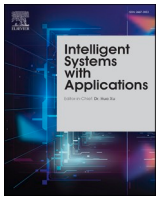
Intelligent Systems with Applications 18 (2023) 200233

Contents lists available at ScienceDirect

Intelligent Systems with Applications

journal homepage: www.journals.elsevier.com/intelligent-systems-with-applications



Using feature maps to unpack the CNN 'Black box' theory with two medical datasets of different modality

Sami Azam [a], Sidratul Montaha [b], Kayes Uddin Fahim [b], A.K.M. Rakibul Haque Rafid [b], Md. Saddam Hossain Mukta [c,*], Mirjam Jonkman [a]

[a] Faculty of Science and Technology, Charles Darwin University, Casuarina, NT 0909, Australia
[b] Health Informatics Research Laboratory (HIRL), Department of Computer Science and Engineering, Daffodil International University, Dhaka 1341, Bangladesh
[c] Department of Computer Science and Engineering, United International University (UIU), United City, Madani Avenue, Dhaka 1212

ARTICLE INFO

Keywords:
Black box
Convolutional neural network
Feature map analysis
Geometric feature
T-test
ANOVA test

ABSTRACT

Convolutional neural networks (CNNs) have been established for a comprehensive range of computer vision problems across several benchmarks. Visualization and analysis of feature maps generated by convolutional layers can be an effective approach to explore the hidden and complex characteristic of a CNN model. Convolutional layers provide diverse feature maps however, the extent of this diversity needs to be explored. This research attempts to provide five insights of the 'Black box' mechanism of CNNs, using skin cancer dermoscopy and lung scan computed tomography (CT) Scan datasets by statistically analyzing layer by layer (three convolutional layers) feature maps using 17 geometrical and 6 intensity-based features to determine the characteristics and level of diversity. Significance and difference of the feature maps layer by layer, black feature maps analysis, difference of the feature maps to each other and to the original image, variations among the feature maps when running the model multiple times and inter-class variation among the feature maps for different iteration are explored. Various statistical methods including T-test, analysis of variance (ANOVA), mean, median, mean squared error (MSE), peak signal to noise ratio (PSNR), structural similarity index (SSIM), root mean squared error (RMSE), dice similarity score (DSC), universal image quality index (UQI) and Spectral angle mapper (SAM) are employed. Experimental results show that for the skin cancer dermoscopy dataset, a large number of black feature maps are produced (20–60%) while the proportion of black feature maps for the CT Scan dataset is comparatively low (2–20%). This demonstrates that for different datasets, feature maps with diverse characteristics can be produced. The layer by layer differences between the feature maps is evaluated using T-tests and ANOVA for seventeen geometrical features and six intensity-based features. For both datasets across most of the geometrical features and across most of the intensity-based features a significant diversity can be observed. The difference of the feature maps to each other and to the original image is quite high, with MSE values for the dermoscopy and CT Scan datasets in the range of 1860–31,399 and 171–6089, respectively, PSNR 3–15 and 10–25, SSIM values of 0.01–0.84 and 0.3–0.81, RMSE values of 0.81–1 and 0.21–1, DSC values of 0.37–0.53 and 0.47–0.75, UQI values of 0.02–0.86 and 0.01–0.88 and SAM values of 0.12–1.53 and 0.19–1.55 for the dermoscopy and CT Scan datasets respectively. When running the model multiple times (three iterations), a notable iteration by iteration diversity is found in terms of mean, median, maximum and minimum values for most of the geometrical features. The inter-class variation among the feature maps for different iterations and layers are evaluated based on the F-value of the ANOVA test. For the dermoscopy dataset, the highest mean F-value is found for layer 1 and iteration 3 while for the CT scan dataset the highest mean F-value is found for layer 3 and iteration 3 indicating that for these feature maps the highest inter-class dissimilarity is generated. The findings of this study may aid in exploring the complex mechanism of convolutional layers, kernels and feature maps.

* Corresponding author.
E-mail address: saddam@cse.uiu.ac.bd (Md.S.H. Mukta).

https://doi.org/10.1016/j.iswa.2023.200233
Received 16 January 2023; Received in revised form 13 April 2023; Accepted 8 May 2023
Available online 10 May 2023

1. Introduction

Convolutional neural networks (CNNs) are used in different automated tasks, such as classification, detection, segmentation, data augmentation (Szegedy et al., 2016). However, CNNs are still considered as a 'Black box' in terms of the underlying mechanism which makes it difficult to interpret the results and have confidence that they provide the optimal solution (Wang et al., 2020) (Park & Yang, 2019) (Li et al., 2022). Though CNNs provides impressive outcomes, several questions still remain unanswered regarding the behavior of the model including how features are learned from the layers, what determines the diversity of the feature maps, what mechanism is responsible for good results and what the causes of failure are. The learned features are challenging to interpret with human vision, resulting in a lack of understanding of the network's core functioning mechanism (Qin et al., 2018). To get more insight, a visualization of CNN can be done using a quantitative/statistical analysis scheme which can provide some patterns of the features derived by CNN. Our interpretation primarily focuses on a layer by layer feature map analysis from different angles. The convolutional layers of the CNN produce distinct feature maps. However, to explain the 'black box' insight in these feature maps is required. The main objective of this research is to analyze the diversity of feature maps from different perspectives. A skin cancer dermoscopy and lung scan computed tomography (CT) Scan datasets are used to derive the resultant images from convolutional layers (feature maps) respectively. The objective of employing two datasets is to present a rigorous assessment of the analyses. While generating the feature maps black images ('Black FM') are also produced in each layer. As the number of these 'Black FM' is substantial, this is an aspect of CNNs which requires further analysis. In CNNs, different layers produce different feature maps and, the degree of layer by layer feature map diversity should therefore be considered. Moreover, the kernels applied by convolutional layers are not consistent while running the model multiple times as a consequence of random seeds (Zeiler & Fergus, 2014). Every time a network is created it uses different arbitrary starting patterns so each time the model trains it learns somewhat altered information owing to randomness. The neural network is initialized with random values of weights, which result in different starting points for every simulation throughout the training period (Rudd-Orthner & Mihaylova, 2019). As a result, for each iteration a different set of feature maps is created (Z. Zhang et al., 2018). Another objective is therefore to analysis the feature maps iteration by iteration. A major research question is why the performance varies for each time a model is trained. This is also explored in this study. The 'black box' nature of CNNs is a key research interest currently where different research attempted to provide different insight. There are several research questions related to the 'black box', such as the inner mechanism of CNN layers, feature interpretation and decision making schemes. Dabliain et al., (Dablain & Jacobson, 2021) presented seven research questions related to the CNN's training performance in terms of classifier retraining, class imbalance issues, relevant features, majority and minority classes and classifier weights. Heinrich et al., (Heinrich et al., 2019) addressed similar issues through two research questions, trying to determine specific schemes to interpret neural networks and incorporating the methods in a framework. Other studies (Zhao et al., 2022) (Brahimi et al., 2018) (Lange et al., 2018) attempted to explain the CNN 'black box' through a feature map visualization approach such as activation maps, saliency map, activation maps etc. However, the concepts such as feature map diversity, kernel operation, and decision making scheme of CNN are generally explained in a more theoretical way in these studies, while we attempt to present a broad experiment using a quantitative and statistical approach. The quantitative analysis includes different aspects of the diversity of feature maps: the extent of this diversity, the possible causes of diversity, kernel operations, inter-class differences based on the diversity of feature maps and more. Though the background and basic concepts of CNN are well-known, a deeper insight in these concepts based on rigorous experimentation can be beneficial in working with CNN architectures. CNN is widely-used due to its performance in computer vision. Behind this notable performance are the convolutional layers which produce the feature maps. While developing CNN architecture, learning the characteristics of the feature maps (layer to layer and iteration to iteration) can, result in a more robust architecture. In this way, the detection and classification performance of a model can be improved. This study uses two datasets having diverse modality and information to investigate how a CNN generates feature maps for different datasets. This may contribute to future studies to developing the optimal CNN model for different datasets. The inner mechanism of the decision-making process of CNNs is still poorly understood, not only by non-technical users but also by experts. This lack of knowledge may cause ambiguity and a hesitance in relying on the predictions of CNNs, especially in critical applications like the medical domain Lange et al. (2018). Opening up the 'black box' can increase the confidence of users such as medical specialists in the results of neural networks (Ferdinand & Mercier, n.d..) (Brahimi et al., 2018) (Dependent et al., 2021). To our knowledge, no study can be found showing the characteristics and level of diversity feature maps of convolutional layers. A new approach is introduced in this study to demystify the 'black box' of CNNs, based on the feature maps of different layers and iterations. As a new concept is raised in this study, in future studies of CNN or 'Black box', the researchers can be highly benefitted from the findings of this paper. Overall, this research aims to dive deeper into the 'Black box', providing insights by analyzing the different characteristics of feature maps. This includes five analyses:

- Analysis-1: 'Black FM' analysis – analyzing the number and proportion of black images in the feature maps of both datasets for all the three layers and iterations.
- Analysis-2: Layer by layer feature maps analysis for each iteration – analysis in terms of geometric and intensity differences of the feature maps for each layer and iteration, employing 17 geometric and 7 intensity-based features extracted from the feature maps after excluding 'Black FM'.
- Analysis-3: Feature maps analysis for a single input image – analysis in terms of the difference of the feature maps and their input image and the with other feature maps, using similarity measures such as mean squared error (MSE), peak signal to noise ratio (PSNR), structural similarity index (SSIM), root mean squared error (RMSE), dice similarity score (DSC), universal image quality index (UQI) and Spectral angle mapper (SAM).
- Analysis-4: Iteration by iteration feature maps analysis – investigating the change in feature maps during three iterations based on the mean, median, max and min values of the 17 geometrical features of the feature maps produced for both datasets.
- Analysis-5: Iteration by iteration feature maps analysis to find inter-class variance – evaluating inter-class difference of the feature maps of the two classes (skin cancer dataset) and three classes (CT Scan dataset) for three iterations using the 17 geometrical features.

In this research, two datasets are used where skin cancer dermoscopy dataset contains two classes of benign and malignant and lung scan CT Scan dataset contains three classes of COVID-19, normal and community-acquired pneumonia (CAP). The objective of considering two datasets is, as the modalities are different, the experiments can be assessed in a more profound way. In this regard, a CNN model is developed with three convolution layers and feature maps are extracted from each layer. As weight initialization uses a random seed, creating different feature maps for each iteration, the model is run three times and feature maps are extracted for each of these three iterations. In analysis-1, the number and proportion of 'Black FMs' for each layer and iteration is evaluated for both datasets. These 'Black FMs' are then removed and the remaining feature maps are used for the rest of the research. In analysis-2, Seventeen geometric features are extracted from the remaining feature maps and T-tests (Xu et al., 2017) (Chaves et al.,

2009) and analysis of variance (ANOVA) tests(Sankur, 2002) (Sampathila et al., 2022) are used to compare the feature maps for different layers and iterations. Moreover, 7-pixel intensity-based features, including pixel brightness, max pixel intensity, number of bright pixels, contrast level, noise level and energy are used to find intensity-based differences among the feature maps. In analysis-3, an original image and its resultant feature maps are then compared deriving the values of MSE, PSNR, SSIM, RMSE, DSC, UQI and SAM (Aslahishahri et al., 2021) (Jin & Chen, 2013) (Montaha et al., 2021). Dissimilarity among the feature maps of a single input image is also evaluated. In analysis-4, Mean, median, maximum and minimum values of the 17 geometrical features are investigated to show the change in feature maps during three iterations and in analysis-5, F-value of ANOVA test using geometrical features of different iterations are carried out to investigate inter-class differences. The methodology and results are briefly explained in the following sections. Section 2 describes prior studies which investigated different aspects of the 'Black box'. In Section 3, a description of both the skin cancer dermoscopy and lung scan CT scan datasets are given. Section 4 gives an overview of the methodology of this study and in Section 5 a description of the CNN architecture used for this research is given. Section 6 contains five sub-sections, describing the five analyses for both datasets, including results and findings. Section 7 provides a short discussion and suggestion for future studies regarding the 'Black box' and in Section 8 the conclusion of this research is given.

2. Literature review

To the best of our knowledge relatively little research has been conducted to unpack the 'Black box' aspect of CNNs, analyzing feature maps. A neuroscience experiment was conducted by Wang et al., (Wang et al., 2020) to understand the layer mechanism of a CNN model using the concept of a Pac-Man video game played by monkeys. Based on their approach, the complex decision-making method of brain can be understood and correlated with CNNs by analyzing the game strategies. Using a video dataset of the game with a CNN model, the authors came to the conclusion that different layers are involved in extracting diverse yet vital features which lead to the ultimate overall assessment. However, no statistical evaluation or the extent of feature map diversity was not assessed in this study. Introducing the concept of matched filtering, Li et al. (Li et al., 2022) demystified the mechanism of CNN by detecting the existing features in the input. They presented a thorough investigation of the learning process of CNN layers and parameters. Several experiments were carried out to explain the inner functionality of CNN. However, limitations include in exploring the impact of kernels on generating feature maps. Moreover, absence of quantitative assessment was observed. Zeiler et al. (Zeiler & Fergus, 2014) presented a visualization method to provide an intuitive insight into the function of CNN layers and the inner procedure of the classifier. An ablation study was carried out to ascertain influence of different layers on the performance. Through ablation study, several layer structures were investigated with which provided the insight of impact of layers on final decision. Though the mechanism of CNN layers and classifier was explained in this study, fewer experiments are found in demystifying the characteristic of kernel. In a recent study (Prijs et al., 2022), the authors attempted to get more insight in the 'Black box' of CNNs by segmenting the fracture lines of ankles using radiographs. After segmenting the fracture lines, a classification approach was carried out to evaluate the performance of CNN more rigorously. With AUC values ranging from 93% to 99% they concluded that CNN has a good discriminatory competence in terms of identifying as well as categorizing the fractures of the ankle. To unpack CNN 'black box' theory, the mechanism and characteristic of layers, kernels and feature maps were less investigated in this research. Moreover, absence of statistical assessment in order to present a comprehensive analysis is observed. Zhang et al. (Q. Zhang et al., 2019) introduced a quantitative analysis for the predictions produced by a CNN for semantic perception employing the inner decision-making mechanism of decision trees. Through analyzing the decision making process of decision trees, they tried to identify which object parts aid significantly in the prediction of a CNN. This research showed the significant region of activation map that contributes greatly in final decision. However, the core components and functionality of CNN such as convolutional layer, kernels and diversity of feature maps were not investigated in this research. Heinrich et al., (Heinrich et al., 2019) presented a broad analysis and constructed a classification system that can help to interpret the neural network. They focused on aspects of CNNs and recurrent neural networks (RNNs) such as Feature Visualization, Activation Maximization, Backward Propagation, Class Activation Map, and Dimensionality Reduction. These core mechanisms of CNN were explained broadly through several experiments. Though a number of experiments were conducted to analyze the characteristics of feature maps, limitations include in the analysis of CNN layers and kernels. A detailed survey is presented by Buhrmester et al., (Buhrmester et al., 2021), to explain the mechanism of neural networks for computer vision. They presented a comparison of studies investigating the 'black box', describing the limitations and providing recommendations for future research. Wei et al., (Wei et al., 2015) analyzed the intra-class variation in fully-connected layers of CNN through a visualization approach. They introduced visualization techniques including parametric visualization model, data-driven patch prior, local and content variation and ensemble and hierarchical encoding. Activation maps were shown for different types of images. It is discovered that along with capturing the local and style variation, CNN encodes these in a hierarchical and ensemble manner. However, no statistical analysis and experiments of CNN layers and kernels were conducted in this research. He et al., (He et al., 2019) presented a comprehensive study to describe CNN's inner mechanisms including learning scheme of CNN, reason of failure and way of improving performance. Three research questions were answered including what the models learn from input data, how and when the models fail to interpret and how to boost the performance of the models. The authors employed VGG16 pre-trained network as deep saliency model for the experiments. The intermediate layers of the model were focused to analysis the learning of individual neuron. Normalized scanpath score (NSS) and mean value of activation maps were used for the analysis of local saliency statistics. It is found that pre-trained models for computer vision tasks have some visual saliency encoded already. On the other hand, the fine-tuned pre-trained models generate irregular responses. However, in this study, more attention was given to the important region of the saliency maps where the diversity of the feature maps based on layer by layer was unexplored. Mahendran et al., (Mahendran & Vedaldi, 2016) analyzed deep CNNs through a number of techniques. Three approaches were introduced inversion where reconstruction of the images was presented, activation maximization and characterization. According to the findings of the study, several layers of CNNs perceive information of images having geometric and photometric invariance of diverse extent. Vaghjiani et al., (Vaghjiani et al., 2020) adopted VGG16 architecture to visualize the feature pattern of CNN. Feature maps with respect to different layers and kernels were interpreted to determine the most optimal features from fundus images in glaucoma detection. Moreover, a number of interpretable notions were developed to determine and investigate the optimal imaging features that contribute most in final prediction. The papers described above have analyzed several aspects of CNN and conducted experiments to understand and visualize the inner mechanism. In some papers, discussion of different layers and parameters were presented. However, limitations of these studies remain in the broader investigation of convolutional layers and their resultant feature maps. Convolutional layers and the learnt features by the layers are the core mechanisms of CNN which requires rigorous investigations to discover some hidden knowledge. The resultant feature maps are yielded by kernels. Very less study has worked with the impact of kernels in producing feature maps. No study is found to present a rigorous statistical analysis like ours in quantitative evaluation of CNN's properties.

Moreover, the analysis of black feature maps is a new finding presented by this study. In recent years, interpreting medical images through AI based approaches has become one of the promising solutions. The details of the medical imaging modalities are so complex, hidden and informative. As a number of computer aided system are proposed to diagnosis diseases using medical images, how CNN interprets these data should be explored. No significant and informative 'black box' study is found to have worked with medical domain. This study introduces two medical datasets of different modalities and characteristics to describe the performance of CNN broadly. Table 1 gives an overview the main objective and strategy presented in these papers. The reason of including the papers below in Table 1, is to give an overview of relevant prior research. While the objectives of our study and previous research may be similar, the strategy can be different.

As can be seen from Table 1, a number of researchers have been working to demystify 'black box' nature of CNNs, focusing on different aspects. However, to our knowledge, no study focused on the feature maps produced by convolutional layers to explain 'black box' mechanism of CNN. As the feature map are one of the most vital mechanisms of a CNN, we expect that diving deeper in this mechanism can create some useful insights in the CNN 'black box'.

Table 1
Overview of the literature.

Paper	Objective	Strategy
Wang et al., 2020	Understanding the layer mechanism of a CNN to explain how different layers are involved in extracting diverse yet vital features.	Using the concept of Pac-Man video game correlated with the decision-making method of brain.
Li et al., 2022	Demystifying the mechanism of CNN by a thorough investigation of the learning process.	Working with CNN layers and parameters.
Zeiler & Fergus, 2014	Providing an insight in the CNN layer functionality and the inner procedure of the classifier.	A visualization method and ablation study to ascertain influence of different layers on the performance.
Prijs et al., 2022	Getting deep insight of the 'Black box' of a CNN model.	Segmenting and classifying the fracture lines of ankles using radiograph images
Q. Zhang et al., 2019	Identifying the most significant features produced by CNNs in accurate prediction.	A quantitative analysis of the CNN model's prediction using a decision tree strategy.
Heinrich et al., 2019	Interpreting the mechanism of neural network.	Explaining several CNN and RNN mechanisms, such as Feature Visualization, Activation Maximization, Backward Propagation, Class Activation Map, and Dimensionality Reduction.
Buhrmester et al., 2021	A survey on the mechanism of neural networks for computer vision.	Discussion of studies related to the 'black box' aspect.
Wei et al	Discovery of intra-class representation in CNN fully connected layers.	Several patch prior and encoding visualization techniques.
He et al	Visualization of saliency models to learn CNN inner mechanism and performance.	VGG16 pre-trained network as saliency model, statistical analysis of activation maps.
Mahendran et al	Understanding layer mechanism and visualization of deep CNN.	Image reconstruction, activation maximization, pattern analysis and layers' learning mechanism.
Vaghjiani et al	Visualization of inherent features produced by CNN.	VGG16 as model architecture, feature analysis of different layers and kernels.
This study	Interpreting the mechanism of a CNN.	Working with feature maps produced by convolutional layers and quantifying the extent of diversity through statistical analysis.

3. Datasets description

To conduct all the experiments, two datasets are employed. The skin cancer dermoscopy dataset (Skin Cancer: Malignant vs. Benign | Kaggle, n.d.) with 3297 images of the ISIC archive is considered for this study. The dataset is collected from the Kaggle repository and contains two classes, benign and malignant, where the benign class contains 1800 images and the malignant class has 1497 images. The pictures are in RGB format having 224 × 224 pixel size. However, not all images are utilized for this research. For both classes, benign and malignant we have randomly taken 200 images. In this regard, all the images are not used as the aim is to extract feature maps from the original image and if all the images are taken, the number of feature maps would be substantially high which would result in higher computational complexity of our analyses. Therefore, a moderate and equivalent number of 200 dermoscopy images are considered. Fig 1 illustrates the original images of two classes of skin cancer dataset.

It can be seen from Fig 1 that there are some artifacts (such as hairs) in the images. These are removed using morphological opening (Montaha et al., 2022).

The lung scan dataset contains 17,104 CT scans from the Kaggle repository(Large COVID-19 CT scan slice dataset | Kaggle., n.d.). There are three classes named Normal, Covid, and CAP where normal class has 6983 images, Covid contains 7593 images and 2618 images are found for class CAP. The CT scans are in PNG format and have pixel size of 512 × 512 which is further resized into 224 × 224. However, like skin cancer dataset, for all the classes of lung scans, we have randomly taken 200 images respectively. Fig 2 illustrates the original images of two classes of lung scan dataset.

The aim of using these two datasets is to present a more rigorous experiment of unpacking the 'Black box' aspect of CNNs, related to the feature maps of the convolutional layers. The two datasets are very different in imaging modality, region of interest (ROI), color format, size and diseases. The ROI of the skin cancer dermoscopy dataset is relatively large whereas the ROI of the lung scans is comparatively small and more complex. The feature maps of these diverse datasets, can provide an idea of how for convolutional layers extract different types of feature maps different modalities.

4. Methodology

In medical AI, the objective is to assist the clinicians in ensuring patient safety and improving healthcare quality through various computer-aided systems for diagnosis, progression and recommendations (Yanase & Triantaphyllou, 2019). Nevertheless, CNN based architectures generally function as black-box models, making it challenging to comprehend the actual relations in the data (Lin et al., n.d.). For medical-based AI systems to be utilized in routine clinical diagnosis, with the help of explainable AI, it is essential to understand how the interpretation is provided and what the CNN learns or extracts

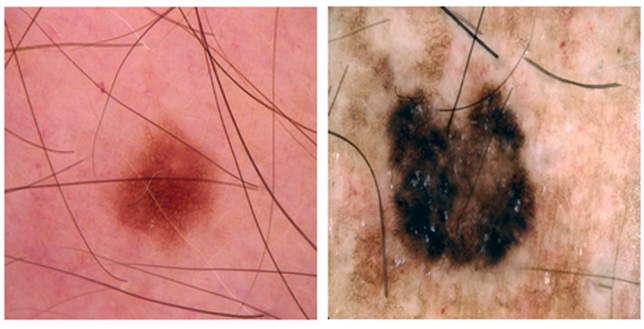


Fig 1. Original images of skin cancer dermoscopy dataset.

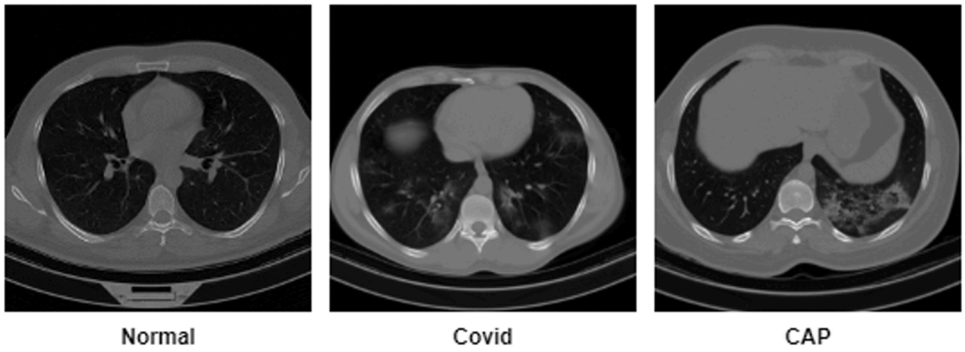


Fig 2. Original images of lung scan CT scan dataset.

from data to reach into a certain decision (Giuste et al., 2022). The rising questions concerning the performance of model interpretability have motivated the development of approaches that benefit users understanding CNN predictions. In this study, to analyze the diverse characteristics of the feature maps, first, a CNN model is constructed. This has three convolution layers each followed by one maxpool layer. As the objective of this study is to evaluate the diversity of feature maps generated by convolutional layers, the model is developed from scratch with three convolutional layers. The motive is to show how the feature maps of one convolutional layer is different from the following convolutional layer. Moreover, as mentioned the experiments are conducted across two different datasets, a shallow architecture having three convolutional layers are developed to present a rigorous experiment while keeping the contents precise and compact. Furthermore, while working with CNN, maxpool layer is added to address the computational complexity. In computer vision, a number of images are generally required to attain an optimal performance from CNN. In this regard, the training time and computational complexity tends to be higher as well. Adding maxpool layer is an effective approach to lower the computational complexity while not compromising the overall performance. Therefore, maxpool layer is another important part of CNN based image interpretation which needs to be investigated in demystifying CNN black box. Some AI scientists have experimented with systematic, experimental, and observational data to discover knowledge that can provide interpretability regarding the relations in the components of a model along with the insightful correlation in experiential data. Lundberg et al., (Lundberg et al., n.d.) proposed a unified approach analyzing feature importance provided by complex deep learning architecture to demonstrate the prediction scheme. Simonyan et al., (Simonyan et al., 2013) presented the visualization of deep CNN in image classification through saliency maps and calculating the class score gradient. Alaa et al., (Alaa & Van Der Schaar, n.d.) introduced symbolic metamodeling framework to convert 'black box' to 'white box' for user understanding. Moreover, in western medicine, the approach tends to consider every system and symptom discretely in diagnosing and treating disease. After analyzing each symptom and health irregularity, a decision is made and necessary advice or healthcare is provided (Lam et al., 2012). As every key component of CNN should be explored in diving deeper into the theory of black box, after each convolutional layer, one maxpool layer is added so that the alteration of feature maps can be addressed considering both convolutional and maxpool layers.

After developing the architecture, the model is run three times for all the classes of both datasets and feature maps from each convolution layer are extracted for three iterations. In CNNs, feature map refers to the output of kernel applied to the previous layer or input layer. The term is used as it is a mapping where a particular sort of feature of the image can be found. Convolutional layers look for features for instance lines, curves, edges using kernels. According to the CNN terminology, the kernel is known as 'filter' or 'feature detector' which slides over the image and computes a dot output refers to 'feature map' or 'activation map'. Feature detectors help to detect several features of an image such as edges, vertical/horizontal lines, bends, etc. The feature map is the result of spotting any of these imaging features. Visualization and explaining the characteristic of feature maps can be one of the most convenient approaches to find some useful insight of 'black box'. The assessment is conducted through quantitative analyses. To evaluate the difference between a set of data or groups, statistical analysis is regarded as the most optimal approach. Moreover, the extent of this diversity can be learnt as well through the scores of several statistical tests. Hence, different statistical tests are conducted on the set of feature maps of different convolutional layers and the diversity is presented in a numerical comparison. From the feature maps of each layer, numerical handcrafted features are extracted and statistical analyses are conducted using the features. The experiments are conducted to show the diversity of the feature maps and the degree of this diversity in a quantitative form. In this regard, two statistical evaluations named T-test and ANOVA test are conducted and T-value and F-value are derived respectively. Based on the T-value and F-value the degree of the feature map diversity is evaluated. Null hypothesis (accept/reject) for both of the tests are also shown. Moreover, the values of MSE, PSNR, SSIM, RMSE, DSC, UQI and SAM are also generated using the handcrafted features from the feature maps. These statistical scores are used to showcase the difference of feature maps from the input image and among themselves of same layer. Finally, Mean, Median, Max and Min values are derived from handcrafted features to conduct the iteration by iteration feature map analysis. However, to address the first four analyses, feature maps from any of the classes can be used as the concerns were related to the difference of feature maps among the layers and iterations. Hence, only the feature maps of the benign class of skin cancer dataset and COVID of CT scan dataset are employed to interpret the first four analyses. For analysis-5 all the classes are needed in order to find the difference among the feature maps of the classes for different iteration. The convolution layers tend to generate different kernels for different iterations which results in different appearances of the feature maps. Three iterations are performed. While generating the feature maps a number of black images are produced. Due to mechanism of CNN layers, deep learning models also use these black images for classification images, as it might be possible that they carry necessary information in terms of the interpretation. All the feature maps of a preceding layer are passed to the next layer to extract more information. The high proportion of black images is an aspect of CNNs which requires more attention.

For Analysis-1, we detect the 'black FMs'. However, it is challenging for human eyes to visually detect whether an image is entirely black. There might be some hidden or tiny information/pixels are easily missed by human eyes. Therefore, the 'black FMs' are detected automatically using the condition 'if all the pixels of an image == 0, detect black image'. For this purpose, feature maps for all the three iterations and convolutional layers are extracted and a possible explanation for the production of such a high proportion of 'Black FMs' is discussed. For the

rest of this research only the feature maps after eliminating 'black FM' are considered.

Two aspects are investigated in analysis 2: (i) geometric features and (ii) pixel intensity features. After removing all the 'black FMs', 17 geometric features (Rafid et al., 2022) (Fatema et al., 2022), 'Area', ' Perimeter area ratio', 'Solidity', 'Equivalent Diameter', 'Convex Area', 'Extent', ' Filled Area', 'Major axis length', 'Minor axis length', 'Mean', 'Standard Deviation', 'Shannon Entropy', 'GLCM entropy', 'Skewness', 'kurtosis', 'LBP energy' and 'gabor energy', are derived from the remaining feature maps. For a better understanding of the diversity of the feature maps in terms of pixel intensity levels, 6 intensity-based features comprising pixel brightness, max pixel intensity, number of brightest pixels, contrast level, noise level and energy values are also extracted. To analysis the image pattern, related information is needed to explain and assess the objects in a particular scenario. The informations are extracted from the images in the form of numeric values, referred as features. Features impact greatly in the applications of computer vision for object detection, segmentation and classification. In the automated study of medical imaging, these features are widely-used to describe the characteristic of the images depending on particular disease (Khan et al., 2020). Both geometric and intensity based features comprising texture, intensity, and shape of the objects are crucial to analyze the biomedical images precisely. The features represent the information of contagious portion of an image which is crucial in pattern analysis (Wani & Raza, 2018). Geometric features represent the features of an object such as size, shape, points, lines, orientation and surfaces. The 17 geometrical features are utilized to determine the geometric properties of the images from different angles. Intensity based features refers to the pixel intensity properties such as brightness, contrast, noise etc. These features describe the pixel by pixel alteration from photometric view. Table 2 provides an overview of these features.

The features both from geometric and photometric perspectives are considered to present a comprehensive analysis regarding diverse characteristic of feature maps. In this regard, along with analyzing the geometric changes of the feature maps the photometric changes are also shown to evaluate the degree of changes rigorously. Statistical comparison of these features for the feature maps of convolution 1, convolution 2 and convolution 3 acquired from all the three iterations is done for the benign class of skin cancer dataset and COVID class of CT Scan dataset using T-test and ANOVA test.

The main objective of the third analysis is to investigate (i) the difference between the original images and its resultant feature maps and (ii) the difference between the resultant feature maps for a single image. MSE, PSNR, SSIM, RMSE, DSC, UQI and SAM values are produced for three original input images of the benign and COVID class and their resultant feature maps for the first convolutional layer and iteration 1.

The difference of feature maps in terms of iteration by iteration is explored next analysis. A statistical comparison deriving the values of mean, median, max and min for the three iterations is presented employing the geometric based features of the benign and COVID class respectively for the first convolutional layer.

Finally, in analysis-5, it is investigated which iteration provides the largest difference between the feature maps of two classes, benign and malignant (for skin cancer) and feature maps of three classes, COVID, non-COVID and CAP (for CT Scan) respectively. The geometric features of all the classes across three iterations of all three layers are compared. Here, the analysis is done for both datasets.

5. Model architecture

This research focuses on the analysis of feature maps produced by different convolutional layers. To conduct the experiments, a shallow CNN model with three convolutional layers, each followed by one maxpool layer, is developed. The input image size is $224 \times 224 \times 3$. All the convolutional layers are configured with 16 kernels having the size of 2×2 and the activation function is rectified linear unit (ReLU). The kernel size for all the maxpool layers is 2×2.

Table 2
Description of the geometric and intensity based features.

No.	Feature name	Description
1	Area	Area of the desired region.
2	Perimeter area ratio	Perimeter area ratio determines the horizontal to vertical pixel ratio of an image.
3	Solidity	Solidarity is the ratio of the contour area and the smallest convex haul which covers the area.
4	Equivalent diameter	The equivalent diameter indicates the diameter of a circle with the same ROI surface area.
5	Convex Area	The equivalent diameter indicates the diameter of a circle with the same ROI surface area.
6	Extent	The area of the segmented object is divided by the area of its convex hull denoted as the extent.
7	Filled Area	The filled area is the interpolated pixel value that covers all the ROI areas.
8	Major axis length	Major axis length is the measurement of the pixel distance between the major axis endpoints of the object area.
9	Minor axis length	Minor-axis length is the lowest length of the targeted pixel area.
10	Mean	Mean is the average pixel intensity of the ROI.
11	Standard Deviation	The standard deviation refers to the measurement of the variation of image gray level intensities.
12	Shannon entropy	The average amount of information contained in the ROI area is estimated using the Shannon entropy.
13	GLCM entropy	GLCM entropy calculates the texture feature contents of the segmented object.
14	Skewness	The skewness is a measurement used to assess the symmetry or asymmetry data distribution in the ROI area.
15	Kurtosis	The kurtosis statistic determines whether the tails of a normal distribution of the ROI area are heavy or light.
16	Lbp energy	Local Binary Pattern energy is a texture primitive descriptor of LBP.
17	Gabor energy	Gabor Energy means convoluting an image with a set of Gabor filters. It's a textual feature of GLCM.
18	Pixel brightness	Brightness is the measurement of the overall pixel intensity.
19	Max pixel intensity	The highest pixel intensity of the image.
20	Number of bright pixels	Numbers of the pixels which remain in a particular range close the highest intensity.
21	Contrast level	Contrast defines the difference between the maximum and minimum pixel intensity of the image.
22	Noise level	Random variation of the color intensity of an image
23	Energy	The term energy implies the regional alteration of the specific image quality.

6. Results and analysis

In this section the outcomes related to the five queries are described. As the number of originals is 200, after applying 16 filters to these images the number of feature maps is $200 \times 16 = 3200$.

6.1. Analysis-1: black FM' analysis

When a kernel is applied to an image, the filter moves along the pixels of the image and outputs a two-dimensional array. Each kernel is designed to find some useful features from the images. However, sometimes the values of these two-dimensional arrays turn out to be '0' which indicates that for that particular kernel no features were detected. We are considering these two-dimensional arrays where all the pixel values are' as 'black FM'. Fig 2 is an illustration of such cases for skin cancer dataset.

As illustrated in Fig 3 three random kernels are applied to each channel (R, G, B) of the input image. However, after convolving the three outputs with the three kernels, the final feature map turns out to be black. One explanation for such case might be this set of kernels cannot find any useful features from the image. However, for other images or

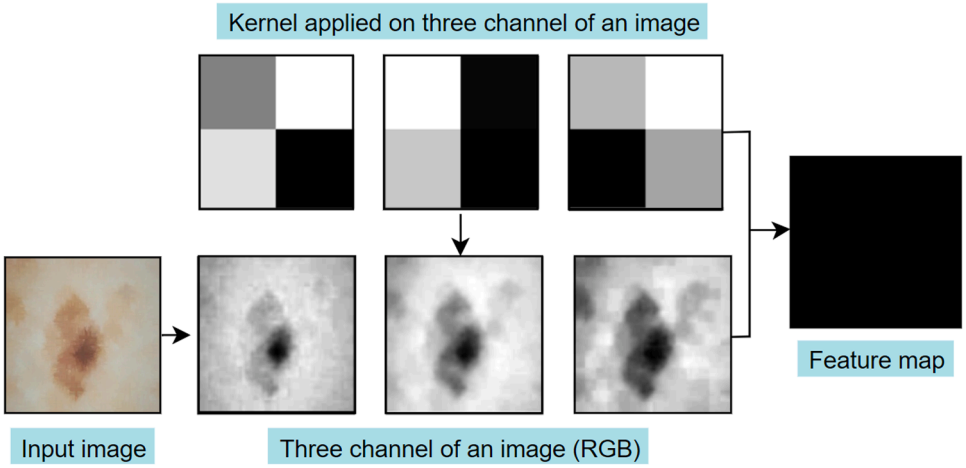


Fig 3. Process of generating feature map.

another dataset these particular kernels might be able to extract features. The reason might be the features these kernels are looking for in an image might not exist in our input image. For instance, if the kernels are designed to extract features like horizontal lines, vertical lines, edges, ridges, corners and more and that image does not contain any horizontal lines etc., then the possible consequence will be that no feature is detected which means a black FM. The main purpose of CNN is to extract different kinds of features from an image but it is possible that not all types of features exist in the images of a particular dataset that a kernel is designed to extract, resulting in a 'black FM' (Mohammed et al., 2022). Nonetheless, these 'black FM' proceed to the next layer of CNN with all the other feature maps achieved from the prior layer. This can be considered an important insight in the CNN 'black box' mystery requiring more analysis.

In this analysis, for each convolution layer, feature maps are generated for three iterations across both datasets. The objective of considering multiple iterations is to conduct a rigorous assessment of the consistency and proportion of 'black FM' outputs for each layer. Fig 4 is an illustration of 32 feature maps for two images of skin cancer dataset for convolution layer 1. The feature maps from 'conv1 0′ to 'conv1 15′ are for image 1, the rest are for image 2.

It can be observed from Fig 4 that along with feature maps containing discrete and useful information, a number of black images are generated. For the first image, 6 out of 16 feature maps are 'black FM' while for the second image 5 out of 16 feature maps are 'black FM' which indicates that the number is high. Therefore, for all the feature maps of each layer and iteration, the number of black images is counted. Table 3 lists the results, where the total number of feature maps is denoted by 'Total FM', the total number of black images is denoted by 'Black FM',

Table 3
Feature map analysis for each convolutional layer and iteration for skin cancer dataset.

layer	class	iteration	Total FM: total feature maps	Black FM: total black images	Remaining FM: total FM – black FM	Proportion of black FM: black FM/ total FM × 100
1	Benign	1	3200	805	2395	25.15%
		2	3200	1454	1746	45.43%
		3	3200	1610	1590	50.31%
2	Benign	1	3200	815	2385	25.46%
		2	3200	1944	1256	60.75%
		3	3200	1022	2178	31.93%
3	Benign	1	3200	1356	1790	42.37%
		2	3200	654	2546	20.43%
		3	3200	1184	2016	37%

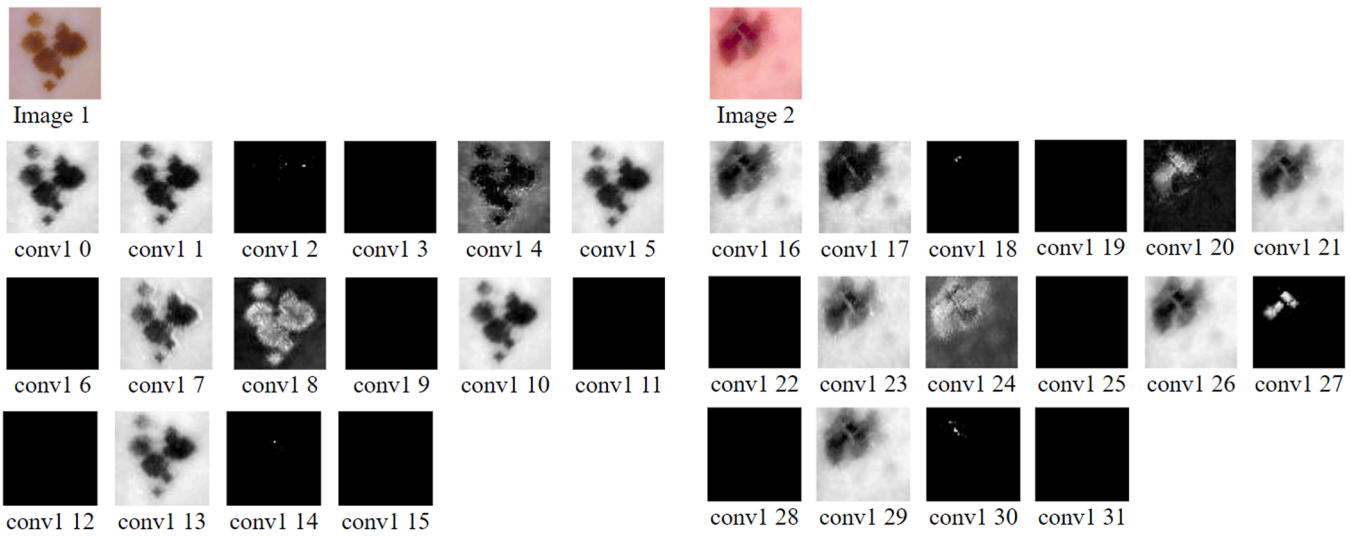


Fig 4. Illustration of 32 feature maps for two input images of skin cancer dataset.

'Remaining FM' indicates the subtraction of 'Black FM' from 'Total FM' and 'Proportion' denotes the percentage of 'Black FM'.

It can be observed from Table 2 that the number of 'black FMs' is not consistent over the iterations and layers. For layer 1, the number of 'black FMs' gradually rises over the iterations. For some cases the proportion is close to 50% or even higher which means half of the feature maps might be resulted in black images. A possible reason for this phenomenon could be the diversity of kernels, which vary depending on the layer and iteration. For example, if a model is run twice with the same dataset, the same set of kernels will not be applied for the two iterations. A 'black FM' is generated when the kernel does not find its targeted feature from an image, based on its structure. As the kernel design changes over iterations, the number of 'black FM' differs for different layers and across multiple iterations. However, these proportions may be different for another dataset having a different modality. Therefore, the 'Black FM' analysis is conducted for an entirely different modality named CT Scan.

In the experiment of using lung scan dataset, the number of black images is comparatively lower than the skin cancer dataset. Fig 5 is an illustration of 32 feature maps for two images of CT Scan dataset for convolution layer 1. The feature maps from 'conv1 0′ to 'conv1 15′ are for image 1, the rest are for image 2.

It can be observed from Fig 5 that along with feature maps containing discrete and useful information, a number of black images are generated. For the first image, 4 out of 16 feature maps are 'black FM' while for the second image 3 out of 16 feature maps are 'black FM' which indicates that the number is quite noticeable, though comparatively lesser than skin cancer dataset. Therefore, for all the feature maps of each layer and iteration, the number of black images is counted. Table 4 showcases the outcomes of 'Black FM' analysis for CT Scan dataset.

It can be observed from Table 4 that the number of 'black FMs' is not consistent over the iterations and layers. For some cases the proportion is close to 20% which means a notable number of feature maps might be resulted in black images. To summarize, a noticeably high proportion of 'black FMs' are found for skin cancer dataset and comparatively low proportion of 'black FMs' are found for the CT Scan dataset. Hence, the phenomenon of 'black FM' which occurs due to the random weight initialization of different kernels, there is no consistency in the number and the phenomenon varies from dataset to dataset. For the rest of our experiments we have worked with the feature maps in the 'Remaining FM' category of both datasets.

Table 4
Feature map analysis for each convolutional layer and iteration for CT Scan dataset.

layer	class	iteration	Total FM: total feature maps	Black FM: total black images	Remaining FM: total FM – black FM	Proportion of black FM: black FM/ total FM × 100
1	Covid	1	3200	277	2923	8.65%
		2	3200	130	3070	4.06%
		3	3200	322	2878	10.06%
2	Covid	1	3200	155	3045	4.84%
		2	3200	278	2922	8.68%
		3	3200	117	3083	3.65%
3	Covid	1	3200	211	2989	6.59%
		2	3200	609	2591	19.03%
		3	3200	76	3124	2.37%

6.2. Analysis-2: layer by layer feature map analysis

With different depth of layers, different types of feature maps are generated, as is illustrated in Fig 6, which also shows the histogram plots.

Fig 6 shows that little similarity exists between the input image and the resultant feature maps over three layers. The accompanied histogram plots illustrate this. The plots show that the outline changes a lot for different layers. Each convolutional layer of the CNN has different filters which result in producing different feature maps from the input image (Prijs et al., 2022). To determine how the feature maps obtained from the three convolution layers are different from one layer to another, a T-test and an ANOVA test are conducted using the 17 geometric features extracted from the feature maps. Convolution layer is developed as a structure having a number of fixed-size kernels that applies complex functions to the input image in order to extract meaningful features. The extraction of the high-level features is performed using consecutive phases of convolutions, nonlinearities, and subsampling mechanisms (Aimar et al., 2019). The feature maps of one convolutional layer are passed through the following convolutional layer (Sarıgül et al., 2019). In this process, the feature maps of the succeeding layer are generated based on the feature maps of the preceding layer. The process continues until the deep-level features are obtained. From the context of this layer to layer relationship, we have

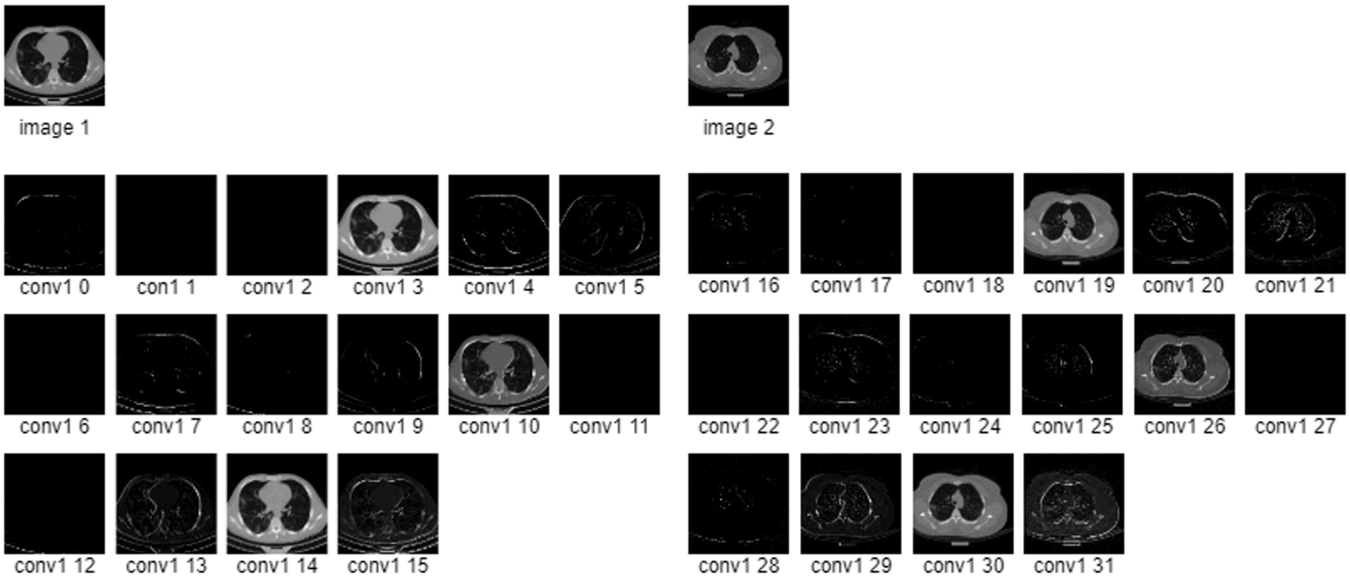


Fig 5. Illustration of 32 feature maps for two input images of CT Scan dataset.

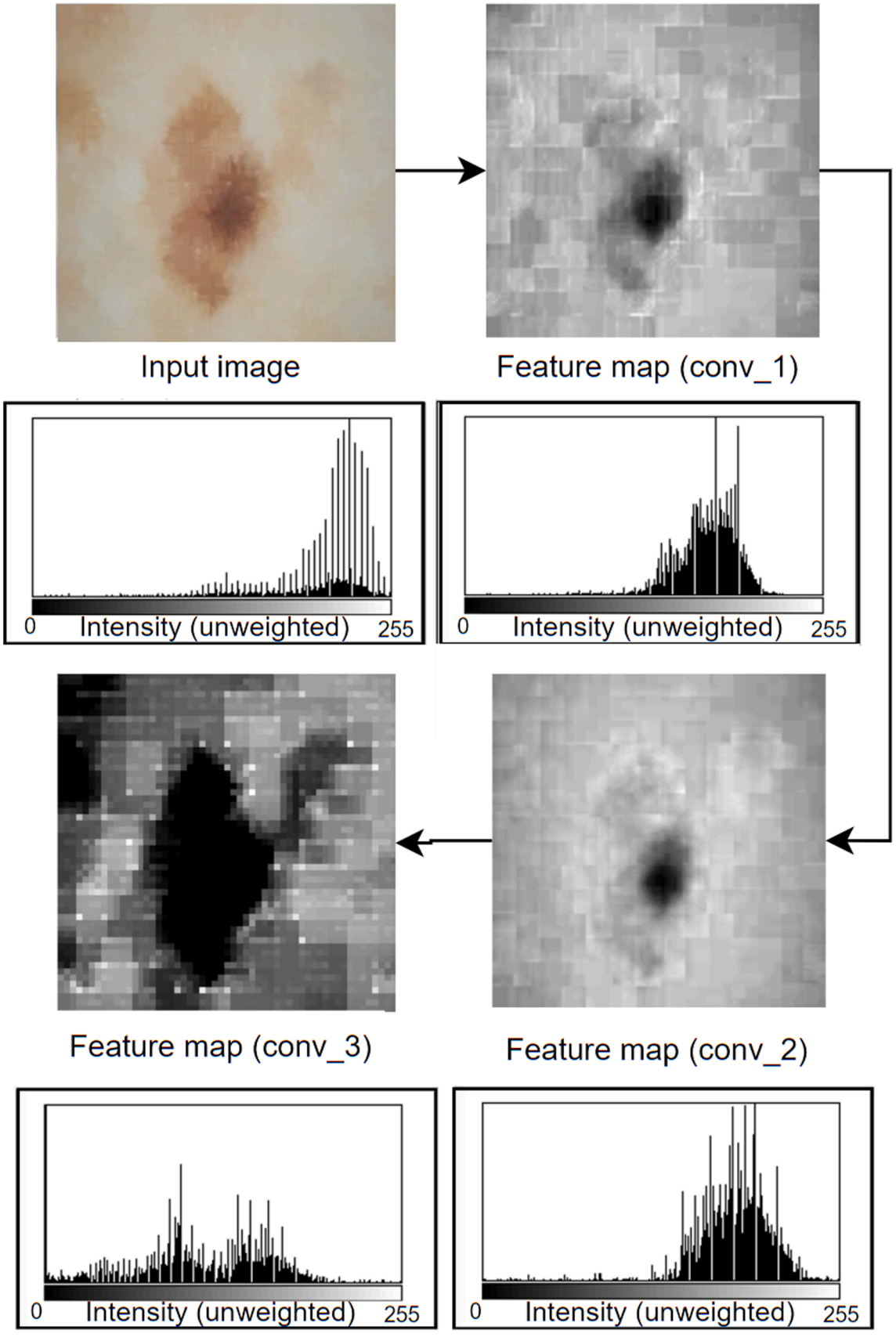


Fig 6. Changes of the feature maps over the layers for skin cancer dataset.

attempted to represent an analysis following a hierarchical manner. The objective to present how the feature maps of the succeeding layer are diverse to the feature maps of the preceding layer. The diversity related to the feature map based on convolutional layers is presented comparing the feature maps of first convolutional layer with second convolutional layer and second convolutional layer with third convolutional layer.

A two sample T-tests is utilized to determine the statistical difference between two distinct groups by comparing their averages. Here, two values are generated, T-value and P-value. The T-value is a method to measure the dissimilarity, the P-value is used to determine whether to accept or reject null hypothesis. Therefore, the T-value primarily provides the difference between two groups and the corresponding P-value is generated to test whether the difference is statistically significant or not (Muriel-Vizcaíno et al., 2017). The null hypothesis is rejected when the P-value ≤ 0.05 (Di Leo & Sardanelli, 2020). In this regard, the null hypothesis is considered as 'similarity between/among the cases. Features maps for the benign class achieved from all the three iterations are considered. For each iteration, interpretation of two cases including T-tests for (i) Conv1_Benign vs Conv2_Benign and (ii) Conv2_Benign and Conv3_Benign are performed. Therefore, for three iterations, a total of six cases are analyzed. Table 5, Table 6 and Table 7 represent the results of the T-tests for the first, second and third iteration respectively.

From Table 5, Table 6 and Table 7, it can be seen that, in terms of geometry, differences between the images are statistically significant for nearly all textural features. Analyzing the three iterations, we see statistically significant dissimilarity for all 17 geometric features for iteration 3 while progressing from conv1 to conv2 and from conv2 to conv3. For iteration 2, dissimilarity is observed for 14 of the 17 features for 'conv1 vs conv2' and for 15 of the 17 features for 'conv2 vs conv3'. For iteration 1, dissimilarity is significant for 15 of the 17 features for as for 'conv1 vs conv2' and for all features for 'conv2 vs conv3'. This demonstrates that the geometrical features of the feature maps change considerably.

An ANOVA test, which determines whether multiple independent groups are statistically equivalent or not, is also done. For each iteration and geometrical feature all the layers are considered. Table 8 represents the results of ANOVA test of class benign for all iterations and layers.

The outcomes of ANOVA tests are that the P values are less than 0.05 in all cases, as can be seen in Table 8 (Kim, 2014). This again demonstrates that the feature maps of different layers are notably different.

The feature maps seemed variable in terms of pixel intensity. Therefore, 7 intensity-based features have also been compared. For every feature map of the three layers and iterations of the benign class, the average values for intensity, RMS pixel brightness, max pixel intensity, number of brightest pixels, contrast level, noise level and energy are calculated. Here, the pixel brightness refers to the overall brightness level of an image, RMS pixel brightness is the RMS of the overall brightness level, Max pixel intensity denotes the highest pixel value presented in an image, the number of bright pixels refer to the count of pixels of which the intensity levels are close to the highest pixel value, the contrast level is the overall contrast of an image, and likewise noise level and energy refer to the overall noise and energy of an image. In order to compute the average numbers of brightest pixels, different threshold ranges were considered for different layers and iterations based on the max pixel intensity value for each layer and iteration. Here, the threshold is calculated by subtracting 20 from the max pixel intensity of each layer. For example, if the max pixel intensity value is 150, then the max threshold is 150 and the min threshold will be 150–20 = 130 and the pixels with intensities is in the range of 130–150 are be counted. All 'remaining FMs' have been used to derive these intensity-based features and the mean value is also calculated. For instance, for conv1, iteration 1, the brightness level for all 2395 feature maps (Table 3) is derived and the average is computed. Table 9 describes the average values of pixel brightness, max pixel intensity, number of bright pixels, contrast level, noise level and energy for each layer, configuration for the benign class.

For a better visualization and understanding the shifting pattern of these features, two bar charts are generated using the values of Table 9. Fig 7 shows the average values of mean pixel brightness, contrast level and noise level and Fig 8 depicts the average values for max pixel intensity, number of bright pixels and energy.

Based on the average values of three features (pixel brightness, contrast and noise), see Table 7 and Fig 7, three findings can be listed:

- The mean pixel brightness for all the layers and iterations are above 100. For the first iteration, the value is found to rise gradually from conv1 to conv3 layer.
- Regarding average contrast level, though a minor fluctuation exists, the values are quite steady over the layers and iterations. Moreover, the average contrast level is higher in layer 3 for every iteration than for the rest of the layers.
- The average noise level gradually increases from layer 1 to 3 and in layer 3 the highest noise level is found (ranging from 5 to 9) whereas in layer 1 the amount of noise is noticeably lower (<1).

Based on the average values of three features (max pixel intensity, number of bright pixels and energy), see Table 8 and Fig 8, three findings can be described:

Table 5
T-test for Conv1_Benign vs Conv2_Benign and Conv2_Benign vs Conv3_Benign.

Features	Conv1_Benign and Conv2_Benign			Conv2_Benign and Conv3_Benign		
	T value	P value	Null Hypothesis	T value	P value	Null Hypothesis
Area	120.69	0	Reject	106.68	0	Reject
Perimeter area ratio	−18.35	9.70E-73	Reject	−60.68	0	Reject
Solidity	−10.01	2.41E-23	Reject	−13.49	1.71E-40	Reject
Equivalent diameter	99.29	0	Reject	89.38	0	Reject
Convex Area	122.64	0	Reject	101.07	0	Reject
Extent	−7.37	1.93E-13	Reject	−13.07	3.27E-38	Reject
Filled Area	72.17	0	Reject	63.32	0	Reject
Major axis length	114.89	0	Reject	95.28	0	Reject
Minor axis length	87.74	0	Reject	72.45	0	Reject
Mean	−6.43	1.32E-10	Reject	−11.97	1.49E-32	Reject
Standard Deviation	−12.14	1.82E-33	Reject	−3.10	0.001927	Reject
Shannon entropy	−0.65	0.51	Accept	−5.55	2.92E-08	Reject
GLCM entropy	5.69	1.31E-08	Reject	4.89	1.02E-06	Reject
Skewness	−0.22	0.82	Accept	9.39	9.47E-21	Reject
Kurtosis	2.58	0.009	Reject	7.52	7.09E-14	Reject
Lbp energy	−4.76	1.99E-06	Reject	9.23	4.20E-20	Reject
Gabor energy	2.97	0.002	Reject	15.42	2.87E-52	Reject

Table 6
T-test for iteration 2: Conv1_Benign vs Conv2_Benign and Conv2_Benign vs Conv3_Benign.

Features	Conv1_Benign and Conv2_Benign			Conv2_Benign and Conv3_Benign		
	T value	P value	Null Hypothesis	T value	P value	Null Hypothesis
Area	74.27	0	Reject	51.68	0	Reject
Perimeter area ratio	−11.51	5.11E-30	Reject	−46.90	0	Reject
Solidity	−5.94	3.21E-09	Reject	−9.59	1.68E-21	Reject
Equivalent diameter	59.80	0	Reject	41.34	0	Reject
Convex Area	77.11	0	Reject	52.03	0	Reject
Extent	−9.08	2.44E-19	Reject	−2.48	0.01	Reject
Filled Area	53.87	0	Reject	41.06	0	Reject
Major axis length	68.12	0	Reject	47.55	0	Reject
Minor axis length	58.05	0	Reject	39.53	0	Reject
Mean	0.003	0.99	Accept	2.30	0.02	Reject
Standard Deviation	−7.49	8.95E-14	Reject	0.85	0.39	Accept
Shannon entropy	2.81	0.004	Reject	−5.21	1.99E-07	Reject
GLCM entropy	6.01	2.07E-09	Reject	−0.51	0.607438	Accept
Skewness	−5.75	9.47E-09	Reject	9.58	3.54E-21	Reject
Kurtosis	−1.61	0.10	Accept	10.39	2.07E-24	Reject
Lbp energy	−6.92	5.32E-12	Reject	10.70	4.91E-26	Reject
Gabor energy	−1.87	0.06	Accept	6.94	4.80E-12	Reject

Table 7
T-test for iteration 3: Conv1_Benign vs Conv2_Benign and Conv2_Benign vs Conv3_Benign.

Features	Conv1_Benign and Conv2_Benign			Conv2_Benign and Conv3_Benign		
	T value	P value	Null Hypothesis	T value	P value	Null Hypothesis
Area	43.4	0	Reject	83.56	0	Reject
Perimeter area ratio	−15.81	1.42E-54	Reject	−47.57	0	Reject
Solidity	−9.47	4.71E-21	Reject	−13.76	3.63E-42	Reject
Equivalent diameter	32.35	0	Reject	63.43	0	Reject
Convex Area	47.47	1.592746e-317	Reject	77.88	0	Reject
Extent	−2.56	0.01	Reject	−5.48	4.48E-08	Reject
Filled Area	36.74	0	Reject	54.52	0	Reject
Major axis length	42.31	0	Reject	67.98	0	Reject
Minor axis length	35.58	0	Reject	57.04	0	Reject
Mean	−7.48	8.84E-14	Reject	−4.28	1.88E-05	Reject
Standard Deviation	−10.55	1.11E-25	Reject	2.33	0.01	Reject
Shannon entropy	−12.96	1.61E-37	Reject	−5.86	4.98E-09	Reject
GLCM entropy	−11.44	9.34E-30	Reject	2.27	0.02	Reject
Skewness	7.86	5.50E-15	Reject	13.63	5.94E-41	Reject
Kurtosis	9.36	2.23E-20	Reject	14.64	1.61E-46	Reject
Lbp energy	10.32	1.40E-24	Reject	12.49	3.68E-35	Reject
Gabor energy	10.08	1.33E-23	Reject	9.91	6.69E-23	Reject

Table 8
Results of ANOVA test for all three layers, for iterations 1, 2 and 3 for class Benign.

Features	Iteration 1			Iteration 2			Iteration 3		
	F value	P value	Null Hypothesis	F value	P value	Null Hypothesis	F value	P value	Null Hypothesis
Area	16,082.45	0	Reject	8295.95	0	Reject	3401.52	0	Reject
Perimeter area ratio	3255.62	0	Reject	2146.70	0	Reject	2133.07	0	Reject
Solidity	287.79	0	Reject	163.37	9.68E-70	Reject	248.83	0	Reject
Equivalent diameter	13,574.71	0	Reject	7462.78	0	Reject	2613.80	0	Reject
Convex Area	16,056.32	0	Reject	8608.60	0	Reject	3921.18	0	Reject
Extent	226.49	5.73E-96	Reject	121.89	1.42E-52	Reject	30.24	8.55E-14	Reject
Filled Area	5623.87	0	Reject	4260.75	0	Reject	2328.01	0	Reject
Major axis length	17,776.97	0	Reject	9752.68	0	Reject	4099.16	0	Reject
Minor axis length	10,121.26	0	Reject	6784.73	0	Reject	2946.93	0	Reject
Mean	166.18	3.37E-71	Reject	4.91	0.007	Reject	69.65	1.24E-30	Reject
Standard Deviation	131.39	1.01E-56	Reject	49.67	4.09E-22	Reject	74.70	9.01E-33	Reject
Shannon entropy	20.44	1.40E-09	Reject	16.88	4.89E-08	Reject	203.57	2.90E-86	Reject
GLCM entropy	54.81	2.42E-24	Reject	36.82	1.28E-16	Reject	107.19	1.82E-46	Reject
Skewness	28.58	4.32E-13	Reject	60.16	1.38E-26	Reject	162.46	1.94E-69	Reject
Kurtosis	16.82	5.15E-08	Reject	37.88	4.49E-17	Reject	159.71	2.64E-68	Reject
Lbp energy	37.32	7.55E-17	Reject	68.47	4.07E-30	Reject	238.69	0	Reject
Gabor energy	168.34	4.28E-72	Reject	33.22	4.49E-15	Reject	189.61	1.41E-80	Reject

- The average max pixel intensity for all the layers and iterations is above 200 and exhibits quite a constant pattern.
- The average number of bright pixels is remarkably higher for the first layer comparing to the last two layers (across all iterations). The average number of bright pixels is significantly greater in layer 1 and

Table 9
Average values for pixel brightness, max pixel intensity, number of bright pixels, contrast level, noise level, energy.

Layer	Iteration	pixel brightness (avg)	Max pixel intensity (avg)	Number of bright pixels (avg)	Contrast level (avg)	Noise level (avg)	Energy (avg)
Conv1	1	121.68	251.51	438.69	40.65	2.104	44.98
	2	119.56	251.56	499.25	39.22	1.717	37.018
	3	105.73	251.84	580.10	38.494	1.068	25.95
Conv2	1	133.28	252.27	244.56	46.41	2.51	65.02
	2	119.11	252.38	239.00	44.11	1.89	56.04
	3	123.66	252.33	207.31	45.65	2.19	61.38
Conv3	1	155.68	253.00	95.82	47.79	5.93	170.97
	2	113.75	251.63	27.53	43.68	8.51	211.37
	3	132.10	252.39	49.45	44.38	5.66	174.28

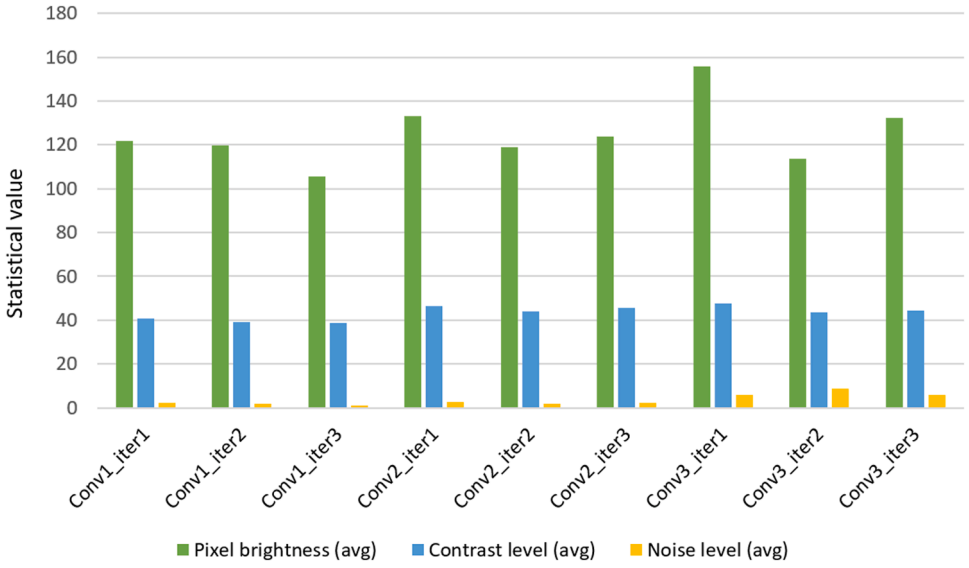


Fig 7. Average values of mean pixel brightness, contrast level and noise level for all the layers and iteration.

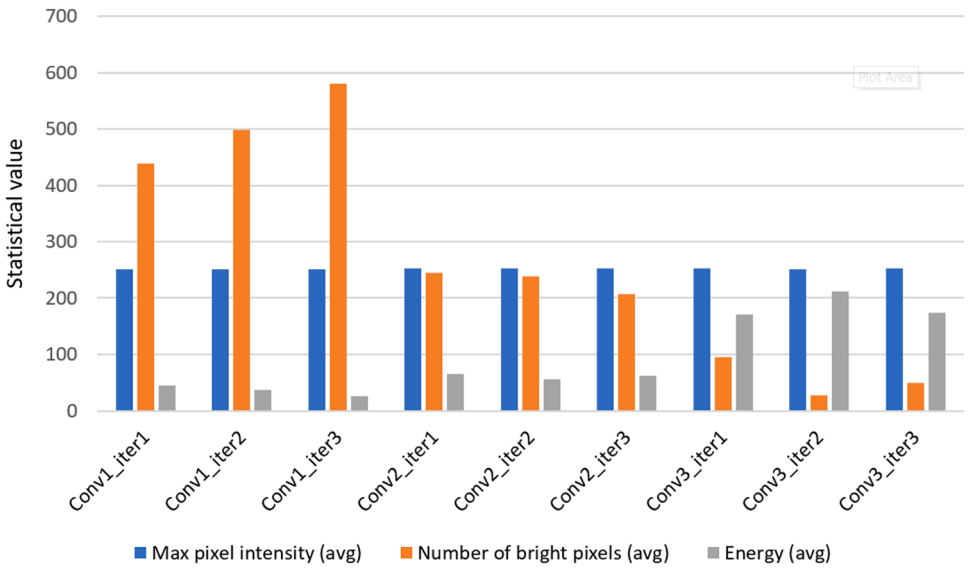


Fig 8. Average values max pixel intensity, number of bright pixels and energy for all the layers and iteration.

dropped markedly for layer 2 and then for layer 3. The reason may be that over the layers the size of the feature maps decreases due to the maxpool layer after each convolution layer.

- The average energy level of the feature maps somewhat increases from conv1 to conv2 and significantly rises in conv3. However, a fluctuation is found for all the layers based on iterations.

The similar experiments are conducted following the same process for CT Scan dataset as well. Fig 9 illustrates the feature maps of three different layers along with their histogram plots.

The feature maps and accompanied histogram plots of Fig 9 show that little similarity exists between the input image and the resultant feature maps over three layers. To analysis this diversity, Features maps

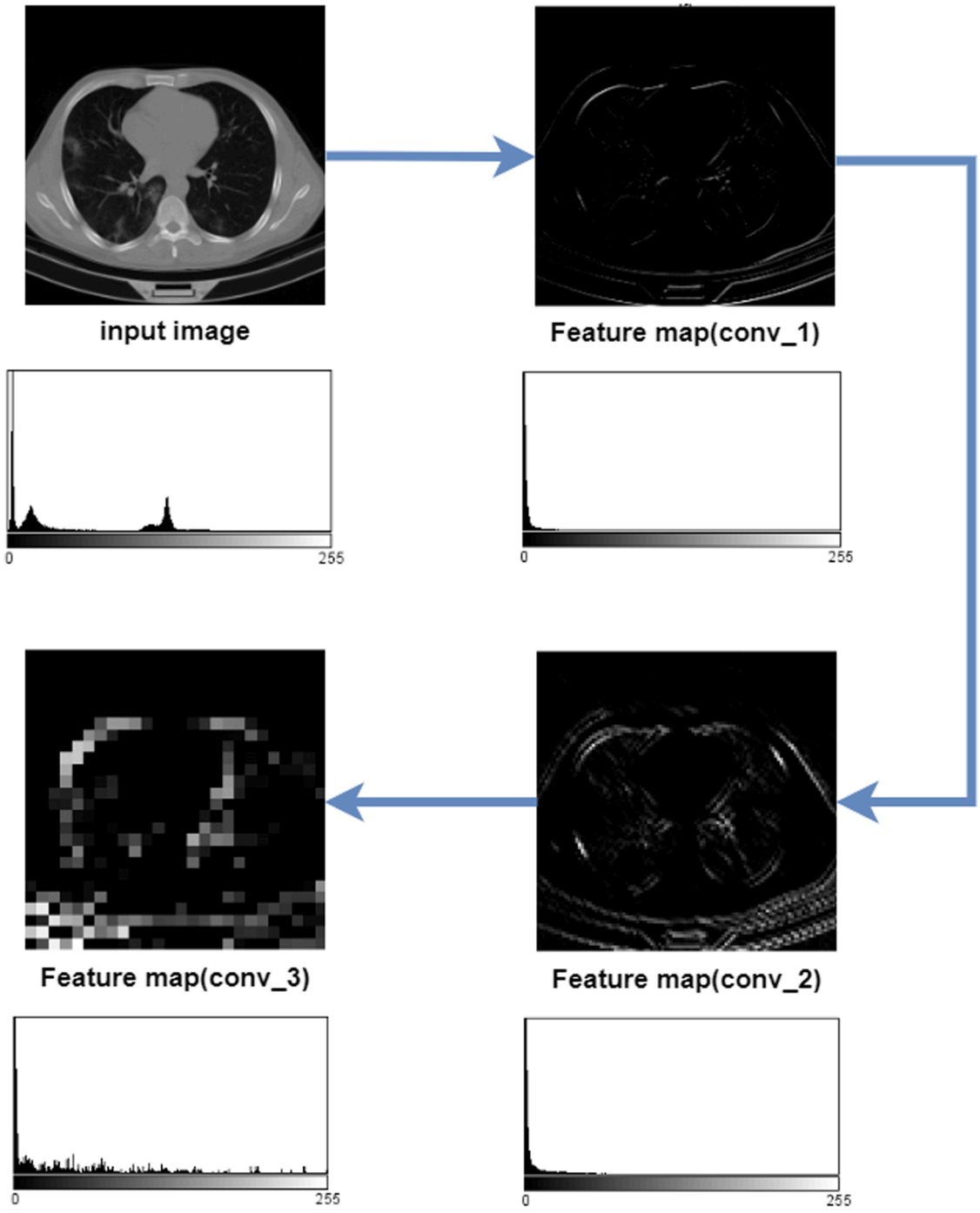


Fig 9. Changes of the feature maps over the layers for CT Scan dataset.

for the COVID class achieved from all the three iterations are considered. For each iteration, interpretation of two cases including T-tests for (i) Conv1_COVID vs Conv2_COVID and (ii) Conv2 COVID and Conv3 COVID are performed. Therefore, for three iterations, a total of six cases are analyzed. Table 10, Table 11 and Table 12 represent the results of the T-tests for the first, second and third iteration respectively.

From Table 10, Table 11 and Table 12, it can be seen that, in terms of geometry, differences between the images are statistically significant for nearly all geometric features. Analyzing the three iterations, we see statistically significant dissimilarity for all 17 geometric features for iteration 3 while progressing from conv1 to conv2 and from conv2 to conv3. For iteration 2, dissimilarity is observed for 14 of the 17 features for 'conv1 vs conv2' and for 15 of the 17 features for 'conv2 vs conv3'. For iteration 1, dissimilarity is significant for 15 of the 17 features for as for 'conv1 vs conv2' and for all features for 'conv2 vs conv3'. This demonstrates that the geometrical features of the feature maps change considerably.

Similarly, ANOVA test is conducted across all the iterations and layers for class COVID. Table 13 represents the results of ANOVA test of class COVID for all iterations and layers.

It can be observed from table 5 that for all the features across three iterations, the null hypothesis becomes rejected. This further validates that the feature maps of different layers are notably different.

Afterwards, for each layer, configuration for the COVID class, the average values of pixel brightness, max pixel intensity, number of bright pixels, contrast level, noise level and energy are derived using the same

Table 10
T-test for iteration 1: Conv1_COVID vs Conv2_COVID and Conv2_COVID vs Conv3 COVID.

Features	Conv1_COVID and Conv2_COVID			Conv2_COVID and Conv3_COVID		
	T value	P value	Null Hypothesis	T value	P value	Null Hypothesis
Area	102.09	0	Reject	99.04	0	Reject
Perimeter area ratio	−37.73	1.92E-273	Reject	−26.92	1.53E-150	Reject
Solidity	−8.75	2.83E-18	Reject	−25.85	8.66E-138	Reject
Equivalent diameter	100.99	0	Reject	103.30	0	Reject
Convex Area	407.47	0	Reject	226.69	0	Reject
Extent	19.07	1.02E-78	Reject	16.07	5.74E-57	Reject
Filled Area	66.75	0	Reject	63.39	0	Reject
Major axis length	299.08	0	Reject	224.59	0	Reject
Minor axis length	0.00	1	Accept	0.00	1	Accept
Mean	−8.99	3.29E-19	Reject	−14.45	1.39E-46	Reject
Standard Deviation	−5.62	2.01E-08	Reject	−29.68	1.97E-181	Reject
Shannon entropy	−7.77	8.87E-15	Reject	−8.22	2.56E-16	Reject
GLCM entropy	−4.18	2.98E-05	Reject	2.58	0.009791	Reject
Skewness	6.18	6.98E-10	Reject	15.91	1.48E-55	Reject
Kurtosis	5.82	6.16E-09	Reject	14.81	3.74E-48	Reject
Lbp energy	1.05	0.293609	Accept	8.18	3.64E-16	Reject
Gabor energy	16.91	9.33E-63	Reject	31.99	2.46E-208	Reject

Table 11
T-test for iteration 2: Conv1_COVID vs Conv2_COVID and Conv2_COVID vs Conv3_COVID.

Features	Conv1_COVID and Conv2_COVID			Conv2_COVID and Conv3_COVID		
	T value	P value	Null Hypothesis	T value	P value	Null Hypothesis
Area	55.10	0	Reject	105.47	0	Reject
Perimeter area ratio	−51.50	0	Reject	−4.62	3.85E-06	Reject
Solidity	−20.54	2.01E-90	Reject	−18.13	7.66E-71	Reject
Equivalent diameter	48.51	0	Reject	119.80	0	Reject
Convex Area	214.12	0	Reject	280.05	0	Reject
Extent	−1.78	0.07589	Accept	12.36	1.23E-34	Reject
Filled Area	48.06	0	Reject	79.29	0	Reject
Major axis length	188.25	0	Reject	160.11	0	Reject
Minor axis length	−1.47	0.142054	Accept	−128.10	0	Reject
Mean	−5.75	9.61E-09	Reject	−9.16	7.16E-20	Reject
Standard Deviation	−23.57	2.22E-116	Reject	−20.98	8.73E-94	Reject
Shannon entropy	−21.18	5.64E-96	Reject	9.89	7.19E-23	Reject
GLCM entropy	−20.49	4.65E-90	Reject	17.73	1.30E-68	Reject
Skewness	29.60	5.12E-171	Reject	−12.78	9.77E-37	Reject
Kurtosis	21.93	2.05E-99	Reject	−15.75	4.32E-54	Reject
Lbp energy	19.84	1.79E-84	Reject	−15.38	2.57E-52	Reject
Gabor energy	31.06	7.20E-197	Reject	19.29	3.72E-80	Reject

Table 12
T-test for iteration 3: Conv1_COVID vs Conv2_COVID and Conv2_COVID vs Conv3_COVID.

Features	Conv1_COVID and Conv2_COVID			Conv2_COVID and Conv3_COVID		
	T value	P value	Null Hypothesis	T value	P value	Null Hypothesis
Area	72.83	0	Reject	72.52	0	Reject
Perimeter area ratio	−34.31	6.25E-233	Reject	−39.85	0.00E+00	Reject
Solidity	−14.57	2.36E-47	Reject	−37.21	2.29E-260	Reject
Equivalent diameter	67.72	0	Reject	60.66	0	Reject
Convex Area	277.61	0	Reject	126.80	0	Reject
Extent	6.11	1.05E-09	Reject	11.33	1.76E-29	Reject
Filled Area	53.09	0	Reject	53.63	0	Reject
Major axis length	221.82	0	Reject	129.74	0	Reject
Minor axis length	1.46	0.144963	Accept	−135.61	0	Reject
Mean	−18.50	2.20E-74	Reject	−28.55	1.99E-168	Reject
Standard Deviation	−16.17	1.25E-57	Reject	−32.47	6.92E-214	Reject
Shannon entropy	−10.32	9.43E-25	Reject	−20.09	6.03E-87	Reject
GLCM entropy	−5.97	2.48E-09	Reject	−9.25	3.26E-20	Reject
Skewness	3.41	0.000645	Reject	22.06	5.23E-102	Reject
Kurtosis	3.76	0.000173	Reject	19.69	9.08E-82	Reject
Lbp energy	2.78	0.005444	Reject	16.91	1.21E-62	Reject
Gabor energy	27.81	1.88E-160	Reject	37.35	2.78E-276	Reject

process described above. Table 14 describes the average values of pixel brightness, max pixel intensity, number of bright pixels, contrast level, noise level and energy.

For a better visualization and understanding the shifting pattern of these features, two bar charts are generated using the values of Table 14. Fig 10 shows the average values of mean pixel brightness; contrast level

Table 13
Results of ANOVA test for all three layers, for iterations 1, 2 and 3 for class Benign.

Features	Iteration 1			Iteration 2			Iteration 3		
	F value	P value	Null Hypothesis	F value	P value	Null Hypothesis	F value	P value	Null Hypothesis
Area	13,491.79	0	Reject	4324.19	0	Reject	6927.93	0	Reject
PA ratio	1846.41	0	Reject	1246.73	0	Reject	2802.00	0	Reject
Solidity	726.85	1.18E-294	Reject	607.49	6.17E-249	Reject	1535.83	0	Reject
Equivalent diameter	17,121.55	0	Reject	5472.33	0	Reject	7407.78	0	Reject
Convex Area	210,049.48	0	Reject	58,724.80	0	Reject	95,465.72	0	Reject
Extent	604.61	4.95E-248	Reject	77.28	5.12E-34	Reject	150.76	3.43E-65	Reject
Filled Area	5311.41	0	Reject	3087.52	0	Reject	3548.64	0	Reject
Major axis length	131,976.94	0	Reject	48,268.28	0	Reject	68,506.46	0	Reject
Minor axis length			Reject	16,909.78	0	Reject	18,447.19	0	Reject
Mean	285.35	4.16E-121	Reject	99.51	1.71E-43	Reject	1171.03	0	Reject
Standard Deviation	622.30	7.60E-255	Reject	851.26	0	Reject	1258.57	0	Reject
Shannon entropy	133.92	4.33E-58	Reject	225.32	2.54E-96	Reject	502.49	3.06E-208	Reject
GLCM entropy	11.05	1.62E-05	Reject	242.87	1.38E-103	Reject	122.14	4.19E-53	Reject
Skewness	217.20	5.52E-93	Reject	669.64	1.03E-272	Reject	297.45	4.75E-126	Reject
Kurtosis	186.28	4.26E-80	Reject	435.88	8.39E-182	Reject	128.98	5.33E-56	Reject
Lbp energy	47.55	2.82E-21	Reject	207.53	6.19E-89	Reject	227.74	2.37E-97	Reject
Gabor energy	1232.76	0	Reject	1105.37	0	Reject	2234.66	0	Reject

Table 14
Average values for pixel brightness, Max pixel intensity, number of bright pixels, contrast level, noise level and energy.

Layer	Iteration	pixel brightness (avg)	Max pixel intensity (avg)	Number of bright pixels (avg)	Contrast level (avg)	Noise level (avg)	Energy (avg)
Conv1	1	25.01	252.30	4.64	30.40	4.92	177.78
	2	14.64	252.20	3.78	20.63	3.91	128.21
	3	19.49	252.31	4.88	26.14	5.09	194.24
Conv2	1	30.17	251.63	2.86	32.41	8.82	304.17
	2	17.01	252.11	2.50	28.18	10.86	368.43
	3	29.88	251.66	2.91	32.52	8.89	327.31
Conv3	1	39.92	251.58	2.48	41.57	17.87	632.42
	2	20.57	250.97	2.12	33.90	17.84	640.72
	3	49.57	251.69	2.83	44.86	19.06	672.94

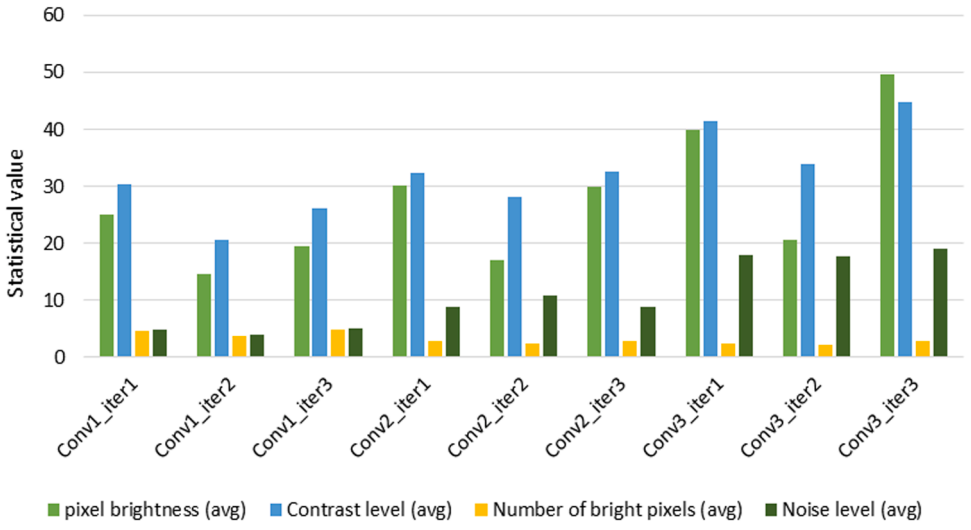


Fig 10. Average values of mean pixel brightness, contrast level, number of bright pixels and noise level for all the layers and iteration.

number of bright pixels and noise level and Fig 11 depicts the average values for max pixel intensity and energy.

Based on the average values of four features (pixel brightness, contrast, number of bright pixels and noise), (see Table 14 and Fig 10), these findings can be listed:

- The mean pixel brightness for all the layers and iterations are above 10 and for most of the cases above 20. For the first iteration, the value is found to rise gradually from conv1 to conv3 layer.
- Regarding average contrast level, though a minor fluctuation exists in the iterations, the values are observed to rise over the layers. Moreover, the average contrast level is higher in layer 3 for every iteration than the rest of the layers.
- The average number of bright pixels is higher for the first layer comparing to the last two layers (across all iterations). For layer 2 and then for layer 3, the number is quite stable.
- The average noise level gradually increases from layer 1 to 3 and in layer 3 the highest noise level is found (ranging from 17 to 19) whereas in layer 1 the amount of noise is noticeably lower (<5).

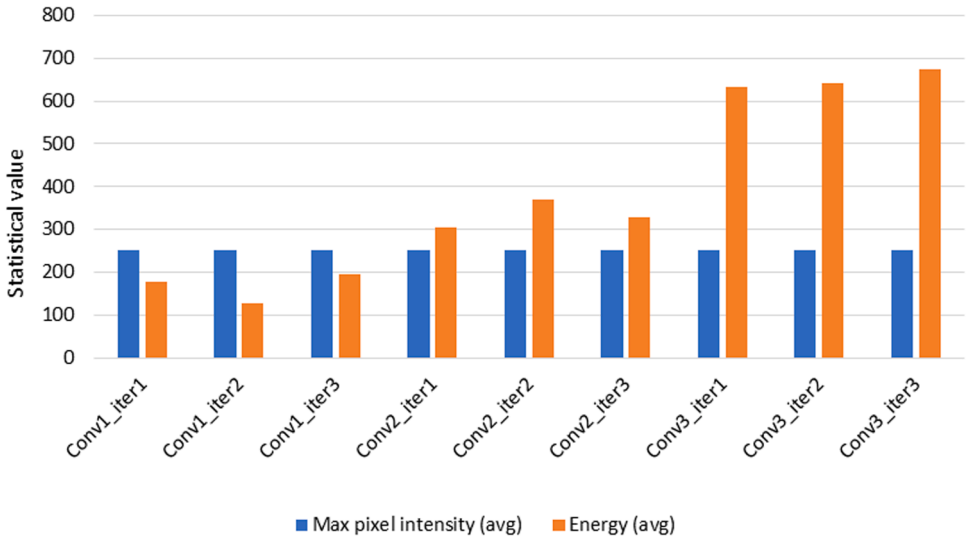


Fig 11. Average values max pixel intensity and energy for all the layers and iteration.

Based on the average values of two features (max pixel intensity and energy), (see Table 14 and Fig 11), these findings can be described:

- The average max pixel intensity for all the layers and iterations is above 200 and exhibits quite a constant pattern.
- The average energy level of the feature maps somewhat increases from conv1 to conv2 and significantly rises in conv3. However, a fluctuation is found for all the layers based on iterations.

To summarize, a photometric/intensity-based alteration occurs for the feature maps of different layers for both datasets. Furthermore, findings of the both datasets based on statistical tests of geometric features and photometric features are closely similar. The explanation for these findings could be that different kernels of different layers extract different types of features, resulting in diversity of the feature maps. For instance, the initial layers of CNN extract the basic features and with increasing the depth more hidden and complex features are extracted. It is therefore predictable that there will be differences among the basic-level features and deep-level features. Moreover, along with extracting diverse features based on geometry, alteration is found for intensity-based features as well. Overall, a CNN extracts different kinds of information from an image and also highlights important regions in different intensity levels. Thus, the learning of the model becomes more efficient which leads to a better classification/detection performance.

6.3. Analysis-3: difference of the feature maps with the original image and with other feature maps

In CNN, applying several kernels to the inputs generates feature maps with an arbitrary pattern that correspond to diverse characteristics of the input tensors, hence various kernels are regarded as distinct feature extractors (Patil & Rane, 2021). To visualize this, histogram

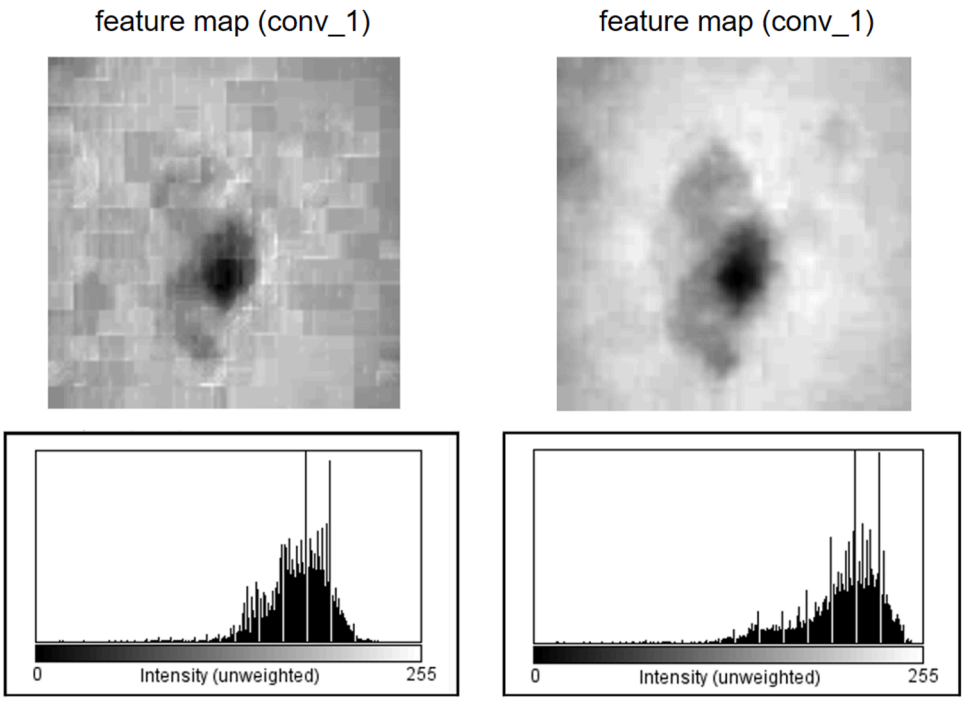


Fig 12. Histograms for two feature maps of same input image of skin cancer dataset.

plots are generated for two feature maps of the same image and the same layer. Fig 12 showcases the two feature maps and their associated histogram plots for skin cancer dataset.

It can be observed that the pattern of the histograms for the feature maps is quite different to evaluate the difference between the original image and the resultant feature maps, seven statistical similarity measures are derived: MSE, PSNR, SSIM, RMSE, DSC, UQI and SAM. Three original images with their 16 feature maps of first convolutional layer of iteration 1 have been used to calculate these values. All the resultant 'Black FM' are excluded. Fig 13 shows three original input images of skin cancer dataset and their resultant feature maps (excluding 'black FM') for which the values are derived.

Table 15 lists the values of MSE, PSNR, SSIM, RMSE, DSC, UQI and SAM for the original image and their resultant feature maps. Here 'FM' denotes the feature maps which are compared with the original images. For example, 'org0-fm1' indicates that the original image 'org0' is being compared with the feature maps 'fm1' (Fig 13).

Alike skin cancer dataset, for CT Scan dataset as well, similar statistical tests are conducted for class COVID. Three original images with their 16 feature maps of first convolutional layer of iteration 1 have been used to calculate these values. Fig 14 shows three original input images of CT Scan dataset and their resultant feature maps (excluding 'black FM') for which the values are derived.

Table 16 showcases the values of MSE, PSNR, SSIM, RMSE, DSC, UQI and SAM for the original image and their resultant feature maps.

It can be observed from Table 15 and Table 16 that for three different images with their respective feature maps of both datasets, the values of MSE, PSNR, SSIM, RMSE, DSC, UQI and SAM differ considerably. If we analyze the values more closely, the MSE values are large for the feature maps of every image, indicating a significant difference of the original image with its subsequent feature maps. Likewise, the values of PSNR range from 3 to 14 for skin cancer dataset and 10 to 24 for CT Scan dataset for the feature maps of every image where a value greater than 30 indicates similarity between two images. Looking at the three other values (SSIM, RMSE and DSC), the results do not meet the required threshold values (SSIM = close to 1, RMSE = close to 0, DSC = close to 1) to affirm the similarity (Aslahishahri et al., 2021). Similarly, a UQI value close to 1 and a SAM value closer to 0 indicate resemblances among two images. Based on that, UQI and SAM values from both tables indicate no similarity between the feature maps and the original images. To summarize, the feature maps are significantly different from their input original image. Furthermore, for a single input image, the corresponding feature maps were providing different similarity measures. The feature maps are notably different from one another. The scores of MSE, PSNR, SSIM, RMSE, DSC, UQI and SAM for the feature maps of a particular image were not close to one another which validate the theory that the kernels applied by the convolutional layers extract different types of information from an image, outputting different feature maps. In addition, the number of 'Black FM' for the feature maps of a particular image is also different. To conclude: feature maps for a particular image are significantly different from the input image and also from one to another.

Table 15
Values of MSE, PSNR, SSIM, RMSE, DSC, UQI and SAM for skin cancer dataset.

FM	MSE	PSNR	SSIM	RMSE	DSC	UQI	SAM
org0-fm1	3437.78	12.76	0.44	0.87	0.52	0.84	0.29
org0-fm2	3123.85	13.18	0.79	0.84	0.53	0.83	0.12
org0-fm3	2146.35	14.81	0.82	0.81	0.53	0.86	0.12
org0-fm4	3008.20	13.34	0.84	0.88	0.52	0.86	0.08
org0-fm5	26,296.46	3.93	0.40	1	0.41	0.12	1.37
org0-fm6	2041.70	15.03	0.81	0.81	0.54	0.89	0.13
org1-fm1	2233.94	14.64	0.68	0.82	0.53	0.16	1.30
org1-fm2	17,982.76	5.58	0.06	1	0.48	016	1.30
org1-fm3	18,637.24	5.42	0.02	1	0.44	0.03	1.38
org1-fm4	3292.83	12.95	0.68	0.93	0.52	0.77	0.19
org1-fm5	18,840.01	5.37	0.01	1	0.43	0.02	1.41
org1-fm6	18,515.73	5.45	0.02	1	0.43	0.11	1.37
org1-fm7	16,776.17	5.88	0.07	1	0.43	0.02	1.11
org1-fm8	19,232.39	5.29	0.004	1	0.43	0.05	1.53
org1-fm8	1860.95	15.43	0.72	0.78	0.54	0.86	0.17
org2-fm1	4311.80	11.78	0.77	0.94	0.50	0.79	0.21
org2-fm2	30,295.85	3.31	0.03	1	0.41	0.10	1.30
org2-fm3	7855.10	9.17	0.72	1	0.49	0.68	0.20
org2-fm4	31,399.37	3.16	0.005	1	0.37	0.03	1.40
org2-fm5	31,051.60	3.20	0.01	1	0.37	0.05	1.37
org2-fm6	27,050.86	3.80	0.11	1	0.38	0.03	0.95
org2-fm7	4046.16	12.06	0.79	0.97	0.50	0.82	0.18

6.4. Analysis-4: iteration by iteration feature map analysis

From Table 3 it can be observed that, for diverse iteration, different numbers of 'Black FM' are produced which strongly indicates that there is divergence in the feature maps for different iterations of the same layer. Fig 15 and Fig 16 illustrate that for each iteration how different feature maps are produced. In this regard, the visualization has been done only for skin cancer dataset.

Fig 15 and Fig 16 depict that for the same image and the same layer, two different feature maps are produced. Closely observing the figures, it can be seen that the kernels were different for each iteration, resulting in the extraction of different feature from the input. In order to evaluate the difference between the iterations more rigorously, Mean, Median, Max and Min values of the 17 geometrical features are derived. Three

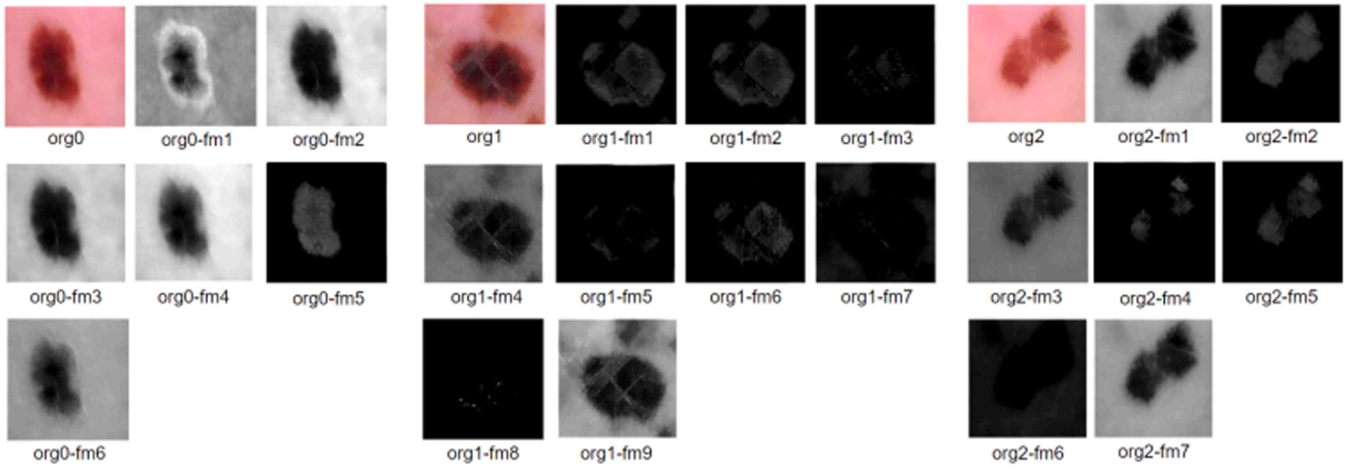


Fig 13. Illustration of 3 original images with resultant feature maps of skin cancer dataset.

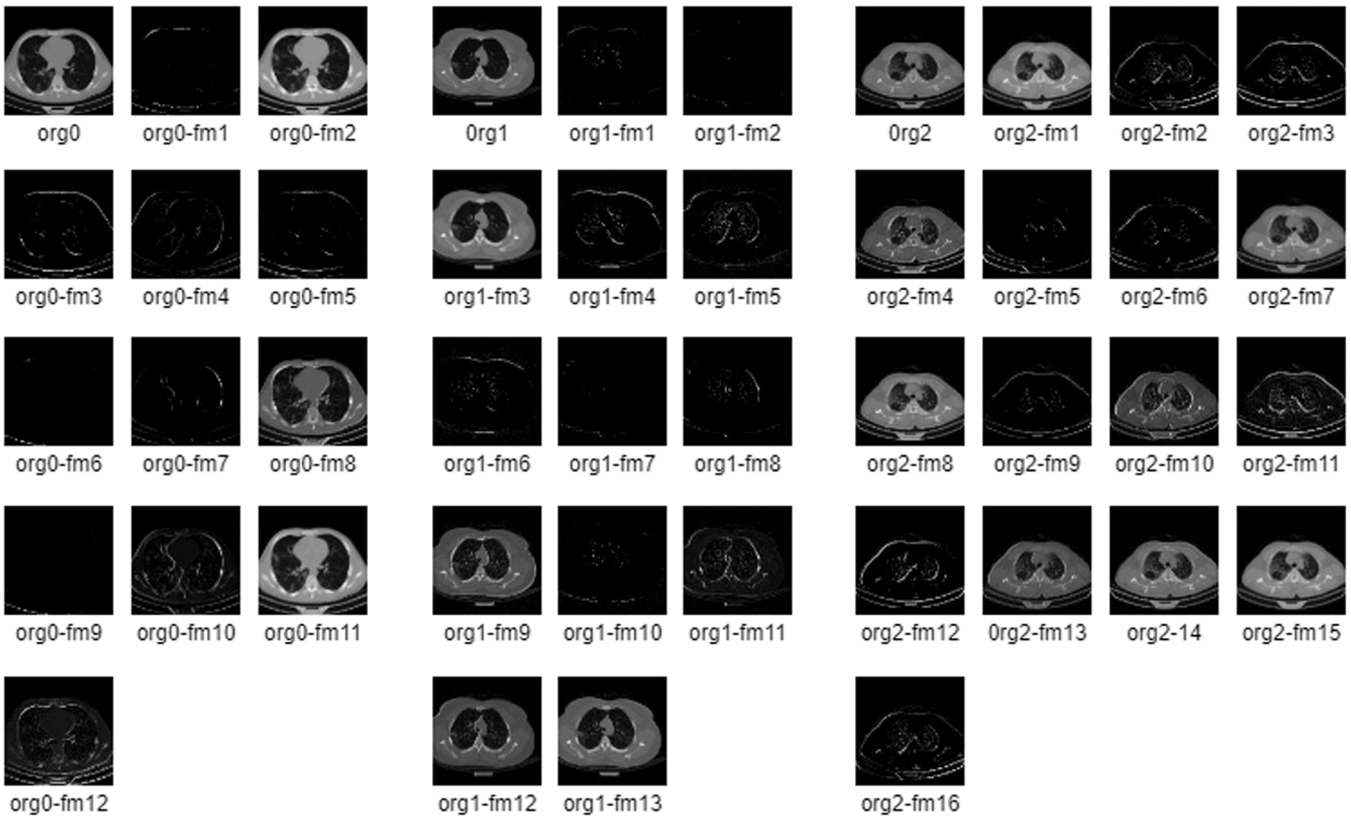


Fig 14. Illustration of 3 original images with resultant feature maps of CT Scan dataset.

iterations for the first layer and for the benign class have been considered. Table 17 illustrates the average mean, median, Max and Min value for the feature maps of the three iterations for Convolution 1 layer.

It is evident from Table 17 that for each feature there are differences among Mean, Median, Max and Min values for the different iterations. For some features, such as Area, Equivalent diameter, Convex Area, Filled Area, Major axis length, Minor axis length, GLCM entropy, Skewness and Kurtosis, the difference is higher than for the rest of the features. The highest diversity is found for Filled Area, Major axis length, Minor axis length and GLCM entropy. However, though the values of Mean, Median, Max and Min were closer in some cases, no exact similarity is found for any of the features. Therefore, it can be concluded that, for different iterations, different kernels are applied resulting in distinct feature maps.

A similar analysis is conducted for CT Scan dataset deriving the mean, median, max and min values of the geometrical features from class COVID. Table 18 illustrates the average mean, median, Max and Min value for the feature maps of the three iterations for Convolution 1 layer.

It is evident from Table 18 that for each feature there are differences among Mean, Median, Max and Min values for the different iterations. For some features, such as Area, Equivalent diameter, Convex Area, Filled Area, Major axis length, Standard Deviation, GLCM entropy, Skewness and Kurtosis, the difference is higher than for the rest of the features. However, though the values of Mean, Median, Max and Min were closer in some cases, no exact similarity is found for any of the features. Therefore, it can be concluded that, as for two different datasets, quite similar outcome is achieved, for different iterations, different kernels are applied resulting in distinct feature maps.

6.5. Analysis-5: iteration by iteration difference between the feature maps of the classes

While running a model multiple times, for each run different kernels are applied to produce diverse feature maps. A common question arises while conducting classification problem using a CNN, 'why is a different accuracy found every time a model is run?' According to previous studies, one thing that impacts on the classification performance is inter-class variance. Therefore, an analysis can be conducted on how much dissimilarity between the classes is produced for each iteration. For each iteration of layer 1, 2 and 3 the difference between the feature maps of Benign and Malignant are derived based on the F-value of ANOVA test, utilizing the 17 geometric features. In this case the 'black FM's are removed (Table 3), from the feature maps of both classes and the interpretation is based on the 'remaining FM'. The average F-value is also calculated for each iteration. A high F-value indicates that the inter-group difference is greater than intra-group difference which means that a statistically significant variance exists in the group means and the feature has a discriminative capability (Kim, 2014) (Ding et al., 2014). Table 19 shows the ANOVA results for two classes and three iterations and for each layer.

Table 19 shows that a substantial variance of two classes is found among the iterations even for the same layer. The higher the F-value, the larger the difference among the classes. Table 20 shows the min and max F values for all features along with their associated layer and iteration. Here, 'Layer (max)' and 'Iteration (max)' denote the layer and iteration for which the maximum F-value is found and similarly 'Layer (min)' and 'Iteration (min)' indicate the layer and iteration with the minimum F value.

Overall, it can be observed from Table 20 that there are significant differences between the max and min F-values and that for the majority of the features the highest F-value comes from layer 1 and iteration 3 and the lowest F-value from layer 2 and iteration 2. The objective of this

Table 16
Values of MSE, PSNR, SSIM, RMSE, DSC, UQI and SAM for CT Scan dataset.

FM	MSE	PSNR	SSIM	RMSE	DSC	UQI	SAM
org0-fm1	5943.03	10.39	0.17	0.99	0.47	0.06	1.43
org0-fm2	515.68	21.01	0.79	0.29	0.59	0.87	0.19
org0-fm3	5695.18	10.58	0.21	0.97	0.47	0.12	1.32
org0-fm4	5728.31	10.55	0.18	0.97	0.47	0.09	1.32
org0-fm5	5892.22	10.43	0.17	0.99	0.47	0.09	1.4
org0-fm6	6089.67	10.28	0.09	1	0.47	0.01	1.55
org0-fm7	5935.12	10.4	0.12	0.99	0.47	0.04	1.43
org0-fm8	918.47	18.5	0.6	0.39	0.53	0.84	0.39
org0-fm9	6058.22	10.31	0.13	1	0.47	0.01	1.52
org0-fm10	4387.65	11.71	0.3	0.85	0.48	0.27	0.92
org0-fm11	569.01	20.58	0.75	0.31	0.59	0.87	0.23
org0-fm12	3940.19	12.18	0.31	0.81	0.49	0.4	0.9
org1-fm1	3981.26	12.13	0.25	0.99	0.49	0.12	1.46
org1-fm2	4030.13	12.08	0.15	1	0.49	0.03	1.54
org1-fm3	226.73	24.58	0.79	0.24	0.67	0.87	0.17
org1-fm4	3896.03	12.22	0.32	0.98	0.49	0.23	1.37
org1-fm5	3857.24	12.27	0.34	0.98	0.49	0.32	1.36
org1-fm6	3977.57	12.13	0.28	0.99	0.49	0.23	1.44
org1-fm7	4034.36	12.07	0.16	1	0.49	0.02	1.54
org1-fm8	3989.53	12.12	0.28	0.99	0.49	0.16	1.46
org1-fm9	429.7	21.8	0.64	0.33	0.51	0.82	0.33
org1-fm10	4027.08	12.08	0.26	1	0.49	0.13	1.5
org1-fm11	2623.18	13.94	0.46	0.81	0.49	0.46	0.84
org1-fm12	326.9	22.99	0.75	0.28	0.61	0.86	0.24
org1-fm13	171.25	25.79	0.77	0.21	0.72	0.88	0.21
org2-fm1	408	22.02	0.8	0.31	0.61	0.85	0.2
org2-fm2	3695.65	12.45	0.43	0.94	0.49	0.33	1.23
org2-fm3	3705.73	12.44	0.43	0.94	0.49	0.3	1.22
org2-fm4	667.94	19.88	0.66	0.4	0.48	0.82	0.41
org2-fm5	4181.79	11.92	0.27	1	0.39	0.08	1.47
org2-fm6	4035.59	12.07	0.34	0.99	0.47	0.2	1.39
org2-fm7	194.21	25.25	0.8	0.22	0.78	0.87	0.19
org2-fm8	304.83	23.29	0.79	0.27	0.65	0.87	0.23
org2-fm9	3859.35	12.27	0.39	0.96	0.49	0.21	1.3
org2-fm10	822.05	18.98	0.67	0.44	0.51	0.79	0.42
org2-fm11	2825.84	13.62	0.5	0.82	0.50	0.47	0.97
org2-fm12	3826.28	12.3	0.43	0.96	0.49	0.33	1.26
org2-fm13	355.86	22.62	0.76	0.29	0.68	0.86	0.26
org2-fm14	187.92	25.39	0.81	0.21	0.75	0.88	0.21
org2-fm15	359.68	22.57	0.8	0.29	0.65	0.86	0.2
org2-fm16	3694.66	12.46	0.43	0.94	0.49	0.33	1.23

analysis is to find the layer and iteration number for which the highest F-value is generated for each feature. for example, if we look at the feature 'area' in Table 19, it can be seen that the highest F-value of 525.29 is found for layer 1, iteration 3 and the lowest F-value of 3.19 is found for layer 2, iteration 2. As a higher F-value indicates that more differences exist among the instances, it can be concluded for feature 'area' the highest difference among the feature maps is generated from layer 1, iteration 3 . Likewise, the lowest difference among the feature maps is generated from layer 2, iteration 2 as indicated by the F-value for this configuration. Observing carefully the columns 'Layer (max)' and 'Layer (min)' in Table 19, it can be found that 12 features out of 17 features yields the maximum F-value for layer 1 and 13 features out of 17 features yields the maximum F-value for iteration 3. The same applies to the remaining two columns, 'Iteration (max)' and 'Iteration (min)', leading to the conclusion that the largest difference among the feature maps can be found for layer 1, iteration 3 and the smallest difference for layer 2, iteration 2. The values of 'Layer (max)', 'Iteration (max)', 'Layer (min)' and 'Iteration (min)' are taken from Table 19 where the 'max' and 'min' indicates the highest and lowest F-value for each feature respectively. Furthermore, the highest mean F-value is found for layer 1 and iteration 3. For a better understanding and to show the pattern of the differences between the classes for different iterations and layers, a bar chart is generated using the mean F-value of the last row of Table 19, see Fig 17.

It can be seen from Fig 17 that for iteration 1, the mean F-value gradually increases over the layers whereas for iteration 3, the mean F-value decreases for layer 1 and 2 before dropping notably at layer 3. However, for iteration 2, a stable trend is found for the F-values across all the layers. Feature map generation by each iteration absolutely follows a random pattern. Due to the randomness, we do not achieve a similar performance from CNN over each time a model is trained newly. This experiment explains how diverse the feature maps can be based on each time a model is trained showing the F-values. The observation concludes that the feature maps can be significantly higher even for the same layer over different iteration. Finally, it is showed that for which layer and iteration, the highest F-value is determined. As the highest mean F-value is found from layer 1 and iteration 3 it can be said that for this configuration the highest dissimilarity is generated between the feature maps of the benign and malignant classes. Since the classification accuracy greatly depends on the inter-class variance, and based on the F-value the difference of the feature maps among the classes can be anticipated, the finding can provide a useful insight of the performance inconsistency of CNN across different iteration.

Similarly, for each iteration of layer 1, 2 and 3 the difference between the feature maps of COVID, non-COVID and CAP of CT Scan dataset are derived based on the F-value of ANOVA test, utilizing the 17 geometric features following the same process. Table 21 shows the ANOVA results for three classes and three iterations and for each layer.

Table 21 shows that a substantial variance of three classes is found among the iterations even for the same layer. Table 22 shows the min and max F values for all features along with their associated layer and iteration.

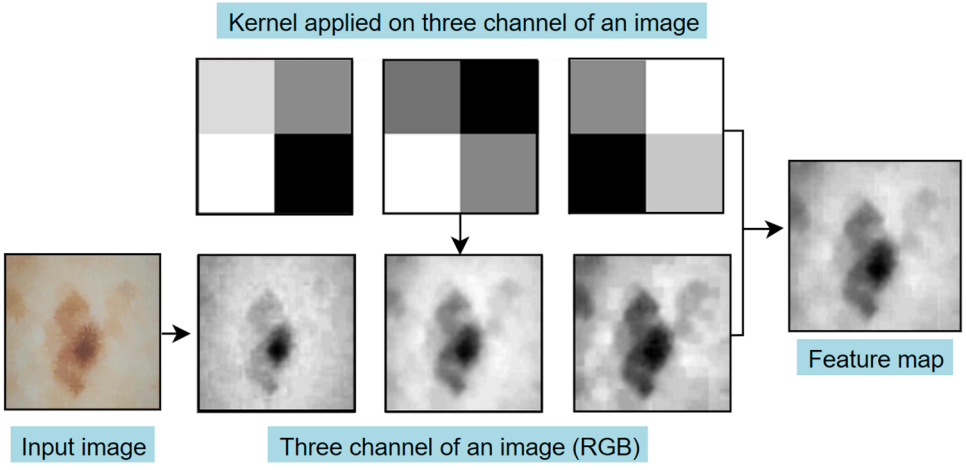


Fig 15. Process of generating feature map (iteration - 1).

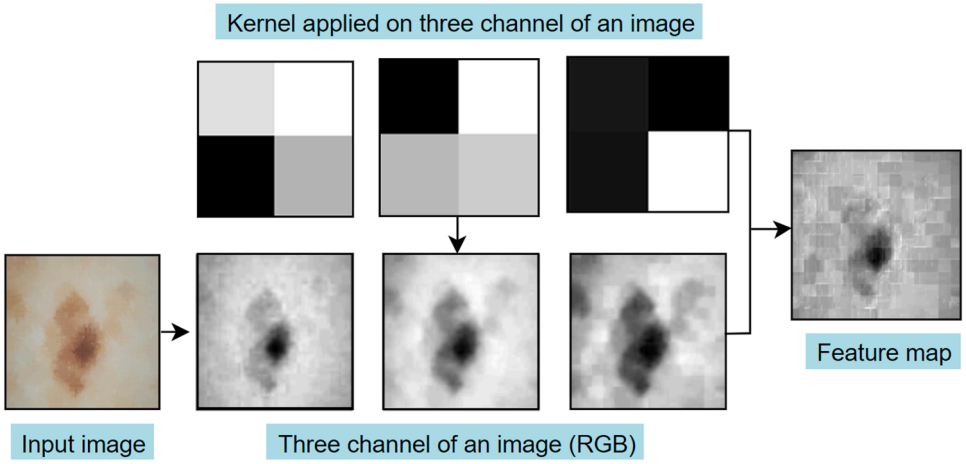


Fig 16. Process of generating feature map (iteration - 2).

Table 17
Mean, Median, Maximum and Minimum value for the feature maps.

Feature	Iteration 1				Iteration 2				Iteration 3			
	Mean	Median	Max	Min	Mean	Median	Max	Min	Mean	Median	Max	Min
Area	10,849.02	12,318	12,321	23	10,078.52	12,318	12,321	25	8249.93	12,315	12,321	26
Perimeter area ratio	0.08	0.06	0.35	0	0.07	0.06	0.32	0	0.07	0.05	0.30	0
Solidity	0.44	0.39	4.12	0.01	0.44	0.39	3.41	0.02	0.42	0.38	4.25	0.009
Equivalent diameter	114.58	125.23	125.25	5.41	107.88	125.23	125.25	5.64	91.45	125.21	125.25	5.75
Convex Area	26,771.72	29,211	36,804	33	23,976.31	27,516	36,732	12	19,374.35	21,741	36,561	15
Extent	0.03	0.02	0.35	0.001	0.03	0.02	0.5	0.004	0.03	0.02	0.41	0.001
Filled Area	896.41	819	3765	12	826.78	771	5259	9	649.81	609	7515	9
Major axis length	170.73	177.58	238.57	7.33	159.65	172.46	237.32	4.91	140.48	158.89	259.22	5.88
Minor axis length	99.98	104.73	157.39	3.82	92.53	99.23	151.98	3.37	79.49	86.58	160.42	3.31
Mean	121.68	136.84	229.19	0.03	119.56	141.41	230.55	0.03	105.73	134.95	233.91	0.03
Standard Deviation	40.65	39.79	101.89	2.13	39.22	39.02	84.43	2.29	38.49	41.38	91.06	2.21
Shannon entropy	6.11	6.66	7.88	0.02	5.67	6.63	7.88	0.02	4.76	6.48	7.84	0.02
GLCM entropy	123.37	135.14	157.11	0.49	113.57	133.59	157.79	0.56	93.42	128.15	155.18	0.59
Skewness	1.07	−0.78	106.59	−5.05	1.71	−0.780	105.41	−5.44	9.32	−0.45	106.59	−6.38
Kurtosis	112.45	1.63	11,651.6	−1.79	130.26	2.35	11,470.2	−1.70	736.84	3.66	11,651.60	−1.76
Lbp energy	0.23	0.15	0.99	0.13	0.28	0.14	0.99	0.13	0.40	0.16	0.99	0.13
Gabor energy	0.36	0.25	0.99	0.13	0.40	0.25	0.99	0.13	0.48	0.29	0.99	0.13

Table 18
Mean, Median, Maximum and Minimum value for the feature maps.

Feature	Iteration 1				Iteration 2				Iteration 3			
	Mean	Median	Max	Min	Mean	Median	Max	Min	Mean	Median	Max	Min
Area	8723.15	9910.50	12,290.0	391.00	6108.33	5599.00	12,280.00	26.00	7505.20	8275.00	12,288.00	31.00
Perimeter area ratio	0.08	0.08	0.30	0.00	0.05	0.04	0.26	0.00	0.07	0.07	0.29	0.00
Solidity	0.25	0.28	0.52	0.02	0.18	0.16	0.74	0.00	0.21	0.23	0.51	0.01
Equivalent diameter	102.08	112.33	125.09	22.31	79.91	84.43	125.04	5.75	91.80	102.65	125.08	6.28
Convex Area	34,827.41	35,985.0	36,963.0	8436.00	33,476.93	35,925.00	36,963.00	42.00	34,013.03	35,898.00	36,963.00	90.00
Extent	0.12	0.09	0.39	0.01	0.08	0.06	0.37	0.00	0.10	0.07	0.39	0.00
Filled Area	4302.50	3348.00	14,556.0	177.00	2812.92	2307.00	13,605.00	12.00	3514.79	2670.00	14,505.00	15.00
Major axis length	167.08	165.72	232.03	86.10	172.59	167.73	335.95	8.51	167.33	164.88	260.44	10.08
Minor axis length	3.65	3.65	3.65	3.65	3.65	3.65	3.65	2.30	3.65	3.65	3.65	3.65
Mean	25.01	15.42	86.82	0.41	14.64	5.31	86.47	0.04	19.49	11.51	84.11	0.05
Standard Deviation	30.40	27.92	71.40	5.16	20.63	16.14	71.69	2.21	26.14	24.00	64.90	2.34
Shannon entropy	4.48	4.93	7.07	0.36	3.21	3.01	7.04	0.03	3.91	4.20	7.12	0.03
GLCM entropy	95.67	104.67	153.82	7.56	69.70	67.75	154.01	0.60	84.42	93.32	153.79	0.68
Skewness	4.39	3.41	31.92	−0.28	12.61	6.52	105.88	−0.28	7.69	3.87	104.89	−0.24
Kurtosis	54.28	13.32	1217.39	−1.58	494.39	55.63	11,543.34	−1.62	230.11	17.49	11,392.09	−1.47
Lbp energy	0.35	0.28	0.95	0.15	0.51	0.46	1.00	0.15	0.43	0.32	1.00	0.15
Gabor energy	0.75	0.81	0.99	0.31	0.84	0.91	1.00	0.30	0.80	0.85	1.00	0.31

It can be observed from Table 22 that there are significant differences between the max and min F-values and that for the majority of the features the highest F-value comes from layer 1 and iteration 2 and the lowest F-value from layer 2 and iteration 1. Furthermore, the highest mean F-value is found for layer 2 and iteration 2. As the highest mean F-value is found from layer 2 and iteration 2 it can be said that for this configuration the highest dissimilarity is generated between the feature maps of the three classes. For a better understanding and to show the

Table 19
F-value comparing two classes for each layer and iteration.

Features	Conv 1 Iteration 1	Iteration 2	Iteration 3	Conv 2 Iteration 1	Iteration 2	Iteration 3	Conv 3 Iteration 1	Iteration 2	Iteration 3
Area	32.25	17.02	525.29	183.74	3.19	404.94	268.78	22.17	12.33
Perimeter area ratio	13.57	1.21	30.41	96.12	53.59	83.74	38.38	15.80	70.11
Solidity	2.36	1.41	14.14	2.63	46.78	227.93	0.36	11.45	20.08
Equivalent diameter	52.26	14.54	474.58	154.58	2.11	330.57	253.60	18.83	3.59
Convex Area	133.35	36.31	494.09	62.92	25.94	63.56	30.16	3.90	60.69
Extent	0.33	40.14	93.95	6.02	3.53	5.93	5.56	15.18	95.19
Filled Area	143.0	18.56	74.10	0.43	0.0003	10.18	563.78	71.58	7.99
Major axis length	78.29	48.83	436.09	52.46	12.36	59.24	102.41	34.29	2.78
Minor axis length	56.92	80.94	497.40	131.05	36.93	28.45	176.02	20.97	19.75
Mean	17.46	9.86	217.31	126.09	21.07	408.67	210.22	19.71	24.30
Standard Deviation	0.003	9.66	2.02	44.16	160.30	642.72	265.57	106.16	0.34
Shannon Entropy	10.96	44.27	480.80	134.88	4.04	389.55	83.48	49.60	5.54
GLCM entropy	24.48	46.39	583.51	187.90	0.13	326.31	154.32	18.44	9.86
Skewness	68.58	12.01	142.85	36.008	15.42	85.05	268.78	22.17	12.33
Kurtosis	47.54	13.91	93.17	10.36	13.27	15.78	38.38	15.80	70.11
Lbp energy	46.89	15.95	502.15	88.23	0.02	334.71	0.36	11.45	20.08
Gabor energy	0.56	33.05	365.68	203.30	0.45	300.61	242.33	48.32	37.33
Mean F-value	42.87	26.12	295.74	89.46	23.48	218.70	158.97	29.75	27.79

Table 20
Max and min F-values of the features.

Features	Max	Min	Layer (max)	Iteration (max)	Layer (min)	Iteration (min)
Area	525.29	3.19	1	3	2	2
Perimeter area ratio	96.12	1.21	2	1	1	2
Solidity	227.93	0.36	2	3	3	1
Equivalent diameter	330.57	2.11	2	3	2	2
Convex Area	494.09	3.90	1	3	3	2
Extent	95.19	0.33	3	3	1	1
Filled Area	563.78	0.0003	3	1	2	2
Major axis length	436.09	2.78	1	3	3	3
Minor axis length	497.40	19.75	1	3	3	3
Mean	408.67	9.86	1	2	2	3
Standard Deviation	642.72	0.003	1	1	2	3
Shannon Entropy	480.80	4.04	1	3	2	2
GLCM entropy	583.51	0.13	1	3	2	2
Skewness	142.85	12.01	1	3	1	2
Kurtosis	93.17	10.36	1	3	2	1
Lbp energy	502.15	0.02	1	3	2	2
Gabor energy	365.68	0.45	1	3	2	2
Mean F-value	295.74	23.48	1	3	2	2

pattern of the differences between the classes for different iterations and layers, a bar chart is generated using the mean F-value of the last row of Table 21, see Fig 18.

It can be seen from Fig 18 that for iteration 2 and 3, the mean F-value gradually decreases over the layers whereas for iteration 3, the mean F-value decreases from layer 1 to 2 before rising notably at layer 3. Hence, it is anticipated that as the highest mean F-value is found for layer 3 and iteration 3, for this case the highest dissimilarity among the classes might be existed.

It is observed from Table 19 (skin cancer dermoscopy dataset) and Table 21 (lung scan CT scan dataset) that though across each iteration and layer notable difference is found among the classes based on the F-value, there is no regularity. For skin cancer dataset, the highest mean F-value is found for layer 1 and iteration 3 and lung scan dataset the highest mean F-value is found for layer 3 and iteration 3. If the model is trained again, the highest mean F-value may be achieved for another

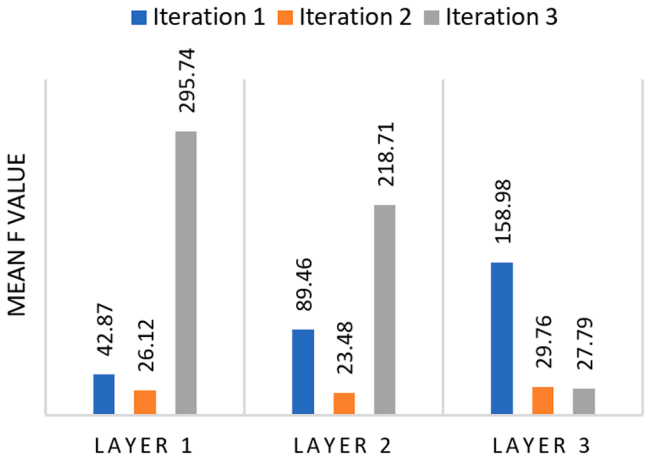


Fig 17. Bar plot of the mean F-values for skin cancer dataset.

layer and iteration. A possible cause of this difference might be, the well-known characteristics of CNN that each time a model is trained, different types of kernels are assigned resulting in dissimilar feature maps across layers and iterations. The differences between the iterations provide an insight into why a single model can perform with different prediction rate for same dataset when the model is trained again. As mentioned previously, the higher the F-value, the greater the difference among the classes and inter-class differences have a major impact on the final prediction. The inter-class difference varies among the feature maps of the classes for different layers and iterations. To summarize, analyzing the F-values for different iterations shows that there is a noticeable diversity between the feature maps of the classes which might lead to different accuracies for different iterations.

6.6. Feature map and overall findings

The objective of this study is to show the difference of feature maps based on five aspects using several statistical approaches. To evaluate this objective more rigorously and precisely, two different datasets are utilized to observe how the findings are co-related across diverse modalities. Five analyses are discussed with feature map visualization and rigorous statistical experiments.

In analysis 1, the appearance of the feature maps are investigated with. According to the findings, both datasets show that 'black FMs' are produced; however, the number is not consistent. The number of 'black

Table 21
F-value between three classes for each layer and iteration.

Features	Conv 1 Iteration 1	Iteration 2	Iteration 3	Conv 2 Iteration 1	Iteration 2	Iteration 3	Conv 3 Iteration 1	Iteration 2	Iteration 3
Area	179.86	412.06	305.86	28.67	197.11	68.66	396.13	331.59	90.60
Perimeter area ratio	66.12	812.68	159.75	10.71	190.92	97.03	312.76	285.31	170.78
Solidity	164.19	420.86	318.87	130.64	167.04	39.47	549.91	209.95	64.03
Equivalent diameter	224.47	378.32	231.81	32.00	206.14	90.83	366.32	321.75	65.06
Convex Area	210.09	51.48	4.36	42.48	511.62	218.13	867.42	64.31	110.90
Extent	271.43	362.83	336.42	257.86	142.00	89.11	48.17	79.24	58.20
Filled Area	266.61	342.09	345.49	298.63	273.27	149.27	609.56	19.76	38.54
Major axis length	204.38	656.00	478.18	212.8	707.09	514.77	56.90	66.41	101.02
Minor axis length	59.92	327.94	297.40	131.05	86.93	228.45	31.57	61.27	103.70
Mean	82.93	394.05	397.69	5.32	697.37	80.19	544.55	470.30	205.98
Standard Deviation	31.27	472.80	293.17	7.07	635.65	257.65	116.26	419.32	565.34
Shannon Entropy	158.36	448.98	337.35	29.99	313.51	96.21	359.91	321.00	105.03
GLCM entropy	171.55	437.40	320.16	39.00	248.46	96.65	321.05	246.93	64.66
Skewness	169.61	195.23	34.40	60.26	297.42	187.31	229.67	304.47	217.90
Kurtosis	86.76	46.07	34.99	72.52	179.22	157.61	215.92	220.38	204.42
Lbp energy	288.45	401.70	280.70	36.74	202.61	85.34	369.04	331.76	68.52
Gabor energy	40.60	271.45	279.55	10.12	480.60	151.79	359.89	179.91	195.32
Mean F-value	157.44	378.34	262.12	82.69	325.70	153.43	338.53	231.39	142.94

Table 22
Max and min F-values of the features.

Features	Max	Min	Layer (max)	Iteration (max)	Layer (min)	Iteration (min)
Area	412.06	28.67	1	2	2	1
Perimeter area ratio	812.68	10.71	2	2	2	1
Solidity	549.91	39.47	3	1	2	3
Equivalent diameter	378.32	32.00	1	2	2	1
Convex Area	867.42	4.36	3	1	3	2
Extent	362.83	48.17	1	2	3	1
Filled Area	609.56	19.76	3	1	3	2
Major axis length	656.00	56.90	1	2	3	1
Minor axis length	327.94	31.57	1	2	3	1
Mean	544.55	5.32	3	1	2	1
Standard Deviation	635.65	7.07	2	2	2	1
Shannon Entropy	448.98	29.99	1	2	2	1
GLCM entropy	437.40	39.00	1	2	2	1
Skewness	304.47	34.40	3	2	1	3
Kurtosis	220.38	34.99	3	2	1	3
Lbp energy	401.70	36.74	1	2	2	1
Gabor energy	480.60	10.12	2	2	2	1
Mean F-value	325.70	82.69	2	2	2	1

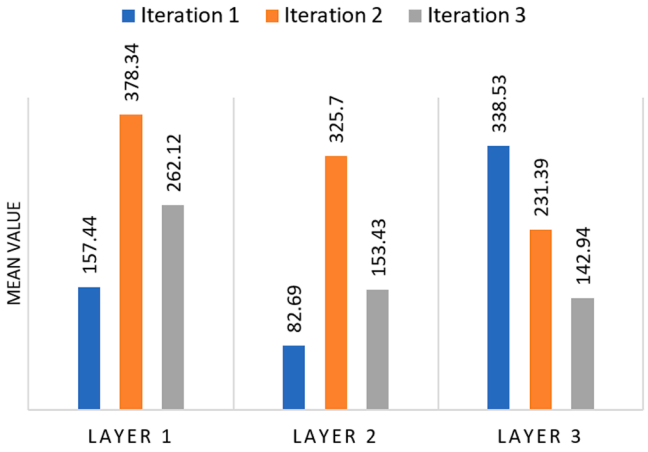


Fig 18. Bar plot of the mean F-values for CT Scan dataset.

FMs' varies depending on different imaging modalities, convolutional layers and iterations. The kernels applied by convolutional layer tend to look for different features to generate feature maps, but these kernels sometimes do not find the targeted feature in an image, resulting in a 'black FM'. The kernel structure changes over multiple iterations and layers, causing an inconsistency in the proportion of 'black FMs' across different iterations and layers.

In analysis 2, difference of the feature maps is explored through statistical tests using the feature maps of layer by layer. Visually, it is observed that different layer produces feature maps of diverse characteristic. However, to what extent this diversity occurs is shown through the statistical analyses. Based on the statistical findings, it is observed that there is a noticeable dissimilarity in both geometrical and intensity-based features of layer by layer feature maps. The T-test results show that the feature maps from first, second and third convolutional layers are different. The ANOVA test shows that the feature maps are different across three convolutional layers. The conclusion comes from the null hypothesis of these two tests as null hypothesis 'reject' indicates to the difference among the groups. These findings provide an insight that the kernels applied by different convolutional layers generate diverse feature maps and this diversity, in terms of geometrical and photometrical properties, is quite high. This validates that the kernels are designed such a way that they can extract meaningful information from an image by altering its geometric and photometric properties.

In analysis 3, it is shown that the convolutional layers produce high diversity of feature maps, not only for feature maps of the different layers but also within the feature maps of a single image. For example, the feature maps for a single image, generated by different layers, are different. If 16 kernels are applied to an image, the resulting 16 feature maps are different to one another and also to the original input image. Though it is a well-known that different kernels produce different feature maps, the degree of dissimilarity was unknown. To discover the degree of diversity, statistical values including MSE, PSNR, SSIM, RMSE, DSC, UQI and SAM are generated. These statistical experiments are used to evaluate the image dissimilarity. Through the statistical analysis, we have quantified the degree of dissimilarity, which may be useful knowledge regarding the 'black box'.

In analysis-4, our investigations show that while training the model multiple times, the characteristics of the feature maps change significantly, as evaluated based on 17 geometric features. Across all three iterations, for the 17 geometric features, average Mean, Median, Max

and Min values are showcased. It is found that Mean, Median, Max and Min values of the 17 geometrical features vary notably over different iteration.

Finally, in analysis-5, the cause of differences in performance of CNNs across multiple trainings is investigated. The results show that for different iterations and layers, the inter-class diversity is different. The insight is determined based on the F-value of ANOVA test. The ANOVA test is conducted across all iterations for each convolutional layer respectively using. In this regard, numeric features extracted from the feature maps were used. As inter-class diversity impacts on the classification performance, this may be a potential reason why different accuracies are achieved when the model is trained multiple times.

7. Comparison with existing literatures

In demystifying the 'black box' nature of CNNs, different studies have focused different aspects. Though the main objective was similar, the methodologies and findings were quite different. Table 23 summarizes the objectives, methodology and findings of these studies and presents a comparison with our research.

Several studies with similar objectives have been conducted investigating different aspects of the 'black box', see table 22. Some researchers worked on the visualization of feature maps, saliency maps, and activation maps to explain the final prediction of neural networks. Other studies experimented with different state-of-the-art deep learning the models to present a comparison of their learning schemes. The research is not limited into image data, but also includes natural language and time-series data. In some studies, the classification performance of the network was improved through understanding the inner mechanism of the CNN. However, though feature maps are visualized in several papers, to the best of our knowledge, no study can be found using quantitative and statistical interpretation of feature map diversity, layer by layer and over multiple iterations. Hence, the strategy and findings of this study can present some new insights in the CNN 'black box' mechanism.

8. Discussion and future work

This study presents a comprehensive insight of CNN's shifting characteristics over different scenarios. The experiments are conducted across two datasets and findings are discussed, which can assist the computer science researchers in generating robust CNN architecture with improved accuracy for various applications. Five research questions relating the convolutional layers, kernel and feature maps are answered through feature map visualization and broad statistical analysis. As the behavior of CNN is learnt through extensive experiments, it will help in the practical implications of designing and training a CNN model. From these analyses, it is found that kernels have a notable influence on producing the feature maps which can impact on the classification/detection performance significantly. Therefore, the findings of this study can help in determining the number and size of kernels. The change of feature maps occurs highly in both geometrical and photometrical features. Therefore, it is crucial that the ROI of input images should not be affected to use as input data in order to obtaining better performance. Two medical datasets are used having different characteristics, and results suggest that the characteristic of feature maps depend on the dataset or the ROI. This is another knowledge discovery that can assist the computer scientists in developing optimal CNN model for different medical datasets. As the convolutional layers produce feature maps which are diverse from input image and among them, each feature map contains different and meaningful information which might be the possible reason of getting optimal accuracy from deep architecture. This finding can help in the real applications to develop the architecture with suitable layer structures. Moreover, from the knowledge of intra-class differences, users will be benefitted learning the potential reason of different performance of CNN over different training period.

Table 23
Comparison with previous studies.

No.	Paper	Objective	Key methodology	Related finding
1	(Dablain & Jacobson, 2021)	Explaining 'black box' from the perspective of how CNNs interpret imbalanced image data showing feature properties, relevance and diversity.	Two cost-sensitive algorithms applied on two datasets.	Answers to seven research question related to how CNN performs with imbalanced dataset on minority and majority classes.
2	(Ferdinand & Mercier, n.d.)	Explaining the 'black box' of neural network's decision strategy by statistical feature extracted from time series natural language data.	Development of Time-Series eXplanation (TSXplain) system using different synthetic and anomaly identification datasets.	The strategy taken by neural networks to provide a correct final decision.
3	(Ramaneswaran et al., 2021)	Classification of acute lymphoblastic leukemia and understanding the inner mechanism of Inception v3 model.	Using hybrid of Inception v3 and XGBoost model.	The feature map only pays attention to the areas that contribute to accurate classification.
4	(Zhao et al., 2022)	Explaining the CNN 'black box' through an image segmentation network.	Gradient-based activation mapping technique, visualization feature map of different layers.	Explaining the CNN segmentation model, based on multiscale features of different layers.
5	(Brahimi et al., 2018)	Comparison of different CNN models in image classification, visualization of internal mechanisms.	Saliency map visualization scheme, six shallow and deep CNN models for plant disease classification.	The saliency maps are able to identify the affected regions automatically.
6	(Dependent et al., 2021)	Classification of breast cancer, visualization of the learned features.	VGG-19 deep learning model, handcrafted features, attention maps.	Deep learning-based features are more impactful than machine learning approaches.
7	(Heinrich et al., 2019)	Interpreting the 'black box' scheme of deep neural networks.	Feature visualization, activation map, forwards and backward propagation.	Solution of two research questions.
8	(Lange et al., 2018)	Understanding neural network's 'black box' by developing a system that can present the mechanism of the layers and decision of deep neural network in human understandable from	VGG19, feature visualization, saliency map, grad-CAM representation.	Discussion of the layer by layer mechanism and output (feature map). CNN extract only relevant features in final prediction.

(continued on next page)

Table 23 (*continued*)

No.	Paper	Objective	Key methodology	Related finding
9	Our research	Understanding and presenting CNN 'black box' theory through layer by layer feature map analysis in a statistical and quantitative way.	Feature map visualization, handcrafted feature extraction, statistical tests using two medical datasets.	Discussion of five research questions related to feature map patterns and diversity, kernels and decision of neural network.

This study explores only the impact of convolutional layer and its components including kernels and feature maps. The other layers of CNN such as batch normalization layer, dropout layer, dense layer and fully connected layer can be some useful aspects to study in future. The impacts of hyper-parameters such as activation function, optimizer, learning rate can also be experimented with. The gradient descent, cost function and back propagation are some inner mechanism of CNN which we aim to research in prospective study. In the extended version of this 'Black box' study, a classification approach can be introduced based on the feature maps extracted from each layer and iteration. The features extracted from the feature maps can be employed for classifying the images using machine learning (ML) algorithms and several feature selection and ensemble techniques can be utilized to acquire the highest possible accuracy. In a further study, the number of images will be kept as low as possible. Moreover, we might attempt to eliminate the max-pool layers while extracting feature maps to discover whether an improvement of performance is found or not. This may help to address the computational complexity and the scarcity of medical images.

9. Conclusion

This study attempts to discover obscure and internal characteristics of convolutional layers through a statistical interpretation of feature maps from layer to layer using two datasets of diverse modality. First, the proportion of 'black FM' for both datasets (benign for skin cancer and COVID for CT Scan) is compared to all the feature maps for each layer and iteration. 17 geometric features and 6 intensity-based features are extracted from the feature maps after the removal of 'black FMs'. Statistical analysis is conducted for these features. The approach of this research reveals a number of insights such as characteristics of feature maps, differences among the feature maps from layer to layer and also with the original image, differences among several iterations, differences between the feature maps of the classes based on different iteration and more. More complex mechanisms of CNN may be uncovered which would be a noteworthy advancement in study of the CNN 'Black box'.

CRediT authorship contribution statement

Sami Azam: Conceptualization, Supervision, Writing – review & editing, Project administration. **Sidratul Montaha:** Conceptualization, Validation, Data curation, Methodology, Writing – original draft, Writing – review & editing. **Kayes Uddin Fahim:** Visualization, Data curation, Writing – original draft. **A.K.M. Rakibul Haque Rafid:** Software, Writing – original draft, Writing – review & editing, Visualization, Validation. **Md. Saddam Hossain Mukta:** Supervision, Validation. **Mirjam Jonkman:** Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Link to the open access datasets has been shared in the manuscript.

References

Aimar, A., Mostafa, H., Calabrese, E., Rios-Navarro, A., Tapiador-Morales, R., Lungu, I.A. et al. (2019). NullHop: A Flexible Convolutional Neural Network Accelerator Based on Sparse Representations of Feature Maps. IEEE Transactions on Neural Networks and Learning Systems, 30(3), 644–656. 10.1109/TNNLS.2018.2852335.

Alaa, A.M., & Van Der Schaar, M. (n.d.). Demystifying Black-box Models with Symbolic Metamodels.

Aslahishahri, M., Stanley, K. G., Duddu, H., Shirtliffe, S., Vail, S., Bett, K., et al. (2021). From RGB to NIR: Predicting of near infrared reflectance from visible spectrum aerial images of crops. In *Proceedings of the IEEE International Conference on Computer Vision, 2021-Octob* (pp. 1312–1322). https://doi.org/10.1109/ICCVW54120.2021.00152

Brahimi, M., Arsenovic, M., Laraba, S., Sladojevic, S., Boukhalfa, K., & Moussaoui, A. (2018). *Deep learning for plant diseases : Detection and saliency map visualisation*. Springer International Publishing. https://doi.org/10.1007/978-3-319-90403-0

Buhrmester, V., Münch, D., & Arens, M. (2021). Analysis of explainers of black box deep neural networks for computer vision: A survey. *Machine Learning and Knowledge Extraction, 3*(4), 966–989. https://doi.org/10.3390/make3040048

Chaves, R., Ramírez, J., Górriz, J. M., López, M., Salas-Gonzalez, D., Álvarez, I., et al. (2009). SVM-based computer-aided diagnosis of the Alzheimer's disease using t-test NMSE feature selection with feature correlation weighting. *Neuroscience Letters, 461*(3), 293–297. https://doi.org/10.1016/j.neulet.2009.06.052

Dablain, D., & Jacobson, K.N. (2021). arXiv : 2210 . 09465v1 [cs . CV] 17 Oct 2022 Understanding CNN Fragility When Learning With Imbalanced Data.

Dependent, M., Breast, H., & Explanation, V. (2021). Conventional Machine Learning versus Deep Learning for Magnification Dependent Histopathological Breast Cancer Image Classification : A Comparative Study with Visual Explanation.

Di Leo, G., & Sardanelli, F. (2020). Statistical significance: P value, 0.05 threshold, and applications to radiomics—Reasons for a conservative approach. *European Radiology Experimental, 4*(1). https://doi.org/10.1186/s41747-020-0145-y

Ding, H., Feng, P. M., Chen, W., & Lin, H. (2014). Identification of bacteriophage virion proteins by the ANOVA feature selection and analysis. *Molecular BioSystems, 10*(8), 2229–2235. https://doi.org/10.1039/c4mb00316k

Fatema, K., Montaha, S., Rony, Md. A. H., Azam, S., Hasan, Md. Z., & Jonkman, M. (2022). A Robust framework combining image processing and deep learning hybrid model to classify cardiovascular diseases using a limited number of paper-based complex ECG images. *Biomedicines, 10*(11), 2835. https://doi.org/10.3390/biomedicines10112835

Ferdinand, K., & Mercier, D. (n.d.). TSXplain : Demystification of DNN Decisions for Time-Series using Natural Language and Statistical Features.

Giuste, F., Shi, W., Zhu, Y., Naren, T., Isgut, M., Sha, Y., et al. (2022). Explainable artificial intelligence methods in combating pandemics: A systematic review. *IEEE Reviews in Biomedical Engineering*. https://doi.org/10.1109/RBME.2022.3185953

Heinrich, K., Zschech, P., Skouti, T., Griebenow, J., & Riechert, S. (2019). Demystifying the black box: A classification scheme for interpretation and visualization of deep intelligent systems. In *AMCIS 2019 Proceedings*. https://aisel.aisnet.org/amcis2019/ai_semantic_for_intelligent_info_systems/ai_semantic_for_intelligent_info_systems/8.

He, S., Tavakoli, H.R., Borji, A., Mi, Y., & Pugeault, N. (2019). Understanding and Visualizing Deep Visual Saliency Models. http://arxiv.org/abs/1903.02501.

Jin, D. S. C., & Chen, J. C. (2013). Low-dose graphic-processing-unit based limited-angle CT reconstruction algorithm development for a home-designed dual modality micro-FT/CT system. In *Proceedings of the IEEE Nuclear Science Symposium Conference Record*. https://doi.org/10.1109/NSSMIC.2013.6829232

Khan, M. A., Rubab, S., Kashif, A., Sharif, M. I., Muhammad, N., Shah, J. H., et al. (2020). Lungs cancer classification from CT images: An integrated design of contrast based classical features fusion and selection. *Pattern Recognition Letters, 129*, 77–85. https://doi.org/10.1016/j.patrec.2019.11.014

Kim, H.-. Y. (2014). Analysis of variance (ANOVA) comparing means of more than two groups. *Restorative Dentistry & Endodontics, 39*(1), 74. https://doi.org/10.5395/rde.2014.39.1.74

Lam, C. F. D., Leung, K. S., Heng, P. A., Lim, C. E. D., & Shun Wong, F. W. (2012). Chinese acupuncture expert system (CAES)-a useful tool to practice and learn medical acupuncture. *Journal of Medical Systems, 36*(3), 1883–1890. https://doi.org/10.1007/s10916-010-9647-0

Lange, T.De, Hicks, S.A., Riegler, M., Pogorelov, K., Anonsen, K.V., Johansen, D. et al.. (2018). Dissecting Deep Neural Networks for Better Medical Image Classification and Classification Understanding. 10.1109/CBMS.2018.00070.

Large COVID-19 CT scan slice dataset | Kaggle. (n.d.). Retrieved November 22, 2022, from https://www.kaggle.com/datasets/maedemaftouni/large-covid19-ct-slice-dataset.

Lin, X., Lei, Y., Chen, J., Xing, Z., Yang, T., Wang, Q. et al. (n.d.). A Case-Finding Clinical Decision Support System to Identify Subjects with Chronic Obstructive Pulmonary Disease Based on Public Health Data.

Li, S., Zhao, X., Stankovic, L., & Mandic, D. (2022). *Demystifying CNNs for Images by Matched Filters*, 1–10.

Lundberg, S.M., Allen, P.G., & Lee, S.-.I. (n.d.). A Unified Approach to Interpreting Model Predictions. https://github.com/slundberg/shap.

Mahendran, A., & Vedaldi, A. (2016). Visualizing deep convolutional neural networks using natural pre-images. *International Journal of Computer Vision, 120*(3), 233–255. https://doi.org/10.1007/s11263-016-0911-8

Mohammed, A.F., Hashim, S.M., & Jebur, I.K. (2022). The diagnosis of COVID-19 in CT images using hybrid machine learning approaches (CNN & SVM). 10(2), 376–387.

Montaha, S., Azam, S., Kalam, A., Rakibul, M., Rafid, H., Ghosh, P. et al.. (2021). BreastNet18 : A high accuracy fine-tuned VGG16 model evaluated using ablation study for diagnosing breast cancer from enhanced mammography images.

Montaha, S., Azam, S., Rakibul Haque Rafid, A. K. M., Islam, S., Ghosh, P., & Jonkman, M. (2022). A shallow deep learning approach to classify skin cancer using down-scaling method to minimize time and space complexity. *PloS one, 17*(8), Article e0269826. https://doi.org/10.1371/JOURNAL.PONE.0269826

Muriel-Vizcaíno, R., Treviño-Garza, G., Murata, C., Staines-Boone, A.T., Yamazaki-Nakashimada, M.A., Espinosa-Padilla, S.E. et al. (2017). En relación con el artículo: Calidad de vida de los pacientes con inmunodeficiencias primarias de anticuerpos. Acta Pediatrica de Mexico, 38(2), 134–138. 10.3389/fpsyg.2015.00223.

Park, Y., & Yang, H. S. (2019). Convolutional neural network based on an extreme learning machine for image classification. *Neurocomputing, 339*, 66–76. https://doi.org/10.1016/j.neucom.2018.12.080

Patil, A., & Rane, M. (2021). Convolutional Neural Networks: An Overview and Its Applications in Pattern Recognition. Smart Innovation, Systems and Technologies, 195, 21–30. 10.1007/978-981-15-7078-0_3.

Prijs, J., Liao, Z., To, M.-. S., Verjans, J., Jutte, P. C., Stirler, V., et al. (2022). Development and external validation of automated detection, classification, and localization of ankle fractures: Inside the black box of a convolutional neural network (CNN). *European Journal of Trauma and Emergency Surgery*. , Article 0123456789. https://doi.org/10.1007/s00068-022-02136-1

Qin, Z., Yu, F., Liu, C., & Chen, X. (2018). How convolutional neural networks see the world — A survey of convolutional neural network visualization methods. *Mathematical Foundations of Computing, 1*(2), 149–180. https://doi.org/10.3934/mfc.2018008

Rafid, A. K. M. R. H., Azam, S., Montaha, S., Karim, A., Fahim, K. U., & Hasan, Md. Z. (2022). An effective ensemble machine learning approach to classify breast cancer based on feature selection and lesion segmentation using preprocessed mammograms. *Biology, 11*(11), 1654. https://doi.org/10.3390/biology11111654

Ramaneswaran, S., Srinivasan, K., & Vincent, P.M.D.R. (2021). Review Article Hybrid Inception v3 XGBoost Model for Acute Lymphoblastic Leukemia Classification. 2021.

Rudd-Orthner, R.N.M., & Mihaylova, L. (2019). Non-Random Weight Initialisation in Deep Learning Networks for Repeatable Determinism. Conference Proceedings of 2019 10th International Conference on Dependable Systems, Services and Technologies, DESSERT 2019, 1, 223–230. 10.1109/DESSERT.2019.8770007.

Sampathila, N., Pavithra, & Martis, R. J (2022). Computational approach for content-based image retrieval of K-similar images from brain MR image database. *Expert Systems, 39*(7), 1–12. https://doi.org/10.1111/exsy.12652

Sankur, B. (2002). Statistical evaluation of image quality measures. *Journal of Electronic Imaging, 11*(2), 206. https://doi.org/10.1117/1.1455011

Sarıgül, M., Ozyildirim, B. M., & Avci, M. (2019). Differential convolutional neural network. *Neural Networks, 116*, 279–287. https://doi.org/10.1016/j.neunet.2019.04.025

Simonyan, K., Vedaldi, A., & Zisserman, A. (2013). Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. http://arxiv.org/abs/1312.6034.

Skin Cancer: Malignant vs. Benign | Kaggle. (n.d.). Retrieved January 3, 2023, from https://www.kaggle.com/datasets/fanconic/skin-cancer-malignant-vs-benign.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2016-Decem, 2818–2826*. https://doi.org/10.1109/CVPR.2016.308

Vaghjiani, D., Saha, S., Connan, Y., Frost, S., & Kanagasingam, Y. (2020, November 29). Visualizing and understanding inherent image features in CNN-based glaucoma detection. *2020 Digital image computing: Techniques and applications, dicta 2020*. https://doi.org/10.1109/DICTA51227.2020.9363369

Wang, B., Ma, R., Kuang, J., & Zhang, Y. (2020). How decisions are made in brains: Unpack "Black Box" of CNN with Ms. Pac-Man Video Game. *IEEE access : practical innovations, open solutions, 8*, 142446–142458. https://doi.org/10.1109/ACCESS.2020.3013645

Wani, N., & Raza, K. (2018). Multiple kernel-learning approach for medical image analysis. *Soft computing based medical image analysis* (pp. 31–47). Elsevier Inc. https://doi.org/10.1016/B978-0-12-813087-2.00002-6

Wei, D., Zhou, B., Torrabla, A., & Freeman, W. (2015). Understanding intra-class knowledge inside CNN. http://arxiv.org/abs/1507.02379.

Xu, M., Fralick, D., Zheng, J.Z., Wang, B., Tu, X.M., & Feng, C. (2017). The differences and similarities between two-sample t-test and paired t-test. Shanghai Archives of Psychiatry, 29(3), 184–188. 10.11919/j.issn.1002-0829.217070.

Yanase, J., & Triantaphyllou, E. (2019). The seven key challenges for the future of computer-aided diagnosis in medicine. In International journal of medical informatics (Vol. 129, pp. 413–422). Elsevier Ireland Ltd. 10.1016/j.ijmedinf.2019.06.017.

Zeiler, M.D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. Lecture notes in computer science (Including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics), 8689 lncs(part 1), 818–833. 10.1007/978-3-319-10590-1_53.

Zhang, Q., Yang, Y., Ma, H., & Wu, Y. N. (2019). Interpreting CNNS via decision trees. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2019-June* (pp. 6254–6263). https://doi.org/10.1109/CVPR.2019.00642

Zhang, Z., Zhou, X., Zhang, X., Wang, L., & Wang, P. (2018). *A model based on convolutional neural network for online transaction fraud detection*. Security and Communication Networks. https://doi.org/10.1155/2018/5680264, 2018.

Zhao, M., Xin, J., Wang, Z., Wang, X., & Wang, Z. (2022). Interpretable Model Based on Pyramid Scene Parsing Features for Brain Tumor MRI Image Segmentation. 2022.