This article has been accepted for publication in a future proceedings of this conference, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI10.1109/I2MTC60896.2024.10560613, 2024 IEEE International Instrumentation and Measurement Technology Conference (I2MTC)

Explaining Deep Learning Models for COVID-19 Detection with Grad-CAM and Novel Use of PCA

Richard Yang
College of Engineering, Design and Physical Sciences
Brunel University London
London, United Kingdom
richard.yang@brunel.ac.uk

Qingping Yang
College of Engineering, Design and Physical Sciences
Brunel University London
London, United Kingdom
qingping.yang@brunel.ac.uk
(Corresponding author)

Ding Chen
Wuhan Union Hospital affiliated with Tongji Medical College
Huazhong University of Science and Technology
Wuhan, P. R. China
D201881391@hust.edu.cn

Fang Wang
College of Engineering, Design and Physical Sciences
Brunel University London
London, United Kingdom
fang.wang@brunel.ac.uk

Yang Qiu
Wuhan Union Hospital affiliated with Tongji Medical College
Huazhong University of Science and Technology
Wuhan, P. R. China
675856494@qq.com

***Abstract*—**Machine learning and more specifically deep learning has achieved remarkable results in a range of computer vision tasks such as classification. Despite this, their black-box nature means researchers are largely unable to explain and interpret the decisions these systems make. Researchers use various techniques to explain deep learning classification models, e.g. Class Activation Maps (CAM) and Gradient Weighted Class Activation Maps (Grad-CAM) which produce heat maps of the input image highlighting the regions that contribute most to the model's decision. In this paper we present a novel technique based on Principal Component Analysis (PCA) to explain deep learning model decisions at a higher level, with results similar to those produced by Grad-CAM. This technique is applied directly to our dataset of COVID-19 blood test images, and we compare the PCA results with Grad-CAM using the convolutional neural network model we developed using the same dataset. As the PCA is applied to the dataset directly, no deep learning model needs to be trained allowing for faster and simpler computation than techniques such as Grad-CAM while producing similar explanation results. The results indicated that the reconstructed PCA map using the first two principal components and Grad-CAM have a similarity score of 85.7% and 71.4% respectively for COVID-19 positive and negative images, with an average similarity score of 78.6%.

Keywords*—*explainablility, principal component analysis, deep learning, COVID-19, Grad-CAM

I. Introduction

Machine learning and deep learning can provide exceptional results, enabling many breakthroughs in a host of computer vision tasks. One such field is COVID-19 detection and diagnostics, where recent studies have shown that deep learning can accurately predict COVID-19 in patients. Many of these papers use image processing based on CT and X-ray images [1]. We have previously developed novel deep learning models for COVID-19 detection using flow cytometry images from complete blood count (CBC) tests [2]. However, their black-box nature and the lack of decomposability into intuitive and understandable components make them difficult to interpret [3]. Consequently, when such kinds of AI systems fail, they do not give the user any warning or explanation of the incoherent output. The interpretability and explainability of AI models are particularly important for their applications in medicine and healthcare.

To build trust in AI systems and have them more integrated into our lives we need to develop models which are 'transparent' and can explain why they predict what they predict [4]. Transparency is useful in three different stages in the evolution of artificial intelligence. When AI performance is weaker than humans and is not considered deployable, the transparency and explanations can pinpoint potential failure modes [5], thus enabling researchers to identify more promising research directions. When AI performance is on par with humans and is considered deployable, the transparency and explanations can help establish trust and confidence in the users. When AI performance is significantly stronger than humans, transparency and explainability can learn from the system and teach humans how to make better decisions [6].

There is generally a trade-off between model accuracy and simplicity or interpretability. Classical rule-based or expert systems prioritize interpretability over accuracy and robustness [7]. Decomposable pipelines are considered more interpretable since each stage is hand-designed and each component offers an intuitive explanation. However, when using deep models, we tend to trade interpretable modules for uninterpretable ones aiming at higher performance through increased abstraction with more layers and tighter integration with end-to-end training.

Currently there are various methods used by researchers for explainability of deep learning classification models, two popular methods are Class Activation Maps and Gradient Weighted Class Activation Maps (Grad-CAM). These techniques produce heat maps that highlight the regions of the input image that have the greatest influence on the model's decisions. In this paper we first apply Grad-CAM to the deep learning model we developed for COVID-19 detection and then compare it with a novel explainability method we developed using Principal Component Analysis (PCA). The application of the PCA method to our dataset images has produced similar maps to those produced by Grad-CAM, thus allowing it to be used for explainability without having to train a deep learning model as well as assist with classification. Our empirical work has established the use of PCA as a reliable alternative method for the general explainability of machine learning (ML) and AI models. This method is simpler and faster to compute than Grad-CAM, and does not depend on

This article has been accepted for publication in a future proceedings of this conference, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.1109/I2MTC60896.2024.10560613, 2024 IEEE International Instrumentation and Measurement Technology Conference (I2MTC)

the specific models, thus suitable for wider ML/AI application scenarios.

II. Deep Learning Model

The deep learning convolutional neural network (CNN) model we previously developed is briefly described in this section to allow its explainability to be studied.

A. Dataset

A dataset was obtained from Wuhan Union Hospital in China. The data is based on a blood test known as Full or Complete Blood Count (CBC) from patients admitted to the hospital during the initial outbreak of the pandemic. This blood test data is of 1744 (799 COVID positive and 945 control) patients' Side Fluorescence Light - Side Scattered (SFL-SSC) images (Fig. 1 and Fig. 2) generated from the Mindray series of Haematology Analysers used for CBC test. The images were used to train convolutional neural networks and also to apply our novel explainability technique based on PCA.

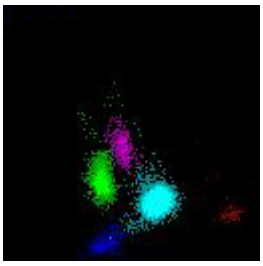

Fig. 1. Sample SFL-SSC Image (2D scattergram) from the CBC Flow Cytometry Dataset.

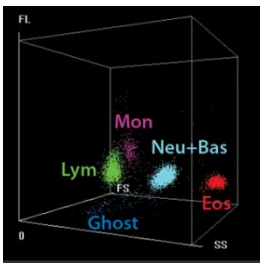

Fig. 2. Labelled Regions of 3D Scattergrams from the CBC Flow Cytometry Imaging.

B. Model Design

For our experiments, a convolutional neural network was trained to classify the images from the SFL-SSC image dataset [2]. The architecture of the CNN model is shown in Fig. 3. This model was trained 10 times and the performance metrics are then calculated for sensitivity, specificity, and accuracy.

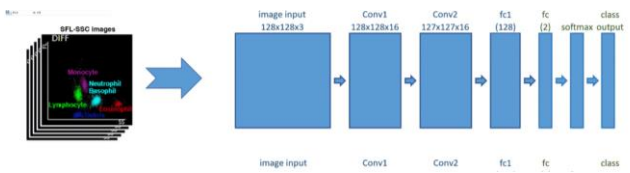

Fig. 3. The Architecture of the CNN Model for COVID-19 Detection Using SFL-SSC Images.

C. Model Results

The CNN produced very strong results, achieving very high sensitivity and specificity, 98.3% and 97.5%, respectively as well as high accuracies 97.8%, based on the test data set (15% of the total dataset), as shown in Fig. 4. Table 1 shows the ten repetitions of the CNN model results.

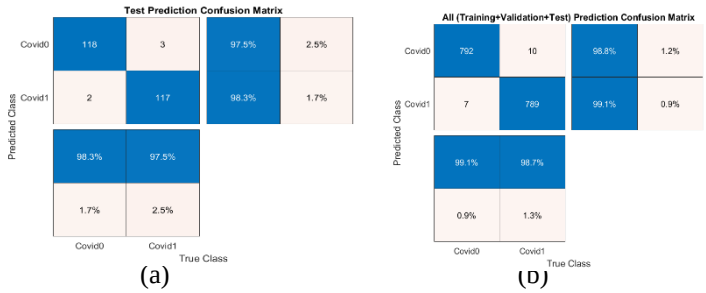

Fig. 4. Prediction Performance of the CNN Models Using SFL-SSC Images. (a) Prediction Confusion Matrix Based on Test Images; (b) Prediction Confusion Matrix Based on All (training+validation +test) Images.

TABLE I. CNN Results (10 repetitions)

	Validation			Training			Test			All		
No.	Sensitivity	Specificity	accuracy	Sensitivity	Specificity	accuracy	Sensitivity	Specificity	accuracy	Sensitivity	Specificity	accuracy
1	96.1%	96.4%	96.3%	97.0%	99.0%	98.3%	92.2%	95.3%	94.1%	96.2%	98.1%	97.3%
2	98.0%	97.6%	97.8%	98.3%	99.5%	99.1%	94.1%	94.1%	94.1%	97.6%	98.4%	98.1%
3	90.2%	96.4%	94.1%	96.6%	98.0%	97.5%	92.2%	98.8%	96.3%	95.0%	97.9%	96.8%
4	90.2%	92.9%	91.9%	97.5%	98.2%	97.9%	94.1%	92.9%	93.4%	95.9%	96.6%	96.3%
5	100.0%	92.9%	95.6%	99.6%	99.7%	99.7%	98.0%	97.6%	97.8%	99.4%	98.4%	98.8%
6	92.2%	98.8%	96.3%	98.7%	99.2%	99.1%	88.2%	97.6%	94.1%	96.2%	98.9%	97.9%
7	96.1%	94.0%	94.8%	97.9%	98.7%	98.4%	98.0%	96.5%	97.1%	97.6%	97.7%	97.7%
8	98.0%	97.6%	97.8%	99.2%	99.7%	99.5%	92.2%	97.6%	95.6%	97.9%	99.1%	98.7%
9	96.1%	91.7%	93.3%	96.6%	98.0%	97.5%	96.1%	96.5%	96.3%	96.5%	96.8%	96.7%
10	94.1%	95.2%	94.8%	97.9%	99.2%	98.7%	94.1%	98.8%	97.1%	96.8%	98.6%	97.9%
Average	**95.1%**	**95.4%**	**95.3%**	**97.9%**	**98.9%**	**98.6%**	**93.9%**	**96.6%**	**95.6%**	**96.9%**	**98.1%**	**97.6%**

III. Explainability using Grad-CAM

A. Class-Activation-Maps(CAM)

Zhou introduced a technique called Class Activation Mapping (CAM) which emerged as a method to identify discriminative regions in image classification using CNN [8]. To apply the CAM to a typical CNN deep learning model, the fully-connected layers before the final output are removed and replaced with a global average pooling (GAP) layer, followed by a fully-connected Softmax layer. The GAP is applied to the last convolutional feature maps, and the resulting pooled values are utilized as features for a subsequent fully connected layer responsible for generating the desired output. Based on this straightforward connectivity structure, it then becomes possible to discern the significance of different regions within an image. This is achieved by projecting the weights of the output layer back onto the convolutional feature maps. The GAP calculates the spatial average of the feature map for each channel at the last convolutional layer. These spatial averages are then combined using weighted sums to produce the final output. A similar weighted sum is employed to compute the feature maps of the last convolutional layer, resulting in the generation of class activation maps.

Below, the process is more formally outlined for the case of SoftMax, though the same approach can be adapted for regression and other loss functions.

For a given image, let $f_k(x, y)$ represent the activation of channel k in the last convolutional layer at spatial coordinates (x, y). For channel k, the result of global average pooling, F_k, is calculated as the summation of $f_k(x, y)$ over all spatial locations (x, y), $i.e.$

This article has been accepted for publication in a future proceedings of this conference, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.1109/I2MTC60896.2024.10560613, 2024 IEEE International Instrumentation and Measurement Technology Conference (I2MTC)

$$F_k = \sum_{x,y} f_k(x,y) \qquad (1)$$

Given a specific class c, the input to the SoftMax activation function for this class, S_c, is computed using F_k as follows:

$$S_c = \sum_k w_k^c F_k = \sum_k w_k^c \sum_{x,y} f_k(x,y) = \sum_{x,y}\sum_k w_k^c f_k(x,y) \quad (2)$$

where w_k^c signifies the importance (weighting) of the feature map f_k for class c.

Introducing the concept of M_c as the class activation map (CAM) for class c, where each spatial element is calculated as follows:

$$M_c(x,y) = \sum_k w_k^c f_k(x,y) \qquad (3)$$

we can simply express the class score S_c as:

$$S_c = \sum_{x,y} M_c(x,y) \qquad (4)$$

Thus $M_c(x,y)$ directly represents the significance or importance of the activation at spatial position (x,y) in relation to the classification of an image into class c. In essence, it explains which regions of the input image are important for the decision that predict the image as a specific class.

Each neuron in the network is designed to be activated by specific visual patterns within its receptive field. Therefore, the feature map f_k can be thought of as representing the activation (or presence) of these visual patterns. The CAM is essentially a weighted linear combination of these visual pattern activations at various spatial locations.

By performing a straightforward up-sampling of the CAM to match the size of the input image, we can locate the regions within the image that are most relevant to a specific class. This process allows us to locate and highlight the image regions that contribute the most to the model's decision for a specific class, providing valuable explanation how the model performs classification.

The main limitation of CAM is that it is only applicable to a specific class of CNN architectures that follow a particular sequence of operations, namely global average pooling over convolutional maps immediately before making predictions. This sequence typically involves convolutional feature maps being processed as follows: convolutional feature maps → global average pooling → SoftMax layer. For other more general CNNs, the architecture will need to be altered to have the above sequence. This constraint means that CAM may not be readily applicable to CNN architectures with different configurations or those that do not incorporate global average pooling in the specified manner. The altering of the architecture can generally degrade the classification performance.

B. Gradient Weighted Class Activation Mapping (Grad-CAM)

The CAM [8] was proposed to identify discriminative regions in the context of image classification using CNNs which do not contain fully connected layers. This means the model will trade off complexity and performance for increased transparency into the workings of the model.

In contrast, Ramprasaath [4] presents a novel approach, known as Grad-CAM, for combining feature maps utilizing the gradient signal, and unlike CAM, it does not need any modifications to the network architecture. This unique feature enables the approach to be applied to a wide range of CNN-based architectures, including those designed for tasks like image captioning and visual question answering. In the case of a fully convolutional architecture, the method effectively simplifies to the conventional CAM (Class Activation Mapping). Therefore, Grad-CAM can be viewed as a generalization of CAM, offering a more versatile and adaptable solution that can be used across a broader spectrum of CNN-based applications.

Deeper layers within a CNN are known to capture higher-level visual constructs [9]. Moreover, convolutional layers inherently preserve spatial information, a characteristic that is often lost in fully-connected layers. As a result, it can be anticipated that the last convolutional layers of a CNN strike the best balance between high-level semantics and detailed spatial information. Neurons in these layers are typically responsible for identifying semantic, class-specific information within an image, such as object parts or distinctive features.

The Grad-CAM leverages the gradient information that flows into the final convolutional layer of the CNN to discern the significance of each neuron for a particular decision or classification of interest [4]. By analysing this gradient information, Grad-CAM can explain which regions of the input image each neuron in the last convolutional layer finds most relevant for making a specific decision. This process helps highlight the key visual cues that guide the network's decision-making process.

To generate the Grad-CAM $M_{Grad-CAM}^c \in R^{u \times v}$, which represents a discriminative localization map, for any class c with a width u and height v, we initially calculate the gradient of the class score for class c, denoted as y^c (computed prior to the softmax operation), with respect to the feature maps A^k originating from a convolutional layer. This gradient is expressed as $\frac{\partial y^c}{\partial A^k}$ and calculated using backpropagation.

The gradients flowing back are subject to global-average-pooling to derive the neuron importance weights, denoted as α_k^c [4]:

$$\alpha_k^c = \frac{1}{N}\sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k} \qquad (5)$$

where N is the total number of pixels in the feature map; i and j index the pixels.

This weight α_k^c signifies a partial linearization of the deep neural network downstream from A and determines the significance or relevance of feature map k for a specific target class c. Subsequently, we execute a weighted combination of the forward activation maps, followed by the application of a Rectified Linear Unit (ReLU) activation function to achieve the following [4]:

$$M_{Grad-CAM}^c = ReLU(\sum_k \alpha_k^c A^K) \quad (6)$$

This will result in a heat-map which is the same size as the convolutional feature maps. ReLU activation function is applied to the linear combination of feature maps $\sum_k \alpha_k^c A^K$ since the focus is solely on the features that exert a positive influence on the class of interest. In other words, the aim is to identify pixels whose intensity should be increased to enhance

This article has been accepted for publication in a future proceedings of this conference, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.1109/I2MTC60896.2024.10560613, 2024 IEEE International Instrumentation and Measurement Technology Conference (I2MTC)

the class score y^c. Pixels with negative values are more likely associated with other categories or classes within the image. Without the ReLU, localization maps may inadvertently highlight not only the intended class but also other undesired elements, leading to diminished localization performance [4]. The inclusion of the ReLU activation function is essential in preventing this issue.

C. Grad-CAM Results

Due to its ease of implementation and good performance for localization and classification, we first applied Grad-CAM to the convolutional neural network we developed and trained for COVID-19 detection. This was applied using the SFL-SFC dataset images with the COVID positive and negative images averaged to produce an average image for both COVID positive and negative samples (Fig. 5).

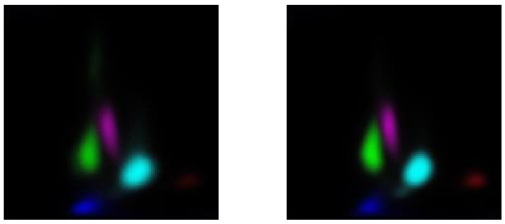

Fig. 5. Average SFL-SSC Images for COVID-19 Positive Patients (left) and COVID-19 Negative Patients (right).

The results of applying Grad-CAM to the average SFL-SSC images of COVID positive and negative patients are shown in Fig 6. Here we can see the regions of the image the network uses to make its decisions for both positive and negative classifications, and they are shown by the redder regions of the heatmap marking it as more important. If compared with Fig. 2, you can see the parts of the image corresponding to the specific blood cell in the patients' blood generated from the haematology analyses. From this we can see that patients who are classified as COVID positive have a higher concentration of Neutrophils and Basophils (Neu + Bas), which is higher if the body is under stress when fighting infections. On the other hand, COVID negative patients have a higher concentration of Eosinophils (Eos) and Monocytes (Mon) in their blood which are white blood cells that supports the immune system.

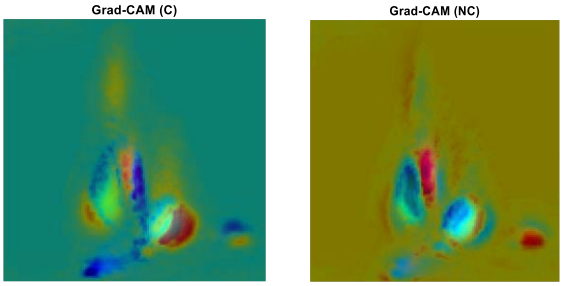

Fig. 6. Grad-CAM of Average SFL-SSC Images for COVID-19 Positive Patients (left) and COVID-19 Negative Patients (right).

D. Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a multivariate method employed for the analysis of a dataset where observations are characterized by multiple interrelated quantitative dependent variables. Its primary objective is to gain crucial insights from the dataset by transforming it into a collection of new orthogonal variables known as principal components to reduce the dimensionality of high-dimensional datasets while preserving as much relevant information as possible. These principal components serve to reveal the underlying structure of the data and facilitate the visualization of the relationships between observations and variables, often presented as points on maps [10].

PCA begins with the computation of the covariance matrix, typically represented as $\mathbf{\Sigma}$, for a dataset containing n observations and p variables. The covariance matrix $\mathbf{\Sigma}$ is calculated as follows:

$$\mathbf{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{x})^T \quad (7)$$

where $\boldsymbol{x_i}$ represents an observation vector and $\overline{\boldsymbol{x}}$ denotes the mean vector of observations. The principal components are obtained by solving the eigenvalue problem for $\mathbf{\Sigma}$:

$$\mathbf{\Sigma}\boldsymbol{v} = \lambda\boldsymbol{v} \quad (8)$$

where $\boldsymbol{v}$ represents the eigenvector and λ is the corresponding eigenvalue. The eigenvectors $\boldsymbol{v}$ are the principal components, and the eigenvalues λ indicate the amount of variance explained by each principal component. The eigenvalue problem can be solved in Singular value decomposition (SVD) or eigenvalue decomposition.

Typically PCA involves the following steps:

1) Centre the data by subtracting the mean of each variable (feature).
2) Calculating the covariance matrix $\mathbf{\Sigma}$.
3) Computing the eigenvalues and eigenvectors of $\mathbf{\Sigma}$.
4) Selecting the top k eigenvectors corresponding to the k largest eigenvalues to form the new feature space.
5) Transforming the original data into this new feature space.

The transformed data retains as much variance as possible while reducing the dimensionality, making it suitable for visualization or subsequent analysis, which is also the basis for its novel use for AI explainability.

E. Using PCA for Explainability

As mentioned previously each image has 128×128 pixels, allowing these images to be converted into a row vector with a length of 128×128×3. To apply PCA to every image in the dataset, we stack each image sample into two large matrices, for COVID positive images and for COVID negative images. The mean is removed from these matrices before PCA is performed. The PCA is applied to all the image vectors in the high dimensional space (each has 128×128×3 = 49152 pixels).

The principal components are essentially the new orthogonal axes with the 1st largest variances, 2nd largest variance, etc. The original values of each image as a vector are the coordinates seen in the original data axes. During the PCA, these coordinates are projected onto the new axes represented by the principal components with the new coordinates as scores. The original image can then be reconstructed using:

$$\boldsymbol{Xrc} = \boldsymbol{S} \cdot \boldsymbol{V}_k^T \quad (9)$$

This article has been accepted for publication in a future proceedings of this conference, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.1109/I2MTC60896.2024.10560613, 2024 IEEE International Instrumentation and Measurement Technology Conference (I2MTC)

where $\boldsymbol{S}$ represents the principal component scores as a matrix and $\boldsymbol{V}$ represents the principal component coefficients as matrix, with column containing coefficients for one principal component. By multiplying the $\boldsymbol{S}$ coordinates with the transposed $\boldsymbol{V}_k^T$ principal axes, the new coordinates in the principal component's axes are converted back to the coordinates in the original axes using the first k principal components approximation. Since each row of $\boldsymbol{S}$ is an image, we can reconstruct individual image using k principal components. These principal components can be displayed to show partially reconstructed images, meaning the mean has not been added back and represent or explain the most important features of the sample images.

F. Explainability Results Using PCA

PCA was applied directly to the combined COVID-19 positive and negative images of the SFL-SFC dataset. The reconstructed average image of both the COVID-19 positive and negative images using the first 10 principal components can be seen below in Fig. 7. Fig. 8 shows the partially reconstructed images for COVID-19 positive and negative using only the first 1, 3 and 10 principal components.

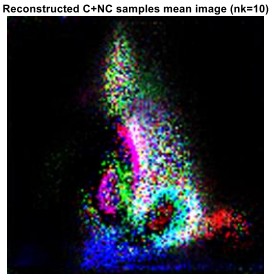

Fig. 7. Reconstructed Mean SFL-SSC Image of COVID-19 Positive/Negative Patients with First 10 Principal Components

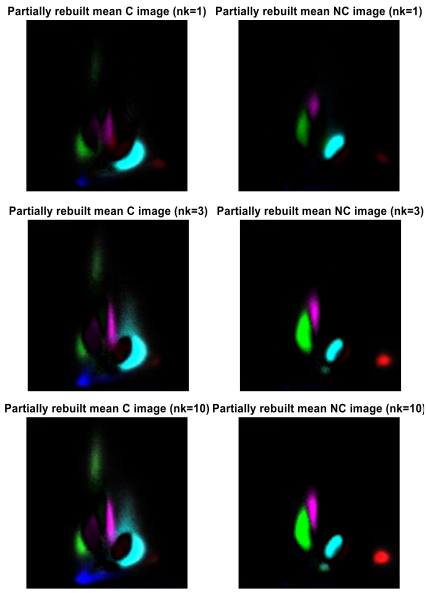

Fig. 8. Partially Reconstructed Images for COVID-19 Positive (left) and Negative (right) Using the First 1, 3 and 10 Principal Components.

It can be seen in Fig. 9 that first 1, 2, 3 and 10 principal components have explained 13.3%, 21.8%, 27.4% and 43.7% of the total variance, respectively. The first two principal components have explained the half of the variance explained by the first ten principal components.

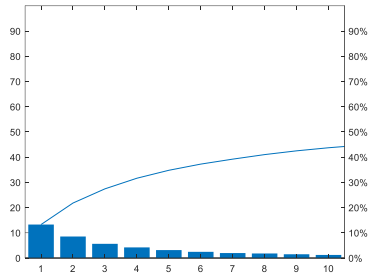

Fig. 9. Percentage of the Total Variance Explained by Each of the First 10 Principal Components.

IV. Discussions

From the above two sections, we can see that there are clear similarities between both the PCA and Grad-CAM results. For example, in the positive COVID images the same regions of the images are highlighted in the heatmap as well as the reconstructed PCA image (Fig. 10). This is also true for the COVID negative Grad-CAM and PCA images (Fig. 11).

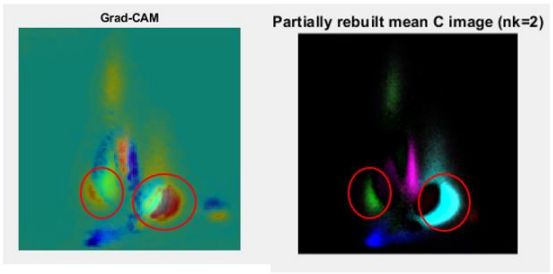

Fig. 10. Highlighted Similarities between COVID-19 Positive GRAD-CAM and Partially Reconstructed PCA Images.

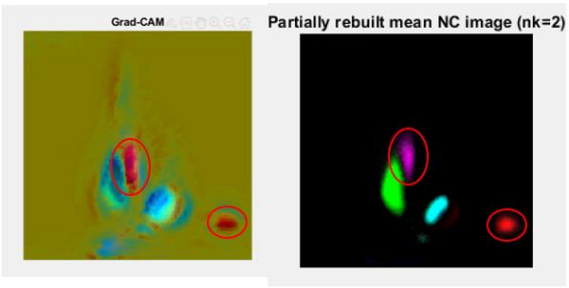

Fig. 11. Highlighted Similarities between COVID-19 Negative GRAD-CAM and Partially Reconstructed PCA Images.

Further close examinations of Fig. 10 and Fig. 11 indicate that both COVID positive and negative Grad-CAM images have 7 sub-regions with significant local maximal values and the PCA maps have located 6 and 5 of these sub-regions, thus giving a similarity score of 85.7% and 71.4% respectively for COVID-19 positive and negative images, and an average similarity score of 78.6%. It is worth noting that the average similarity score of 78.6% is achieved with only the first two principal components.

These similarities demonstrated that PCA can achieve similar results to those using Grad-CAM for explainability. While PCA is directly applied to the data and not the model itself therefore not directly explaining the model's decision

This article has been accepted for publication in a future proceedings of this conference, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.1109/I2MTC60896.2024.10560613, 2024 IEEE International Instrumentation and Measurement Technology Conference (I2MTC)

like Grad-CAM, it can be seen as a higher-level explanation of the model. In a way, its explanation can be regarded as the interpretation of the overall operation of the AI model and Grad-CAM.

As well as offering a higher-level explanation of the model, due to being directly applied to the data, PCA is computationally much faster and simpler to perform than training convolutional neural networks and applying Grad-CAM to the model. It can be used to gain important insights into the data even before the model is trained and it can also assist in classifications.

V. Conclusions

In this paper, we discussed a novel use of PCA to explain a convolutional neural network's classification in comparison with the well-known Grad-CAM method. The proposed PCA method was applied to a dataset comprising COVID-19 blood test results using CBC SFL-SSC images and tested with the CNN model we previously developed.

Based on the significant local maximal values in the Grad-CAM and the PCA reconstructed map using two principal components, the PCA map and Grad-CAM have a similarity score of 85.7% and 71.4% respectively for COVID-19 positive and negative images, with an average similarity score of 78.6%. These results have confirmed that the proposed PCA method can reliably generate heatmaps closely resembling those produced by Grad-CAM, which means that the proposed PCA method can be a reliable alternative for the general explainability of ML/AI models. Since the proposed PCA is simpler and faster to compute than Grad-CAM and it does not depend on the specific models, it is suitable for very wide application scenarios using various ML/AI models.

References

[1] R. Kumar, R. Arora,V. Bansal, V.J. Sahayasheela, H. Buckchash, J. Imran, et al., "Accurate prediction of COVID-19 using chest X-ray images through deep feature learning model with SMOTE and machine learning classifiers," *MedRxiv*, pp. 2020-04, 2020.

[2] R. Yang, D. Chen, Q. Yang, Y. Qiu, F. Wang, "Accurate COVID-19 detection using full blood count data and machine learning," in *Conf. International Congress on Measurement, Quality and Data Sceience*, Bordeaux, 2023.

[3] Z. C. Lipton. "The mythos of model interpretability," *Queue*, vol. 16, no. 3, pp. 31-57, Jun 2018.

[4] R.S. Ramprasaath, C. Michael, D. Abhishek,V. Ramakrishna, P. Devi, and B. Dhruv, "Grad-CAM: visual explanations from deep networks via gradient-based localization," in Proceedings of *IEEE International Conference on Computer Vision*, Venice, Italy, 2017, pp. 618-626.

[5] D. Hoiem, Y. Chodpathumwan and Q. Dai, "Diagnosing error in object detectors," in *European Conference on Computer Vision*, Florence, Italy, Oct. 2012, pp. 340-353.

[6] E. Johns, O. Aodha, amd J.B. Gabriel, "Becoming the expert-interactive multi-class machine teaching," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*," 2015, pp. 2616-2624.

[7] P. Jackson. *Introduction to Expert Systems*, United States, 1986.

[8] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*," Boston, MA, USA, 2016, pp. 2921-2929.

[9] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Transactions on Pattern Analysis and Machine Intelligence*,vol. 35, no. 8, pp.1798–1828, 2013.

[10] L. Smith, *Principal Components Analysis*, Department of Computer Science, University of Otago, New Zealand, 2002.