important regions. From [4], SHAP can be mathematically expressed as:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} (f(S \cup \{i\}) - f(S)) \tag{1}$$

where: ϕ_i is the SHAP value for feature i, N is the set of all features, S is a subset of features excluding i, $f(S)$ is the model's prediction based on the subset S, $|S|$ is the size of subset S, and $|N|$ is the total number of features that are not included in set S

$$val_x(S) = \int_{\hat{f}(x_1, \ldots, x_p)} dP_{x \notin S} - \mathbb{R}[\hat{f}(X)] \quad [4] \tag{2}$$

4.2 LIME

For image classification tasks, LIME generates local explanations by systematically perturbing the input image and analyzing the model's response to these perturbations to identify which regions of the image are most influential in the model's prediction. LIME for images functions differently compared to its application in tabular or text data. In images, instead of perturbing individual pixels which would have little effect since predictions are typically influenced by larger regions, LIME segments the image into superpixels. A superpixel is a group of connected pixels that share similar properties, such as color or texture. Perturbations are then introduced by randomly masking or turning off these superpixels, often replacing them with a neutral value.

Once the perturbed images are created, the original model is queried to generate predictions for each modified instance. By correlating how the presence or absence of certain superpixels affects the model's output, LIME constructs a linear surrogate model to estimate the contribution of each superpixel to the final prediction. This locally interpretable model assigns importance scores to superpixels, allowing for a clear visualization of the most critical regions that influenced the model's decision. The use of superpixels ensures that explanations are meaningful, as they capture the collective contribution of spatially coherent regions rather than individual pixels, which might not independently influence the outcome. The entire process of LIME on images can be expressed as:

$$\arg \min_{g \in G} \sum_{i=1}^{m} \pi_x(z'_j)(f(h_x(z'_j)) - g(z'_j))^2 + \Omega(g) \tag{3}$$

where:

- $f(x'_i)$: Prediction of the black-box model for the perturbed image x'_i.
- $g(z'_j; \mathbf{w})$: Prediction of the surrogate model (linear model) for the binary vector representation z'_j of the perturbed image with weights $\mathbf{w}$.
- $\pi_x(x'_i)$: Kernel function (Proximity score) that gives higher weights to perturbed images close to the original image x.
- $\Omega(g)$: Regularization term to enforce sparsity in the surrogate model for interpretability.

4.3 Intersection Between LIME and SHAP

Both LIME and SHAP techniques aim is to provide insights into the model's decision-making process, but they do so from different perspectives and through different mechanisms. In image analysis, the intersection of LIME and SHAP entails integrating their advantages to produce a more reliable interpretation to a model's unique prediction. Consistent patterns of feature importance can be found by comparing the explanations offered by LIME and SHAP. This increases confidence in the interpretation of the model. For example, when the same superpixels or regions in an image are frequently highlighted by both LIME and SHAP as being crucial for a prediction, it strengthens the confidence in the significance of those regions.

5 Results and Discussion

As shown in Fig. 3, the application of the SHAP library to the COVID-19 CT scan data reveals the specific regions and features that exert the greatest influence on the model's prediction. SHAP quantifies the contribution of each region of our image data to the model's prediction by assigning Shapley values. These values indicate how much each region adds to or detracts from the classification decision. In essence, SHAP breaks down the CNN's prediction into parts, showing which regions have the most significant positive or negative impact on the final output. This detailed breakdown allows for a "region-level" understanding of the model's inner workings. By highlighting the most influential areas, SHAP helps us see exactly what the CNN is focusing on. The highlighted regions in

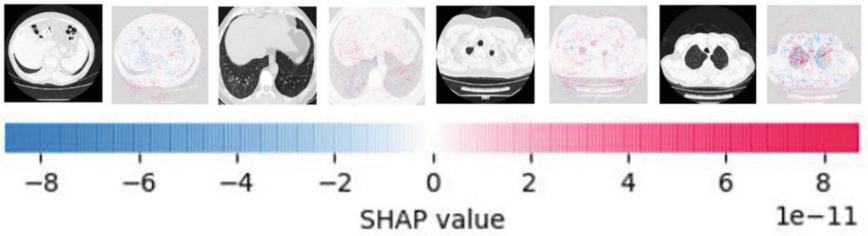


Fig. 3. CT scan images with SHAP explanations for COVID-19 classification, highlighting blue regions as less important and red regions as important for increasing the likelihood of a positive prediction.

red represent the areas that have the most significant influence on the CNN's predictions, indicating that these regions are critical for the model's decision-making process. In contrast, the blue-highlighted areas correspond to regions that have little to no impact on the model's prediction, suggesting that they are not relevant for the classification outcome.

Figure 5 demonstrates the application of LIME (Local Interpretable Modelagnostic Explanations) on our COVID-19 CT scan images. The yellow contours overlaid on the CT images highlight the regions that are of critical importance to the model's prediction. Specifically, LIME enables identify localized areas within the images that contributed the most to the model's decision.

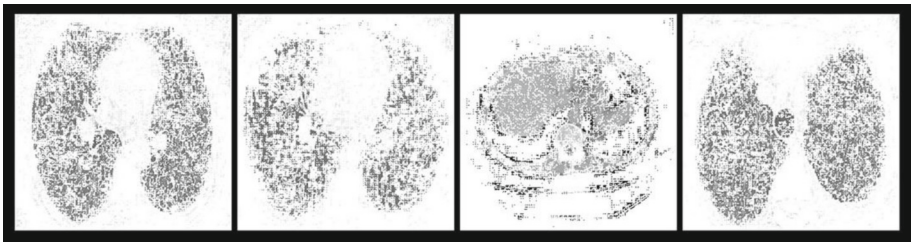


Fig. 4. CT scan slices after applying SHAP, highlighting the regions most critical to the model's prediction.

After thoroughly identifying the most important features using SHAP along with LIME, we carefully modified the original CT scan images by masking out the remarkably non important regions. The areas carefully masked in black, per Fig. 4, clearly show the important regions SHAP identified. Afterward, the CNN model was retrained with the generated images that only had the most important features. As presented in Table 1, we applied SHAP and LIME, to the entire dataset of both positive and negative COVID-19 images. The generated images were then divided into training and testing datasets. After training our model on these processed images, the SHAP identified important regions yielded an accuracy of 96.13%, while the LIME identified regions achieved an accuracy of 95.47%. The intersection of both SHAP and LIME-identified regions recorded an accuracy of 89.78%. These results underscore the model's reliance on the key regions

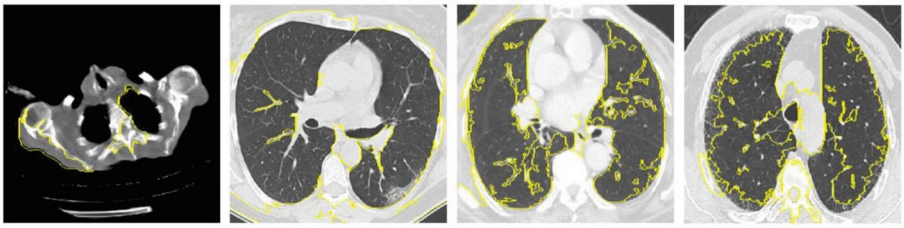


Fig. 5. CT scan images displaying LIME-generated explanations for COVID-19 classification, where yellow contours highlight key regions identified as significant contributors to the model's prediction.

for decision-making, as removing non-influential features did not significantly impact its performance. This further validates the effectiveness of XAI-based feature selection in enhancing model interpretability without compromising predictive accuracy.

5.1 Comparison Results

We compared our results with other models for the detection of COVID-19 from CT scan images such as Deep COVID DeteCT [18], VGG16 [19], VGG19 [19] and and Deep-COVID [16]. These Deep COVID DeteCT [18] and Deep-COVID [16] models utilized famous InceptionV3 [18] and NASNetLarge [16] pre-trained weights techniques and demonstrated substantial testing accuracy when trained on the dataset, mod dev dataset. However, applying these interpretability techniques, SHAP and LIME showed a decline in their accuracies. From Table 1, we can see that our COVID-CNN model [14] used for the prediction of COVID-19 CT scans performed better compared with other existing models when SHAP, LIME, and the intersection between both techniques were applied. This demonstrates the effectiveness of interpretability in highlighting important features, thereby optimizing model accuracy and reliability.

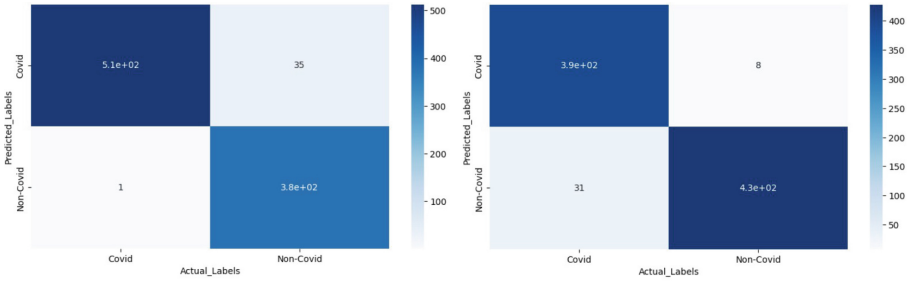
Table 1. Comparison results with existing models

Models	SHAP		LIME		Intersection		All Regions	
	Accuracy %	F1-Score	Accuracy %	F1-Score	Accuracy %	F1-Score	Accuracy %	F1-Score
COVID-CNN (Ours)	96.13	96.13	95.47	95.47	89.18	88.72	97.85	97.85
VGG19 [19]	90.37	90.31	93.83	93.82	84.24	84.21	88.55	87.42
VGG16 [19]	92.26	91.80	89.94	89.40	76.43	76.10	87.12	86.43
Deep COVID DeteCT [18]	86.23	69.99	69.42	71.49	70.39	70.21	70.89	70.08
Deep-COVID [16]	74.62	74.40	78.25	77.60	72.31	71.00	95.59	84.67

5.2 Evaluation Metrics

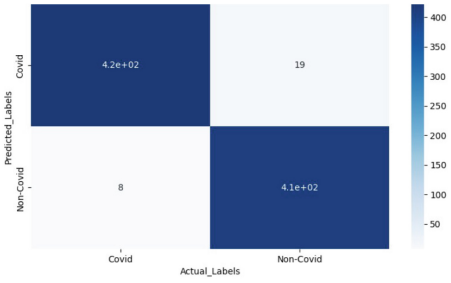
To evaluate the effectiveness of our model, Accuracy, Recall, Precision and the F1-score are used [20]. Accuracy represents the proportion of correct predictions out of the total predictions made. The F1-score, calculated as the harmonic mean of precision and recall, ensures a balance between precision which is the correctness of positive predictions and recall applies to the model's ability to identify all positive instances. These metrics offer a comprehensive evaluation of the model's performance, as illustrated in Eq. 4 [21].

$$\begin{aligned} \text{Accuracy} &= \frac{TP + TN}{TP + TN + FP + FN} \\ \text{Precision} &= \frac{TP}{TP + FP} \\ \text{Recall} &= \frac{TP}{TP + FN} \\ \text{F1 - score} &= \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \end{aligned} \tag{4}$$



(a) Confusion Matrix for SHAP Generated 4649 Images

(b) Confusion Matrix for LIME Generated 4649 Images



(c) Confusion Matrix for Intersection between SHAP and LIME on 4649 Generated Images

Fig. 6. Confusion Matrix for SHAP, LIME, and Intersection Using our CNN Model

6 Conclusion

In this paper, we applied explainable AI techniques SHAP and LIME to our developed COVID-CNN model [14] specifically developed to predict COVID-19 from CT scan images. After applying SHAP and LIME to our adopted model, the results provide valuable insights into the model's decision-making process. SHAP captured feature importance, demonstrating how the model focused heavily on specific regions where COVID-19 abnormalities typically occur. LIME, on the other hand, offered local interpretability by highlighting features critical to the predictions. The intersection features between SHAP and LIME confirmed consistent feature attributions, demonstrating features that are consistent to both interpretability techniques. Using all images of important regions identified by both SHAP and LIME, we retrained our adopted model using only the important regions masking out the non-important regions. Retraining on SHAP identified regions resulted in a 96.13% accuracy, while LIME based regions yielded 95.47% accuracy. This demonstrated how focusing only on relevant features reduces overfitting and model complexity. Intersection between both SHAP and LIME achieved an accuracy of 89.18%. The model retrained on SHAP identified regions achieved a prediction accuracy of 49.14% when tested on the original images, while the model retrained on LIME based regions attained an accuracy of 50.54%. This depends on setting a threshold $\in$ to determine which SHAP values are considered significant. We set $\in$ equal to or greater than 1e-5 for both SHAP and LIME to filter out insignificant feature contributions.

Although SHAP and LIME identify important features from the input data for the model prediction, the explainable ability of those techniques also depend on model prediction accuracy. If model performance is lower on given dataset, even if we use XAI techniques to identify important features from the input data, that might not be useful and accurate. All implementations developed in this paper can be found at: https://github.com/kobinasam/Explainable-Convolutional-Neural-Network-for-COVID-19-Images (Fig. 6).

Acknowledgment. This work is supported by the U.S. National Science Foundation under award 2434487 and U.S. National Institutes of Health U24 HG013013. We thank anonymous reviewers for their insightful comments and inputs.

References

1. Ng, M.-Y., et al.: Imaging profile of the covid-19 infection: radiologic findings and literature review. Radiol. Cardiothorac. Imaging **2**(1), e200034 (2020)
2. Wang, L., Lin, Z.Q., Wong, A.: Covid-net: a tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images. Sci. Rep. **10**(1), 19549 (2020)
3. Ribeiro, M.T., Singh, S., Guestrin, C.: Why should i trust you?: explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, pp. 1135–1144, 2016
4. Lundberg, S.M., Lee, S.-I.: A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems, NIPS, pp. 4765–4774, 2017
5. Bhandari, M., Shahi, T.B., Siku, B., Neupane, A.: Explanatory classification of cxr images into covid-19, pneumonia and tuberculosis using deep learning and xai. Comput. Biol. Med. **150**, 106156 (2022)

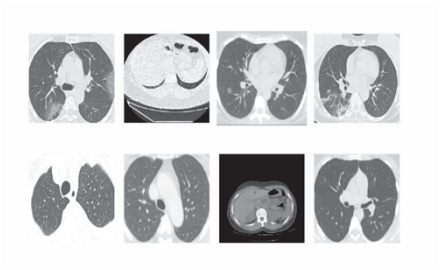
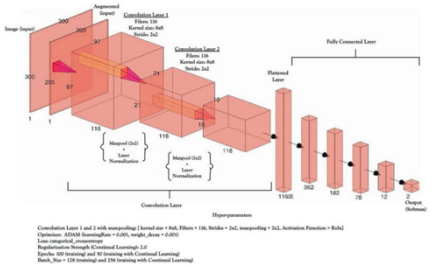
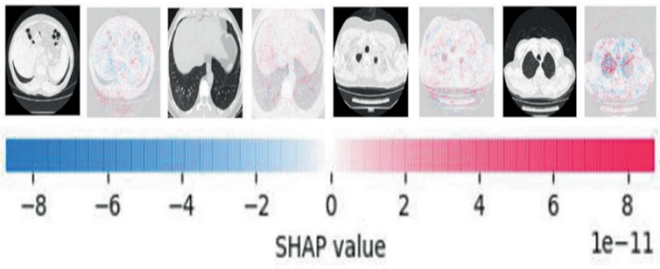
6. Ong, J.H., Goh, K.M., Lim, L.L.: Comparative analysis of explainable artificial intelligence for covid-19 diagnosis on cxr image. In: 2021 IEEE International Conference on Signal and Image Processing Applications (ICSIPA), pp. 185–190, 2021
7. Ahsan, M.M., Gupta, K.D., Islam, M.M., Sen, S., Rahman, M.L., Shakhawat Hossain, M.: Covid-19 symptoms detection based on nasnetmobile with explainable ai using various imaging modalities. Mach. Learn. Knowl. Extr. **2**(4), 490–504 (2020)
8. Ahsan, M.M., Nazim, R., Siddique, Z., Huebner, P.: Detection of covid-19 patients from ct scan and chest x-ray data using modified mobilenetv2 and lime. Healthcare **9**(9) (2021)
9. Sarp, S., et al.: An xai approach for covid-19 detection using transfer learning with x-ray images, Heliyon (2023)
10. Togacar, M., Muzoglu, N., Ergen, B., Yarman, B.S.B., Halefoglu, A.M.: Detection of covid-19 findings by the local interpretable model-agnostic explanations method of types-based activations extracted from cnns. Biomed. Signal Process. Control **71**, 103128 (2022)
11. Abeyagunasekera, S.H.P., Perera, Y., Chamara, K., Kaushalya, U., Sumathipala, P., Senaweera, O.: Lisa: enhance the explainability of medical images unifying current xai techniques. In: 2022 IEEE 7th International conference for Convergence in Technology (I2CT), pp. 1–9, 2022
12. Gaur, L., Bhandari, M., Razdan, T., Mallik, S., Zhao, Z.: Explanation-driven deep learning model for prediction of brain tumour status using mri image data. Front. Genet. **13**, 822666 (2022)
13. Zhang, Z., Damiani, E., Hamadi, H.A., Yeun, C.Y., Taher, F.: Explainable artificial intelligence to detect image spam using convolutional neural network. In: 2022 International Conference on Cyber Resilience (ICCR), pp. 1–5, 2022
14. Annan, R., Qin, H., Qingge, L.: Generalized deep learning models for COVID-19 detection with transfer and continual learning. In: Proceedings of the 16th International Conference on, vol. 101, pp. 58–72, 2024
15. Soares, E., Angelov, P., Biaso, S., Froes, M., Abe, D.: Sars-cov-2 ct-scan dataset: a large dataset of real patients CT scans for sars-cov-2 identification, 2020
16. Ghaderzadeh, M., Asadi, F., Jafari, R., Bashash, D., Abolghasemi, H., Aria, M.: Deep convolutional neural network–based computer-aided detection system for COVID-19 using multiple lung scans: design and implementation study, vol. 23, p. e27468, 2021
17. Maftouni, M., Law, A.C.C., Shen, B., Kong, Z.J., Zhou, Y., Yazdi, N.A.: A robust ensemble-deep learning model for covid-19 diagnosis based on an integrated ct scan images database. In: IIE Annual Conference. Proceedings, Institute of Industrial and Systems Engineers (IISE), pp. 632–637, 2021
18. Lee, E., et al.: Deep covid detect: an international experience on covid-19 lung detection and prognosis using chest CT. NPJ Digit. Med. **4**, 11 (2021)
19. Karim, M.R., Dohmen, T., Cochez, M., Beyan, O., Rebholz-Schuhmann, D., Decker, S.: Deepcovidexplainer: explainable COVID-19 diagnosis from chest x-ray images. In: 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 1034–1037, 2020
20. Yang, P., Sturtz, J., Qingge, L.: Progress in blind image quality assessment: a brief review. Mathematics **11**(12) (2023)
21. Miao, J., Zhu, W.: Precision–recall curve (PRC) classification trees. Evol. Intel. **15**(3), 1545–1569 (2022)

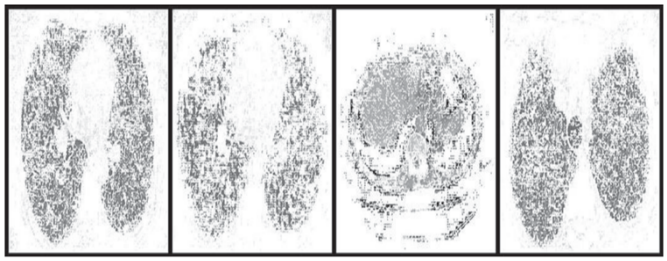
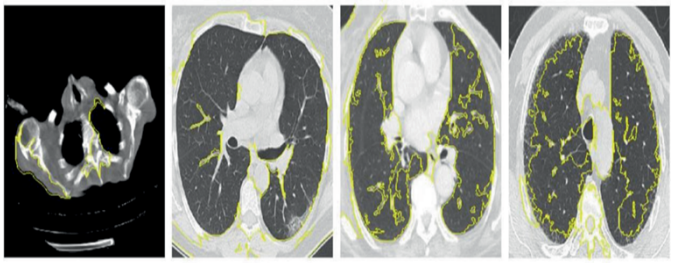
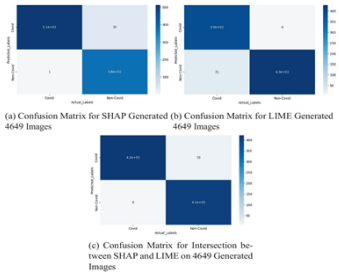
Author Queries

Chapter 17

Query Refs.	Details Required	Author's response
AQ1	Please check and confirm if the authors Given and Family names have been correctly identified.	
AQ2	Please check and confirm if the inserted citations of Figs. 1 and 6 are correct. If not, please suggest an alternate citations.	

Alternative Texts for Your Images, Please Check and Correct them if Required

Page no	Fig/Photo	Thumbnail	Alt-text Description
5	Fig1		A series of eight medical CT scan images showing cross-sectional views of the chest. The scans display various lung and thoracic structures, with differences in tissue density and contrast. Some images highlight lung abnormalities, while others show normal lung tissue. The scans are arranged in two rows, each containing four images.
6	Fig2		Diagram of a convolutional neural network architecture. It starts with an augmented input image, followed by two convolutional layers, each with 116 filters, kernel size 8x8, and stride 2x2. Maxpooling and layer normalization are applied after each convolutional layer. The network progresses to a flattened layer and a fully connected layer, ending with a softmax output. Hyperparameters include ADAM optimizer, learning rate 0.001, weight decay 0.001, and specific settings for loss, regularization strength, epochs, and batch size.
9	Fig3		CT scan images of a human chest are displayed in a sequence, each showing different cross-sectional views. Below the images is a color gradient bar representing SHAP values, ranging from -8 to 8, with blue indicating negative values and pink indicating positive values. The SHAP value is labeled, with a notation of 1e-11 on the right. The images and bar illustrate the impact of features on a model's output.

Page no	Fig/Photo	Thumbnail	Alt-text Description
9	Fig4		A series of four grayscale medical imaging scans, likely CT or MRI, showing cross-sectional views of a human torso. Each image displays different sections, highlighting variations in tissue density. The scans are arranged in a row, with subtle differences in shading and texture across the images, indicating anatomical structures.
10	Fig5		CT scan images of the chest showing cross-sectional views of the lungs and surrounding structures. Each image highlights areas with yellow outlines, indicating specific regions of interest. The scans display variations in tissue density and structure, useful for medical analysis.
11	Fig6	 (a) Confusion Matrix for SHAP Generated 4649 Images (b) Confusion Matrix for LIME Generated 4649 Images (c) Confusion Matrix for Intersection between SHAP and LIME on 4649 Generated Images	Three-panel figure showing confusion matrices for COVID-19 image classification. Panel (a) displays the confusion matrix for SHAP-generated images, with values indicating true positives, false positives, true negatives, and false negatives. Panel (b) shows the confusion matrix for LIME-generated images, with similar metrics. Panel (c) presents the confusion matrix for the intersection between SHAP and LIME on the same dataset. Each matrix includes axes labeled "Actual Labels" and "Predicted Labels," with categories "Covid" and "Non-Covid." Color intensity represents the frequency of predictions.