Explaining facial skin diseases classification using LIME

Table of Contents

1. Introduction:

In recent years, the prevalence of skin diseases and associated treatment have become a burden on the global healthcare system. The average life expectancy and cost related to the treatment of skin conditions can be an economic threat to some healthcare systems [1]. According to the latest report published by the Allied Market Research, the global dermatological market is categorized into acne, alopecia, dermatitis, psoriasis, Rosacea, skin cancer, and others [2]. As estimated, these skin diseases will be the most prevalent in 2030.

In this study, we are looking at one of the aforementioned skin conditions i.e. Rosacea, to provide a computer-aided early diagnosis. Rosacea is an undesirably neglected skin condition in computer-aided clinical diagnosis/decision-making systems. Rosacea is a chronic skin condition that goes through fading and relapse. Hence, it requires early diagnosis and regular consultation with dermatologists. Generally, people in the age group above 30 are affected by Rosacea.

To support the early diagnosis of Rosacea, there are only a few existing studies that discuss the computer vision and machine learning techniques for Rosacea classification. Usually, these studies rely on deep learning models for classification. In computer vision, the accuracy of a clinical diagnosis system is often calculated based on the classification of its input dataset, i.e. images. To perform this classification, the algorithms are designed in such a way that they identify particular features of interest. Items could be a piece of the segmented region of interest in the input image or a set of pixels, or a group of features in N-dimensional

space [3]. Because most medical images are complex, versatile and error-prone, it is hard to choose the features of interest in the images while performing a classification problem. Although these studies conducted on the classification of Rosacea or any skin diseases in general present some reasonable level of accuracy in classifying, the reasons behind such numerical values of accuracy are unspecified due to the Blackbox nature of deep learning models. The concept of Explainable AI (XAI) may help understand the causes behind the obtained accuracy of these classification models. The concept of explanation (Explainability) is a gateway between AI and the practical world. It is a powerful tool for detecting flaws in the models and biases in the data to verify the predictions, improve the model, and gain new insights into the problem[4].

There are various explanation patterns discussed in the AI literature to make computer-aided diagnosis viable. It is studied that the technique of explainability varies on the type of input dataset. In this work, Rosacea images are the input data. Hence the possible explainability methods that apply to images are considered. In this work, the main focus is the local explainability due to the nature of the skin disease dataset, which is discussed in the next section. There are various techniques available for explaining the medical images[5][6][7].

This study looks at the Rosacea classification that includes additional explanations using Local Interpretable Model-agnostic Explanations (LIME) [8]. Several advantages that influenced the explainability using the LIME method are:

- Among many applications of XAI techniques, LIME can be applied to high dimensional domains, i.e. image data.
- LIME provides local explainability, which means it gives the root cause behind the individual predictions that the model makes. This feature is indispensable as it helps reverse engineering and helps understand the model functionality.
- LIME trains an inherently interpretable model on a new dataset made from the permutation of samples and the corresponding prediction of the black box.
- LIME also shows which feature contributes to the decision making and by how much. It allows replacing the underlying “black box” model by keeping the same local interpretable model for the explanation [7].
- LIME uses “superpixels” for the images that are more likely to correspond to semantically different parts of an image. At the same time, occlusion perturbs image patches in a systematic, uniform way, ignoring possible semantic similarity between adjacent pixels [9].
- A predicted output image can be explained/interpreted by randomly changing individual pixels of the image. LIME uses one of the image segmentation methods, which helps segment the predicted output image into superpixels. Superpixels are the combination of pixels with similar colours. These superpixels are turned on and off according to the specified probability in each selection set.

However, there are some limitations of LIME discussed as follows:

- As LIME is an approximation model, and the local model might not cover the complete attribution due to the generalization, it might be unfit for cases where we legally need thorough explanations of a decision.
- There is no consensus on the neighbourhood boundary for the local model; sometimes, it provides very different explanations for two nearby data points [7].
- If the sampling process is repeated, the outcome of the explanations can be different and unstable for long run applications. This instability may eventually lead to doubtful events and critical analysis of the explanations.

1.1. Why LIME can be a suitable model for our problem compared to other model-agnostic techniques:

Other model-agnostic techniques share the same characteristics with LIME, such as Gradient methods and Shapley values. Gradient-based methods [10] offer local explanation vectors as local gradients of the probability function of the trained model for the positive class. The probability function is learned from the input examples that belong to the explanation vector for the classified test point in the local gradients. The partial derivates of a function give the main functionality of the method at a given input data/example. The number of output features of such gradient-based methods is very high, which leads to the high-dimensionality problem. Shapley values help enhance interpretability/explainability by computing the important values for each feature for individual predictions. As per the comparative analysis by Lundberg et al.[11] Shapley values provide many output features that can be computationally costly and offer very complex explanations. In contrast, LIME is more specific towards the individual images' local feature points, giving smaller outputs than input. This functionality reduces the computational time and cost and is easily understandable [12].

This study is focused on the skin disease diagnosis and explanation (or interpretability). Even though most skin features are similar in humans, the local characteristics of skin diseases are unique. Indeed, the features of skin diseases such as shape, size, colour, area, lesion height and depth, erupted vessels, affected tissues etc., vary from person to person, and so do they from one photograph to another. These kinds of variations can be standardized as local variations. In this study, a few specific features such as colour and area of disease are taken into consideration. When skin diseases are diagnosed by a dermatologist, the affected region, such as a particular locality, is considered for the diagnosis/ examination. Hence, in this study, we will look at each image and the model explanation for the specific images, given the classification results.

2. Related work on LIME using Skin diseases images:

In this section, we will discuss some previous work conducted on skin disease image classification using deep learning algorithms and explainability using LIME.

Stieler et al. [13] used the HAM10,000 skin image dataset LIME as an explainer for skin image classifiers. The dataset consists of 10,015 dermoscopic images of 7 types of skin cancer lesions. Only two classes i.e. nevus (nv) and melanoma (mel) are included for the experimentation. The transfer learning approach with a pre-trained MobileNet model was used for the classification task. Along with LIME as an explainer, the ABCD rule of dermatological diagnosis is considered to support the interpretability. To make interpretability medically viable Stieler et al. [13] suggests that some necessary dermatological feature extraction methods should be considered, along with the explainability methods of AI. This way the model can learn to distinguish between medically relevant and medically irrelevant features for diagnosis.

Xiang et al. [14] have conducted a classification task on the HAM10,000 dataset. Only three classes are considered for the classification task, summing it 8900 samples in total. The CNN models that are used for classification are VGG16 [15], DenseNet [16], Xception [17], and Inception-ResNet v2 [18]. Further, the classification results by Inception-ResNet v2 are demonstrated using LIME. The results show that the significantly affected regions, i.e. cancerous lesions, are highlighted through ‘positive boundary’ and ‘positive regions’ on the classified images. Appropriately highlighted regions validate that the cancerous lesion features are taken as important features for the classification.

Meske et al.[19] have discussed the explainability of a CNN model with malaria detection. A VGG-16 and VGG-19 [20]architectures are used for classification. In this study, LIME highlights the affected and unaffected regions of malaria. Although LIME could

successfully detect the affected and unaffected regions, there is a downside for some instances: not all noticeable regions were highlighted.

Albeit Explainable AI is a young field, there are still so many unknown areas to catch up on. It is palpable that very few studies are conducted on LIME applying to skin disease image diagnosis.

As motivation from previous studies, we consider some deep learning classification models such as Inception v3 as base model for classifying facial skin conditions and then apply LIME as a surrogate model for explaining the predicted output. The skin diseases classification is mainly focused on skin cancer lesions. So far, most of the work carried out using LIME uses dermoscopic skin lesion images. In this study, we may explore LIME on a novel and limited dataset of rosacea called "rosacea full face-300 (rff-300)". It is evident from the name that the 'modality' of the images is different from the cancerous skin lesion images. A detailed explanation of the dataset will be provided in the Methodology section.

3. Importance of Superpixels with LIME:

LIME relies on the segmentation of the image into superpixels. A superpixel carries more information than individual pixels [21]. The corresponding pixels of an image, with similar properties such as colours and spatial attributes, are grouped into a more significant segment of pixels called superpixels, an outcome of image over-segmentation. The random perturbations of superpixels are important to identify the relevant superpixels that were important in decision making for certain label or class. As LIME is meant for exploring local explanations from the input data and model, superpixels supports in interpretation by highlighting the local features of an image. The main advantages of Superpixels in LIME are:

(i) Lower complexity- Superpixels approach reduces the complexity of the image due to the small number of entities and use less processing power.

(ii) Significant Entities – Gives articulated meaning to the pixel group with the same texture and colour distribution.

(iii) Marginal information loss – Although important information is highlighted, some insignificant regions also walk in during the over-segmentation process. It is advantageous compared to other segmentation techniques as only minor insignificant areas come into the case. [22].

One of the superpixels methods is the **Quick-shift algorithm** [23], which segments RGB channel images by identifying clusters of pixels in the joint colour and spatial dimensions. Quick shift has four advantages: (i) simplicity; (ii) speed; (iii) generality; (iv) a tuning parameter to trade off under and over-fragmentation[23]. The Quick-shift algorithm supplements the advantages of Superpixels Explanation. The first step of LIME, by default, uses the Quickshift algorithm for a given image *I* to generate superpixels.

4. Methodology:

In this section, the workflow of LIME, the choice of datasets and classification models are discussed.

4.1. LIME workflow:

The key idea of LIME is to split the input image I into superpixels that create new examples out of the input image. LIME helps explain a Blackbox model, i.e., a deep learning classification model. For image classification tasks, LIME helps reveal the region of an image with the set of superpixels with a firm reference with a prediction label. The workflow of LIME for the Image classification model is as follows:

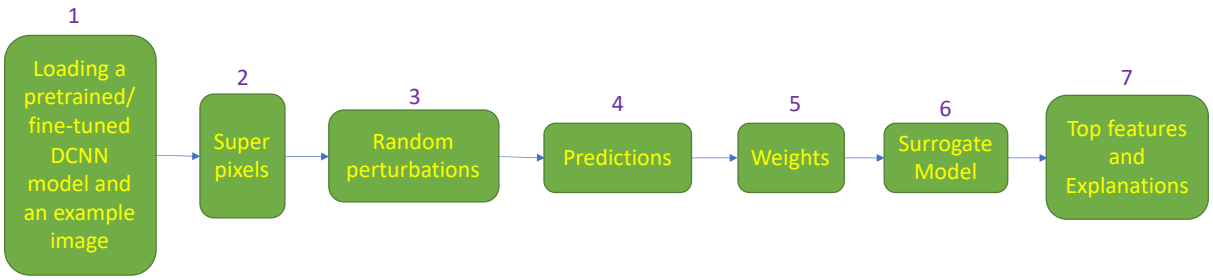

Figure 1: An illustration of LIME workflow.

(i) **A pretrained/fine-tuned DCNN model and an example image** are loaded to the LIME explanations pipeline.

(ii) **Superpixels:** LIME uses the Quickshift algorithm for the input image I to produce superpixels over 5-dimensional space. The obtained number of superpixels is calculated and mapped on the image.

(iii) **Random Perturbations**: For the same I, some numbers of perturbations are created. Perturbations are the generated from random sampling of the superpixels and from the desired number of perturbations. Hence the outcome of the perturbations indicates the superpixels that are on and off (on and off are represented by the binary numbers 1 means on and 0 means off). A bigger given number of perturbations leads to greater the reliability of explanations with more interpretable features.

(iv) **Predictions:** Furthermore, the prediction for each generated perturbation is calculated against the rest of the classes/labels in the trained classification model.

(v) **Weights:** The cosine distance between each randomly generated perturbation and I is calculated to help the explanation further. As a distance metric is associated with each induced perturbation, a kernel width value (between 0 and 1) can be set based on expected distance values. An important note is that the weights depend only on the number of inactivated superpixels.

(vi) **Surrogate Model:** Based on the previously calculated interpretable features (perturbations, predictions, and weights), a weighted linear regression model is fitted, and interpretable coefficients are computed. Each coefficient in the linear model corresponds to each superpixel in the segmented image. These coefficients represent the degree of prediction of given classes. These coefficients are presented by an array, in which each column corresponds to a given class.

(vii) **Top features and Explanation**: To find the top features that can contribute to explanation, it is necessary to sort (by using a sorting algorithm) the interpretable coefficients to figure out which superpixels have a larger impact on the prediction of a given label. The choice for the number of top features should be less than the number of obtained superpixels. This step would allow for highlighting only the top features of a given image I and masking out other details that did not help map the predictions.

4.2. Datasets and Data preparation:

Rosacea-Full-Face-300 (rff-300): The "rff-300" is a collection of images from various sources such as Irish Dataset (confidential) [24] , SD-260 [25], teledermatology websites and google image search results.

Flicker Faces High Quality (FFHQ): FFHQ [26] is a collection full-face images of people with normal skin.

Hence these two datasets were considered for training a binary classification model where one class belongs to normal faces (FFHQ), and another belongs to rosacea faces (rff-300).

4.3. Classification:

The pre-trained Inception v3 [27] is used as a base model to perform the classification task. Further, this inception v3 is fine-tuned with rff-300 and FFHQ datasets. The datasets are resized to 299x299 and split in the ratio of 70:20:10 for training, validation, and test, to fine-tune the Inception v3 model. *The final test accuracy of the model is 92.5%.* The training and validation loss and accuracy are depicted in Figures 2. As can be observed in figure 2, the model loss increases after 8 epochs; and the accuracy declines after 7 epochs due to the limited/ short data supply. *Hence it is vital to study the performance of LIME when a DCNN classification model is trained with small number of sample images.*

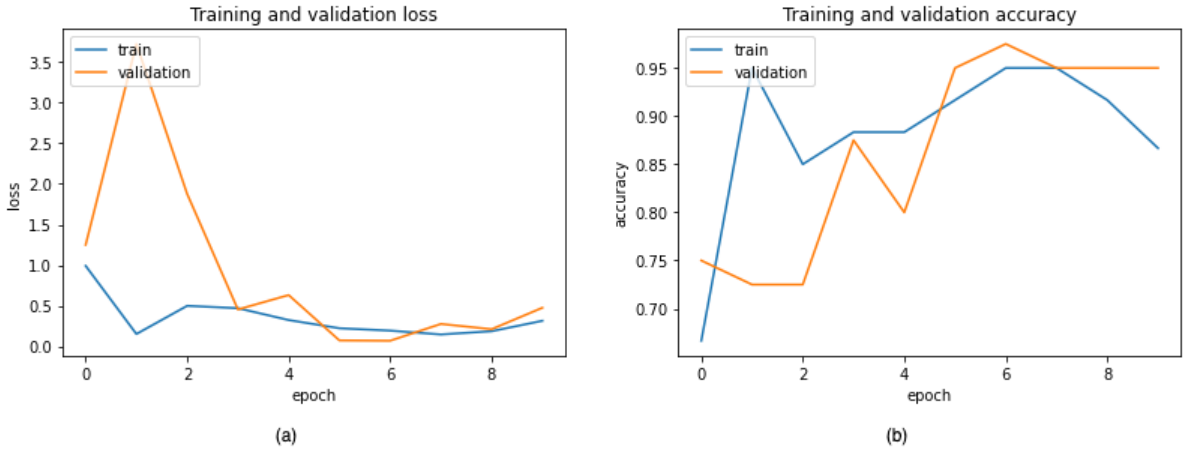

Figure 2

4.4. LIME workflow1 :

According to the LIME workflow discussed previously, the fine-tuned Inception v3 model was loaded with the LIME pipeline for generating explanation. Figure 3 is an output from the LIME pipeline for a Rosacea face. Fig. 3(a) is an example image from the rff-300 dataset, Fig. 3(b) is an illustration of **73 superpixels** found in the image, and Fig. 3(c) is an illustration of random perturbations. In Fig. 3(d), the top features that were part of the training and the predicted label 'Rosacea' are visible, and the masking hides the irrelevant parts; relevant top features are the superpixels that are turned on, and the irrelevant features are the ones where the superpixels are turned off by masking the regions. Given the output, a few observations are:

1. Most rosacea regions are exposed as top features, and healthy skins are masked.
2. However, part of the healthy skin on the chin, left side of the forehead and nose are still exposed. This means LIME could not detect and mask a few blocks of superpixels for normal skin.
3. As it is comprehensible that the idea of superpixels is to group and segment the image according to the colour and spatial attributes of the pixels; some parts of the skin could have been misinterpreted due to regions of the skin disease (rosacea) are subtle that it has the similar colour and spatial attribute as the normal skin. This is one of the anticipated drawbacks of LIME.

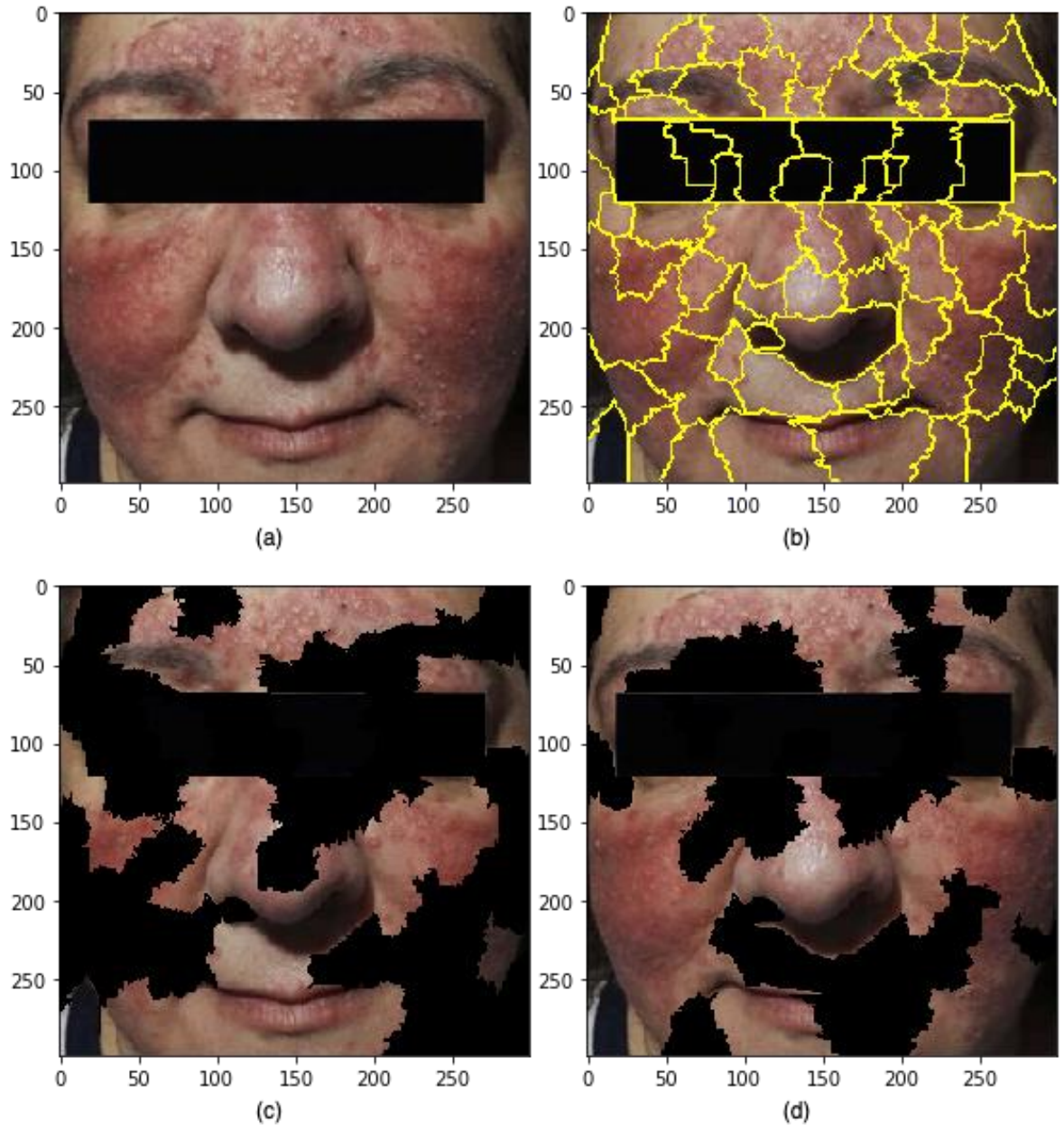

Figure 3

4.5. Pre-processing task and classification:

To alleviate the chances of miscalculation of relevant and irrelevant superpixels, a pre-processing with the input images was necessary to investigate. As the typical nature of Rosacea is the redness of the face, increasing the redness through contrast enhancement can help control the miscalculation to a certain extent. So, as a pre-processing, the contrast of each image in "rff-300" was increased by *label 2* to maintain the standardization. However, no changes were made to the normal face images of the FFHQ dataset.

Another dataset is prepared with contrast-enhanced images of rff-300. Inception v3 model is fine-tuned with the contrast-enhanced datasets. The noted final test accuracy was 94.99 % (~95%).

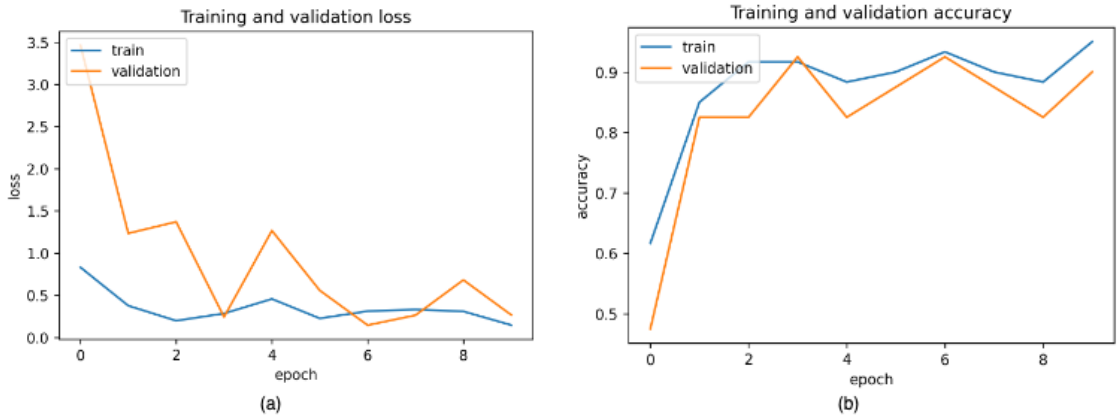

Figure 4

Based on the Fig. 4 graphs and test results, a few observations are made in comparison to the previous section:

(i) The final test accuracy was improved by 2.5%, which is notable.

(ii) Another notable improvement is, in the experiment with contrast enhancement, on and after 8 epochs, the model loss and accuracy are declining and inclining, respectively, in contrast to the previous experiment without contrast enhancement.

(iii) Although the volume of the input dataset is small, it is noticeable that the preprocessing task (through contrast enhancement) has essentially improved the model training and the accuracy.

4.6. LIME workflow2 (applying LIME to contrast enhanced image):

Figure 5 is the illustration of outcomes with LIME. Compared to the previous explanation by LIME, a few notable differences can be listed as follows:

(i) In contrast to the previous experiment, there are 82 superpixels obtained for the same image as shown in Fig 5(b), which is an increase of 9 more superpixels due to contrast enhancement.

(ii) In contrast to the previous results in figure 3(d), 5(d) illustrates that the coverage/masking of irrelevant features such as normal skin on the chin region of the face is masked more accurately than the previous one.

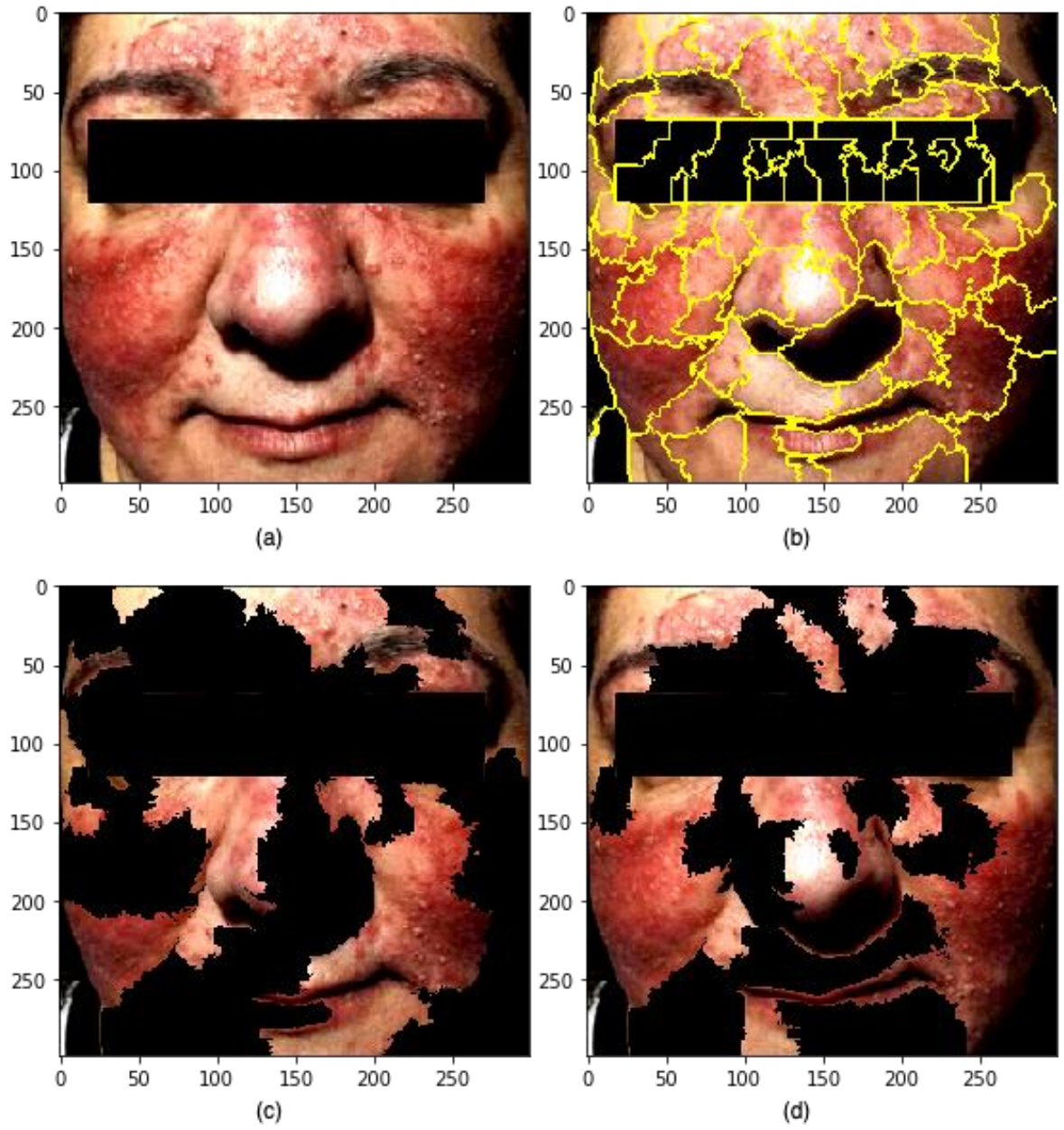

Figure 5

4.7. LIME workflow with normal face:

Understandably, Rosacea features can be associated with a colour, so the right features are picked up. As LIME effectively detects the relevant features for rosacea, it is important to test the efficacy of LIME for normal faces. Hence a few points observed in fig. 6 are:

(i) There is a total of 54 superpixels obtained from the given image I (fig 5(b)).

(ii) In fig.5(d), all the irrelevant features such as the image's background and most of the hair are masked. In contrast, most parts of the face are displayed as the top features. Hence it is distinctive that normal skin is taken as a crucial feature.

(iii) Conversely, the rosacea-affected skin is the top feature in Rosacea images. Hence it is evident that the binary classification model Inception v3 has learnt two important features for two given labels, i.e. Normal faces and Rosacea faces.

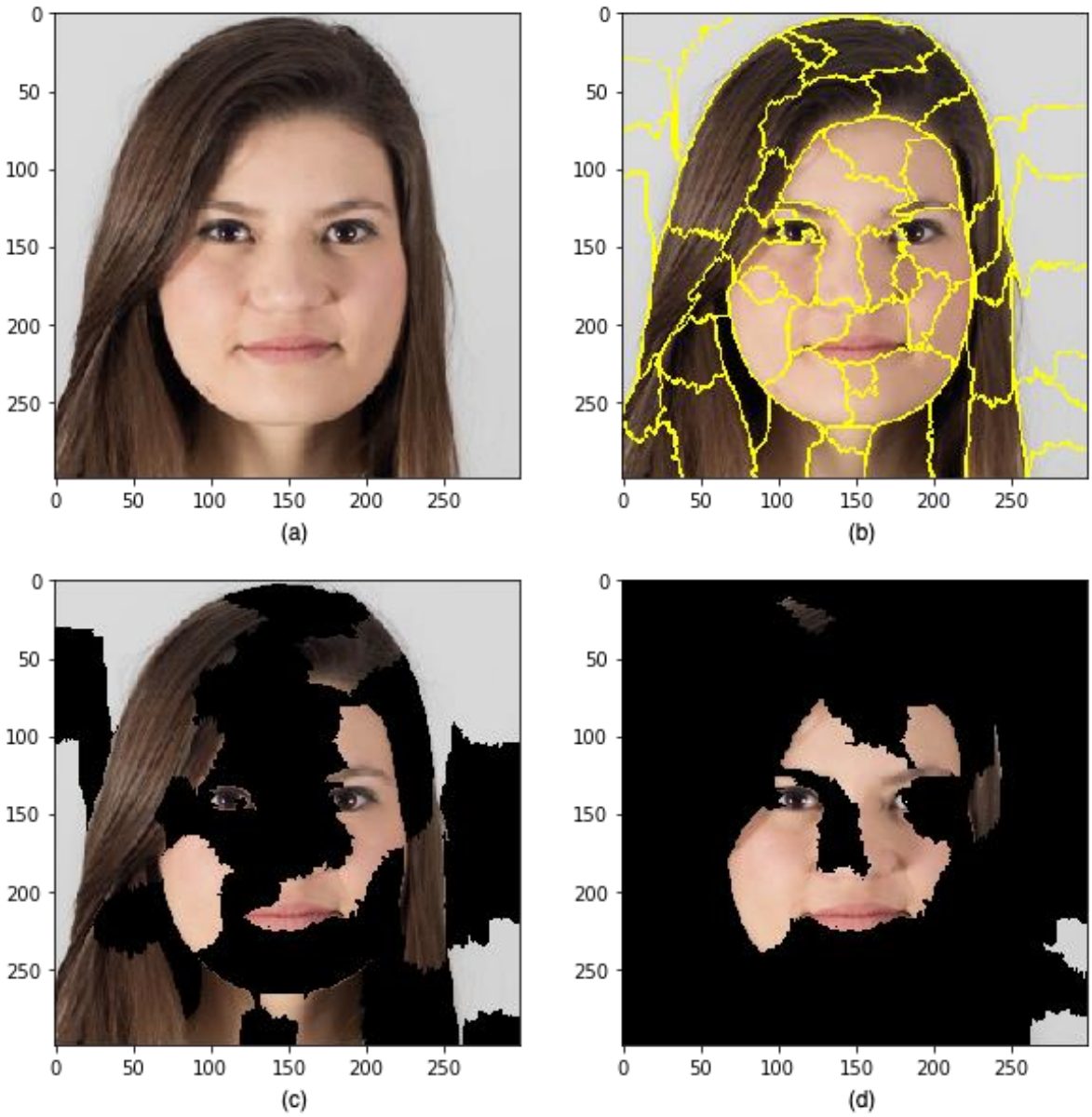

Figure 6

5. Discussion and Futurework:

5.1. Are superpixels suitable for medical imaging?

As per the discussed related work, experiments, and results, it is observable that superpixels can be dependent on the imaging domain. A few points need to be taken care of when applying LIME for explanation in medical imaging.

(i) Most of the work on LIME is implemented on CIFAR10 [28], ImageNet [29] and other real-world datasets. In this case, LIME is applied to medical datasets such as skin diseases, a considerable shift in the domain.

(ii) These experiments helped explain a black box model such as Inception v3 by differentiating between two crucial features.

(iii) The image pre-processing (through contrast enhancement) improved the LIME explanation by increasing the number of superpixels for the given input image.

(iv) The full face images from the FFHQ dataset had some background information. It is observable that faces have been taken as top features while backgrounds have been taken as irrelevant features. However, it is not the case for every image on the dataset. Hence it is arguable that Inception v3 is not entirely successful in picking up complex features (such as various colour channels) from a small dataset, but only a few sets of spatial/colour features.

(v) It can be concluded that LIME can be used in medical imaging domain explanation. Still, it may require domain/ imaging modality-specific pre-processing tasks to enhance the quality of explanation by improving the distinctiveness of the features that may help pick the right number of superpixels.

(vi) Superpixels cannot be represented as gradients in this application. Among various existing methods of superpixels, LIME uses the Quick-shift algorithm for implementing superpixels in the explanation pipeline. Quickshift does not support gradient representation in this scenario as the superpixels are hugely irregular in shapes and are not coherent with the neighbouring superpixels. The simple linear iterative clustering (SLIC) superpixel method [30] works efficiently with gradient representation.

(vii) As the LIME explanation pipeline uses the Quickshift algorithm for generating superpixels, it does not allow for precise control over the size or number of superpixels. This practice requires several non-intuitive parameters to be tuned and does not offer control over the amount or compactness of superpixels [30]. This is a notable limitation in gradient representation for the obtained superpixels. However, a few studies [30] [22] claim that the SLIC superpixel method performs much better. The SLIC has advantages over Quickshift due to showing a better correspondence between superpixels and relevant areas, which should be integrated with the LIME pipeline instead of the Quick-shift algorithm for generating superpixels. The colour similarity is the primary influence for segmentation in the Quickshift, while the clustering is the primary influence in segmentation in SLIC. Although the application of superpixels via LIME is demonstrated expected results for rosacea, the choice of Superpixels methods should be flexible for datasets from

various medical domains. For example, instead of Quickshift, SLIC can be incorporated with LIME for the explanation.

5.2.Fidelity of Explanation:

Although the intention of incorporating explanation is to understand the decision-making process/prediction of black box models, it is essential to analyse and understand the explanation provided by the integrated explanation pipeline/methodology. Recently, Fidelity Evaluation is one of the approaches to verify the XAI algorithms and assess the potential of the explanation methodology. In the XAI domain, fidelity is an open problem and refers to the quality of explanation [31]. As mentioned in the original work on LIME[8], there is interpretability vs fidelity trade-off. In any medical imaging domain that comprises a small dataset for training, decision making and explainability, the fidelity can be measured based on various criteria such as:

(i) Classification is performed on the small volume of dataset.
(ii) What features are learned during the training and classification.
(iii) Images that are involved in generating explanation.

5.3.Is the explanation by LIME viable with a small dataset and other medical imaging domain:

In most cases, it is unlikely to get a higher training accuracy from the DCNN model with a small number of images. Albeit only 600 images are used for training and classification purposes, the initial experiments and explanation outcomes are obtained. As discussed in the previous section, the workflow of explanations is useful by picking up the rosacea and normal features of the facial images. These functionalities were achievable based on the input data with the specific modality (full front face images) on the Inception V3 model trained. It is understood that LIME explains the images based on the similar colour attributes of the images. Theoretically, the explanations were picked up on the colour feature of the faces rather than the spatial features of the skin disease. Although it could explain the simple features picked up through binary classification, it is not suitable for clinical application yet. The LIME pipeline cannot be standardized across all medical imaging domains. But it can be optimized to work well for individual domains. The implementation of LIME can be varied in the range of different medical imaging modalities such as dermoscopic, MRI, CT, X-ray, histopathological imaging etc.

Based on the obtained results from this work, further implementation and improvement in LIME workflow can convey the trustworthiness of medical imaging classification and decision making. Given the simple nature of the explanation structure of LIME, it can be helpful for local explanations in early diagnosis and decision-making situations. However, numerous gaps are to be bridged.

Some open questions:

(i) Does the model accuracy affect explainability?
(ii) Are limited/small number of medical images suitable for achieving explanability?
(iii) Can explainable pipeline be fooled by generative adversarial attack? [32]

6. Conclusion:

This study discusses the application of LIME with a small dataset of Rosacea skin conditions. In part of this study, the understanding of the LIME algorithm was explored through various experiments, and some observations were made. A binary classification model was trained on the normal and Rosacea faces to generate the LIME explanation for Rosacea faces. Secondly,

the fine-tuned model was integrated into the LIME pipeline to generate explanations based on the crucial features on which predictions were made in the classification model. Hence the experimentations helped in understanding the features the classification model took. Furthermore, a few limitations were listed. Based on a few limitations observed, some future directions are recommended.

References:

[1] C. Flohr and R. Hay, "Putting the burden of skin diseases on the global map," *Br. J. Dermatol.*, vol. 184, no. 2, pp. 189–190, 2021.

[2] O. S. Linu Dash, "Dermatologicals Market by Disease (Acne, Dermatitis, Psoriasis, Skin Cancer, Rosacea, Alopecia, and Others), Type (Prescription-based Drugs, and Over-the-Counter Drugs), and Route of Administration (Topical Administration, Oral Administration, and Parente," 2022. [Online]. Available: https://www.alliedmarketresearch.com/dermatological-drugs-market#:~:text=The global dermatologicals market size,11.5%25 from 2021 to 2030. [Accessed: 10-May-2022].

[3] D. Gunning, "Explainable Artificial Intelligence (XAI)," 2017.

[4] W. Samek, T. Wiegand, and K. R. Müller, "Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models," *arXiv*, 2017.

[5] A. Singh, S. Sengupta, and V. Lakshminarayanan, "Explainable deep learning models in medical image analysis," *J. Imaging*, vol. 6, no. 6, pp. 1–18, 2020.

[6] E. Tjoa and C. Guan, "A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 32, no. 11, pp. 4793–4813, 2021.

[7] S. R. Islam, W. Eberle, S. K. Ghafoor, and M. Ahmed, "Explainable Artificial Intelligence Approaches: A Survey," pp. 1–14, 2021.

[8] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should i trust you?' Explaining the predictions of any classifier," *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, vol. 13-17-Augu, pp. 1135–1144, 2016.

[9] D. T. Huff, A. J. Weisman, and R. Jeraj, "Interpretation and visualization techniques for deep learning models in medical imaging," *Phys. Med. Biol.*, vol. 66, no. 4, 2021.

[10] D. Baehrens, T. Schroeter, S. Harmeling, M. Kawanabe, K. Hansen, and K. R. Müller, "How to explain individual classification decisions," *J. Mach. Learn. Res.*, vol. 11, pp. 1803–1831, 2010.

[11] S. M. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions," *Adv. Neural Inf. Process. Syst.*, vol. 2017-Decem, no. Section 2, pp. 4766–4775, 2017.

[12] D. Garreau and U. von Luxburg, "Explaining the Explainer: A First Theoretical Analysis of LIME," no. c, 2020.

[13] F. Stieler, F. Rabe, and B. Bauer, "Towards Domain-Specific Explainable AI: Model Interpretation of a Skin Image Classifier using a Human Approach," pp. 1802–1809, 2021.

[14] A. Xiang and F. Wang, "Towards Interpretable Skin Lesion Classification with Deep Learning Models," *AMIA ... Annu. Symp. proceedings. AMIA Symp.*, vol. 2019, pp. 1246–1255, 2019.

[15] A. Kwasigroch, A. Mikołajczyk, and M. Grochowski, "Deep neural networks approach

to skin lesions classification - A comparative analysis," *2017 22nd Int. Conf. Methods Model. Autom. Robot. MMAR 2017*, pp. 1069–1074, 2017.
[16] G. Huang, Z. Liu, and K. Q. Weinberger, "Densely Connected Convolutional Networks," *CoRR*, vol. abs/1608.0, 2016.
[17] F. Chollet, "Xception: Deep Learning with Depthwise Separable Convolutions," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1800–1807.
[18] C. Szegedy, S. Ioffe, and V. Vanhoucke, "Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning," *CoRR*, vol. abs/1602.0, 2016.
[19] C. Meske and E. Bunde, "Transparency and trust in human-AI-interaction: The role of model-agnostic explanations in computer vision-based decision support," in *International Conference on Human-Computer Interaction*, 2020, pp. 54–69.
[20] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc.*, pp. 1–14, 2015.
[21] P. Neubert and P. Protzel, "Superpixel benchmark and comparison," in *Proc. Forum Bildverarbeitung*, 2012, vol. 6, pp. 1–12.
[22] L. Schallner, J. Rabold, O. Scholz, and U. Schmid, "Effect of superpixel aggregation on explanations in LIME–a case study with biological data," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 2019, pp. 147–158.
[23] A. Vedaldi and S. Soatto, "Quick Shift and Kernel Methods for Mode Seeking," in *Computer Vision -- ECCV 2008*, 2008, pp. 705–718.
[24] F. C. Powell, "The Powell Lab." [Online]. Available: https://www.ucd.ie/charles/research/researchgroups/thepowelllab/.
[25] K. W. Xiaoxiao Sun, Jufeng Yang, Ming Sun, "SD-128,198,260 Recognition of Clinical Skin Disease Images."
[26] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of stylegan," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 8107–8116, 2020.
[27] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2016-Decem, pp. 2818–2826, 2016.
[28] "Cifar10 dataset."
[29] O. Russakovsky *et al.*, "ImageNet Large Scale Visual Recognition Challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, 2015.
[30] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "SLIC superpixels compared to state-of-the-art superpixel methods," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2274–2282, 2012.
[31] M. Velmurugan, C. Ouyang, C. Moreira, and R. Sindhgatta, "Developing a Fidelity Evaluation Approach for Interpretable Machine Learning," *arXiv Prepr. arXiv2106.08492*, 2021.
[32] J. Heo, S. Joo, and T. Moon, "Fooling neural network interpretations via adversarial model manipulation," *Adv. Neural Inf. Process. Syst.*, vol. 32, 2019.