J Inf Process Syst, Vol.20, No.4, pp.558~573, August 2024
https://doi.org/10.3745/JIPS.04.0318
ISSN 2092-805X (Electronic)

JOURNAL OF INFORMATION PROCESSING SYSTEMS JIPS

An Explainable Deep Learning-Based Classification Method for Facial Image Quality Assessment

Kuldeep Gurjar[1], Surjeet Kumar[2], Arnav Bhavsar[3], Kotiba Hamad[2], Yang-Sae Moon[1], and Dae Ho Yoon[2,*]

Abstract

Considering factors such as illumination, camera quality variations, and background-specific variations, identifying a face using a smartphone-based facial image capture application is challenging. Face Image Quality Assessment refers to the process of taking a face image as input and producing some form of "quality" estimate as an output. Typically, quality assessment techniques use deep learning methods to categorize images. The models used in deep learning are shown as black boxes. This raises the question of the trustworthiness of the models. Several explainability techniques have gained importance in building this trust. Explainability techniques provide visual evidence of the active regions within an image on which the deep learning model makes a prediction. Here, we developed a technique for reliable prediction of facial images before medical analysis and security operations. A combination of gradient-weighted class activation mapping and local interpretable model-agnostic explanations were used to explain the model. This approach has been implemented in the preselection of facial images for skin feature extraction, which is important in critical medical science applications. We demonstrate that the use of combined explanations provides better visual explanations for the model, where both the saliency map and perturbation-based explainability techniques verify predictions.

Keywords

Explainable Deep Learning, Face Image Quality Assessment, Image Classification, MobileNet, Transfer Learning

1. Introduction

Image classification, one of the most fundamental areas of computer vision, is the process of categorizing image data into one of many specified classes. This is the basis for other computer vision tasks such as localization, detection, and segmentation. Categorization of face images is a typical application for image classification, and the identification of applicable or non-applicable image capture conditions is critical in facial recognition applications, such as security checks and skin analysis. In addition, in medical image capture applications such as ultraviolet (UV) light-based cancer detection, the user must have a clear image of the face without anything covering the face and keep their eyes closed to avoid the harm of UV light. Therefore, to implement standard image capture guidelines, images with eyeglasses, covered faces,

※ This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/3.0/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Manuscript received December 27, 2022; first revision March 21, 2023; accepted April 8, 2023.

* **Corresponding Author:** Dae Ho Yoon (dhyoon@skku.edu)
[1] Dept. of Science, Kangwon National University, Chooncheon, Korea (mail2kuldeepgurjar@gmail.com, ysmoon@kangwon.ac.kr)
[2] School of Advanced Material Science and Engineering, Sungkyunkwan University, Suwon, Korea (surjeetkumar@skku.edu, hamad82@skku.edu)
[3] School of Computing and Electrical Engineering, Indian Institute of Technology Mandi, Mandi, India (arnav@iitmandi.ac.in)
Kuldeep Gurjar and Surjeet Kumar contributed equally to this work.

Copyright© 2024 KIPS

unknown objects, or faces that are not clear (blurred) need to be rejected at the time of capture.

The application areas for facial image capture, such as security checks and medical skin analysis, are growing significantly. Moreover, maintaining quality is a challenge in any area where high-quality facial images are required. With the improved quality of smartphone cameras, facial image-based applications are moving towards mobile-phone-based image capture. However, the most common issues with the quality of facial images are blurry images, eyeglasses, a covered face, and unknown objects. The existing deep convolutional neural network (CNN)-based facial image capture methods used in the above-mentioned application areas lack a thorough understanding of the models, which makes predictions about facial features doubtful. To use these models for critical application areas such as healthcare and security checks, we need to have more trust in the models. In this study, we developed a model for selecting appropriate facial images for medical analysis and security operations. The model helps capture an acceptable image for analysis to generate a high-accuracy output. Given the challenges in identifying faces using smartphone-based facial image capture applications, this research has important implications for applications such as security checks and medical skin analysis, where high-quality facial images are essential for accurate analysis. Conventionally, the use of deep learning methods for quality assessment has been limited by the black-box nature of these models, which raises concerns regarding the reliability of their predictions. To address this, we developed a new approach that combines gradient-weighted class activation mapping (GradCAM) and local interpretable model-agnostic explanations (LIME) to provide visual evidence of the active regions within the image on which the deep learning model makes a prediction. A combination of these techniques eliminates the drawbacks associated with both techniques and provides improved visualizations for interpretability. This approach was implemented in a preselection system for facial images that selects clear and applicable images for further processing. To build a trustworthy model, we applied two explainability techniques, GradCAM and LIME, and a new technique that uses a combination of these techniques to provide a better explanation for the model. To demonstrate the applicability of our work, we presented a workflow for the preselection of images in medical skin analysis, where UV light-based images are used to detect melanoma in the skin. In addition, dermatologists use UV images for skin analysis from a cosmetic point-of-view. A crucial aspect of UV light is that it is harmful to the eyes. Therefore, a preselection system must verify that a clear facial image is present without eyeglasses or anything else covering the face. Fig. 1 shows the workflow of the proposed system. Here, a transfer learned model is used to initially classify the image into five categories (blur, eyeglasses, covered face, unknown object, and applicable good image). If the image contains a clear face, further processing is permitted. After this prediction, an applicable image goes through another check, in which closed and open eyes are detected. We implemented traditional image-processing techniques to verify whether the user's eyes were closed or open. When it is confirmed that a clear face with closed eyes is detected, UV light imaging can be performed for skin analysis.

1.1 Application and Research-based Contributions

- Facial image quality assessment and poor-quality image removal.
- Make face images uniform in terms of quality and format.
- A novel approach for deep learning model explainability with a combination of GradCAM and LIME.

1.2 Related Work

CNNs have shown great potential in image classification since AlexNet [1] won the ImageNet Challenge [2]. A general trend has been to introduce deeper networks [3–6]. Over time, these models started to take up more storage space and became computationally costly. Owing to the increasing interest in building small and cost-effective neural networks [7,8], a new class of models called MobileNet was designed [9]. These models are ideal for mobile and embedded applications, with a relatively high accuracy for light and computationally efficient models.

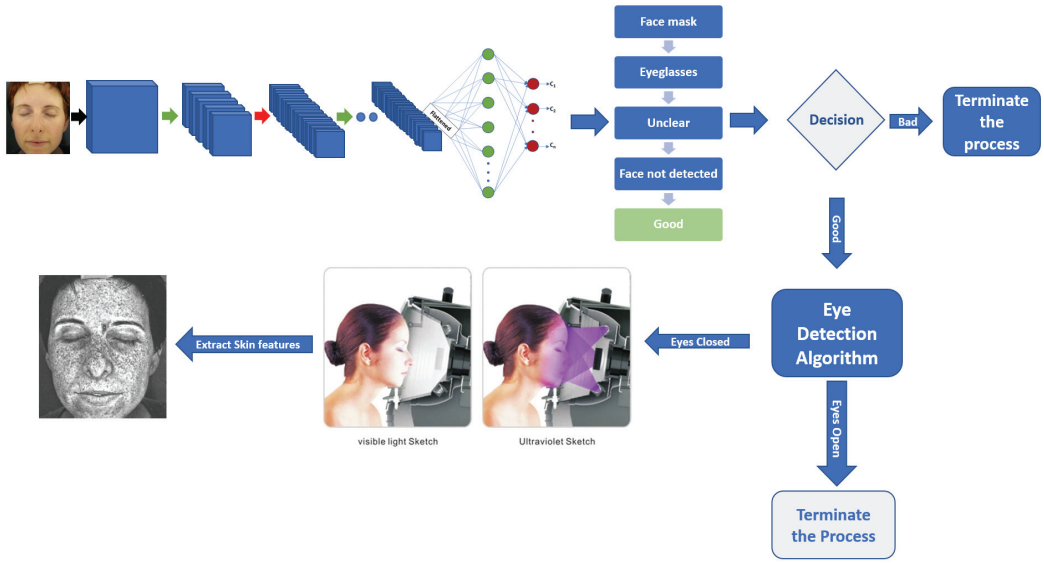


Fig. 1. Overview of the preselection process before facial feature extraction in UV-based skin analysis. The skin-analyzing device-based part of this image is taken from a commercial website named aliexpress.com.

The application domain for image classification is wide, and contributes to various stages of human life. From image-based healthcare applications to remote sensing and the automotive industry, there have been several research projects. We list some of the related works from the perspective of the application domain point-of-view. In healthcare, much work has been conducted with the help of image classification [10–13]. We focused on skin analysis, and many similar studies have been conducted in the past few years [14–20]. Anggo and Arapu [21], proposed a face recognition method using the principal component analysis (PCA) method and Fisher's linear discriminant (FDL) method. Winarno et al. [22] proposed a face recognition-based attendance system using a hybrid feature extraction method. Min et al. [23], proposed a face recognition method using the principal component analysis (PCA) method. Priadana and Habibi [24] proposed face detection using the Haar cascades method. Cao et al. [25] proposed a beauty prediction method using a residual-in-residual (RIR) structure in the neural network. As previously mentioned, deep learning has made a significant contribution to many industries. Now that almost all technological advancements are dependent on artificial intelligence (AI) in general and deep learning in particular, explainability and understanding of their decisions and internal processes are becoming important. Recently, explainable AI has become a hot area of research in diverse fields of applied AI

[26–30]. Two of the most famous among several explainability techniques are LIME, which explains a model's predictions using another local interpretable model [31], and GradCAM [32], in which a coarse localization map is produced from the last convolutional layer in a CNN model, highlighting the important regions in the image for predicting the concept. The use of LIME and GradCAM is becoming increasingly popular for explaining models in various fields [33–39]. Recently, Schlett et al. [40] provided a review of facial image quality assessment.

1.3 Dataset

We collected images belonging to five categories, namely "covered face, eyeglasses, blur, unknown objects, and good face images." These images were stored in the appropriate format for image classification, that is, training, validation, and test set format for the MobileNet architecture. 80% of the data were used for training, and 20% for validation and testing.

1.4 Model for Transfer Learning

MobileNet was selected as the base model for the transfer learning. MobileNet was introduced by Google researchers to improve the performance of mobile models on state-of-the-art computer-vision tasks and benchmarks. MobileNet was designed for mobile and embedded vision applications to reduce the intensive computations involved in earlier versions of deep neural networks. These models have a streamlined architecture that uses depth-wise separable convolutions for lightweight deep neural networks. MobileNets have shown their effectiveness in many applications. Depthwise convolutions are important building blocks for an efficient CNN architecture. In this technique, the convolutional operator is replaced by a factorized version that separates the convolution into two layers. The first layer comprises a depthwise convolution that performs a lightweight filter for each input channel. The second layer performs a 1 × 1 convolution, called pointwise convolution. This layer is responsible for building novel features by calculating the linear combination of input channels. Depth-wise separable convolutions have a reduced computational cost compared to conventional convolution operations.

A standard convolution on a $h_i \times w_i \times d_i$ input tensor L_i uses a convolutional Kernel $k \in R^{k \times k \times d_i \times d_j}$ to give $h_i \times w_i \times d_j$ output tensor L_j has a computational cost as given in Eq. (1):

$$h_i . w_i . d_i . d_j . k . k \tag{1}$$

Depth-wise separable convolutions with similar performance have a cost, as given in Eq. (2):

$$h_i . w_i . d_i (k^2 + d_i) \tag{2}$$

For MobileNet with k = 3, the cost is approximately eight to nine times more efficient than regular convolutions. We used the original MobileNet Architecture by Howard et al. [9].

In MobileNet, all the layers are followed by Batchnorm and ReLU nonlinearity, except for the last fully connected layer. Fig. 2 shows the contents of the convolutional layer in the MobileNet. This model is suitable for mobile applications because it helps obtain very memory-efficient inferences.

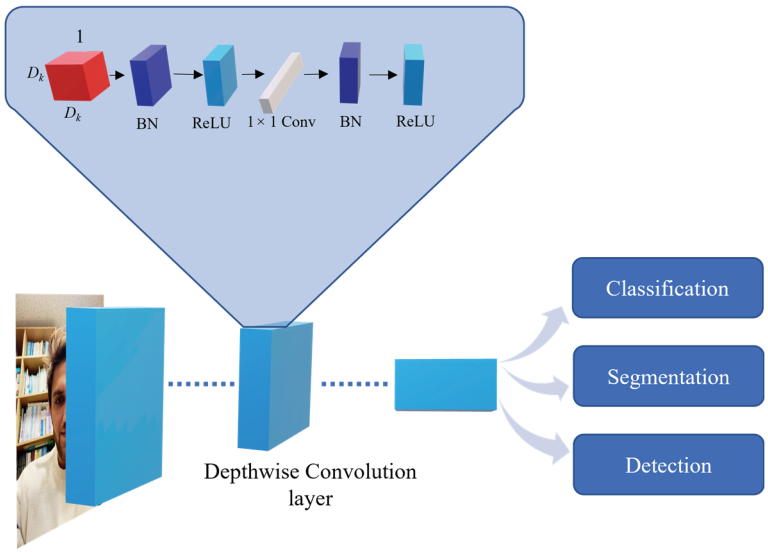


Fig. 2. Details of the depthwise convolution layer in MobileNet.

1.5 Interpretability

When using AI models for healthcare and medical applications, we need to evaluate the model deeply and not just rely on the accuracy of the model on the training datasets. AI algorithms and hardware have evolved significantly in the last decade, but they do not provide useful information regarding the dataset, and the model can learn certain biases that decrease its trust. To obtain trust in the model, explainability techniques were used.

In this study, we combined two popular techniques for image classification model explainability. We used a combination of heatmaps generated from GradCAM and applied LIME to determine how the heatmaps and superpixels correlate with each other. In this study, we used the LIME technique to verify the model. A method combining GradCAM and LIME for interpreting the results.

Lime uses the following steps at first to obtain an interpretable model for model f for example E.

1. Construct superpixels d in E, which are small homogenous patches in the image.
2. Generate n new images $x_1, x_2, \ldots x_n$ by turning on and off the superpixels.
3. Use these images to generate new predictions y from model f. $y = f(x_i)$.
4. Build a local weighted model B fitting the y_is to the presence or absence of superpixels.

According to LIME, every coefficient of the local weighted surrogate model corresponds to a superpixel of the original image E. A simplistic way to observe this is by visualizing only the superpixels with the highest positive coefficient for B_n, blocking the remaining superpixels.. Therefore, we combined GradCAM and LIME to provide more precise explanations for the model. In 2017, Selvaraju et al. [32] observed that convolutional layers can capture spatial information from the input data lost in the fully connected layers. Thus, the last convolutional layer contains both high-level semantics and detailed spatial information. GradCAM uses the gradient of features flowing into the last convolutional layer to produce a localization map that highlights the important regions in the image for prediction. This technique provides visual proof of the functioning of the trained neural network, which allows us to

investigate the model, which is traditionally considered a black box. GradCAM uses the gradients of any target image in a classification network passing through the last convolutional layer to create a localization map that shows the image regions that are significant for the prediction. This technique is applied to almost all CNN model families. GradCam can be used to explain the predictions from a model using heat maps. To obtain a discriminative localization map Grad CAM $L^c_{Grad-CAM}$ of width u and height v for the target class, we follow these steps.

1. Compute the gradients for the class c, y^c in the last layer before the softmax layer, that is $\frac{\partial y^c}{\partial A^k}$.
2. We used these backflowing gradients and applied global average pooling over the width i and height j dimensions to obtain neuron importance weights a^c_k.

$$a^c_k = \frac{1}{z}\sum\sum\frac{\partial y^c}{\partial A^k_{ij}}. \tag{3}$$

3. Find the weighted combination of forward activation maps and apply ReLU to obtain:

$$L^c_{Grad-CAM} = ReLU\left(\sum_k a^c_k A^k\right). \tag{4}$$

This result is a course heatmap of the same size as that of the convolutional feature map. ReLU was applied to the linear combination of maps because of our interest in the features that have a positive role in the prediction of a particular class.

2. Methodology

The dataset contained five different categories of images, which were used for transfer learning on the pre-trained MobileNet architecture. The prediction for each class from the model was first explained using LIME. LIME shows the most significant superpixels responsible for prediction. Once a prediction gives us the preselection of images, we use the combined explainability technique using GradCAM and LIME to explain the model's performance in choosing the right image. Once a proper face image is classified, a traditional Haar cascade classifier is used to check the image with eyes closed or open.

2.1 Training

A MobileNet model with pre-trained ImageNet weights was loaded using the TensorFlow deep-learning framework. The model was slightly modified to provide predictions for the five classes. The last layer in the model was replaced by a softmax activation layer with five output neurons. The remaining layers of the model were not trained to achieve faster convergence and significant feature reuse.

2.2 Explainability with LIME

To validate the image classification model, we applied LIME. Fig. 3 illustrates this process. A mask was generated that marked the boundaries of the superpixels in the image. Perturbed image samples are generated by turning different regions, called superpixels ON and OFF, within the image. Predicting all the samples and selecting the samples predicted to belong to the class, we must interpret the image.

Compute the distance metric to evaluate the difference between the perturbed samples and the original image.

The distance metric here is the cosine distance because the images and samples are multidimensional vectors. After calculating the cosine distance metric, the kernel function maps the distance to a weight between zero and one. Then, a weighted linear model was trained using the samples, weights, and prediction vectors. This gives us a coefficient that shows the effect of each superpixel on the prediction of the target class. From these coefficients, we select the superpixels that contribute most to the correct prediction of the target class.

2.3 GradCAM and LIME

GradCAM generates a heatmap highlighting the important regions in the image for predicting the target class. Whereas LIME first creates superpixels within the target image and then selects the most important superpixels to give a visual interpretation for the prediction. So far, these techniques have mostly been used separately. A combination of these techniques makes solid visual evidence that is verified by the two most famous explainability techniques in computer vision. Here, we generate the heatmaps using GradCAM and the gradients for the classes in the last convolutional layer of the model. Based on these

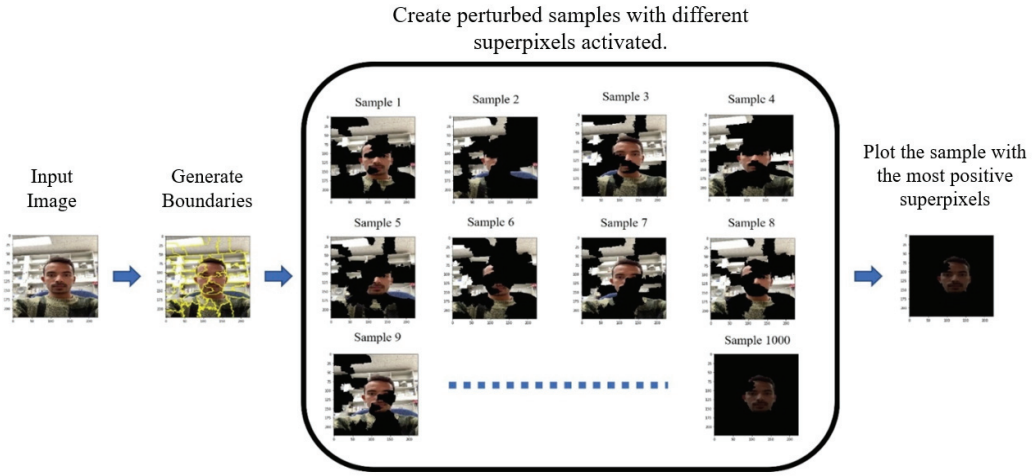


Fig. 3. Method for selective the superpixels that positively contribute during the prediction of the target class.

gradients, the neuron importance weights are calculated. Then the weighted combination of the forward activation maps is obtained, and ReLU activation is applied. A heatmap is generated and overlapped over the original image to obtain areas of higher activation. After this step, superpixels are generated in the image. The most important superpixels are selected using the method explained below.

1. Calculate the gradients $\frac{\partial y^c}{\partial A^k}$ for the target class c in the last convolutional layer during prediction.
2. Use these backflowing gradients and apply global average pooling over the width *i* and height *j* dimensions to obtain neuron importance weights a_k^c.

$$a_k^c = \frac{1}{Z} \sum \sum \frac{\partial y^c}{\partial A_{ij}^k}. \tag{5}$$

3. Find the weighted combination of forward activation maps, and apply ReLU to obtain.

$$L^{c}_{Grad-CAM} = ReLU\left(\sum_{k} a^{c}_{k} A^{k}\right) \tag{6}$$

4. Overlap this heatmap over the original image to get a new image E that shows the features that have a positive influence on the prediction of the target class.
5. Now generate superpixels d in the E.
6. Now generate 1,000 new images, $x_1, x_2, x_3, \ldots x_{1000}$ by turning ON and OFF the superpixels.
7. Calculate the distance metric for each sample from the original image. Given that the samples are just the perturbations of the original image, and all these are multidimensional vectors, cosine distance is used as a distance metric.
8. Map these values to a weight between zero and one with the help of a kernel function.
9. Use these weights to generate a weighted linear model that fits the perturbations and output.
10. Based on this model, the coefficient for each superpixel is obtained, representing the effect of superpixels in the prediction.
11. Select the top superpixels and plot them for visualization.

The combination of GradCAM and LIME produces a result that has a heat map and only important superpixels. Defining the most important areas in the image by combining these two techniques.

2.4 Haar Cascade

To show the applicability of this work to skin feature analysis, we used an open-source Haar cascade model in OpenCV [41]. To detect the presence of eyes in the image and then predict if the eyes are open or closed. We integrated this model with our system. The applicable face image selected by the Mobilenet model was given to the cascade model to predict open or closed eyes within the image.

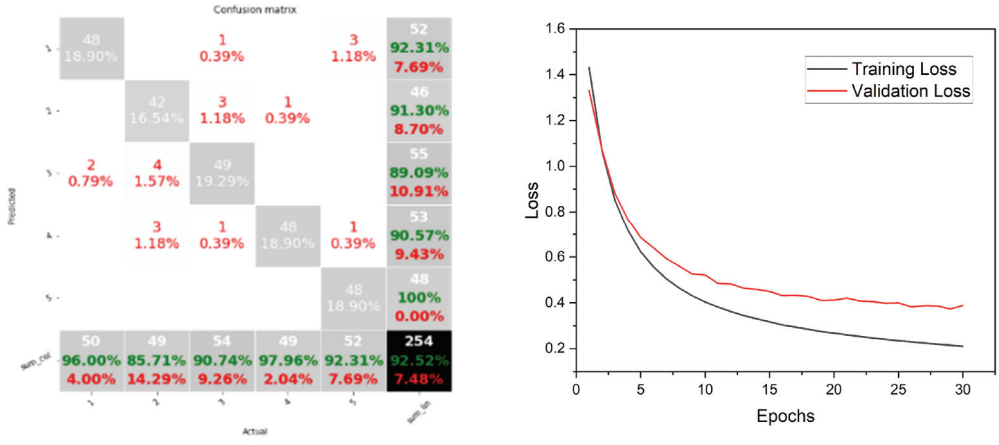
2.5 Implementation Structure

The implementation structure of the approach includes using a transfer learning-based model for the initial classification of images into five categories: blur, eyeglasses, covered face, unknown object, and applicable good image. The realization process of the approach involves training the transfer learning-based model using a large dataset of facial images with varying illuminations, camera quality variations, and background-specific variations to ensure that it can accurately classify images into the five categories. The combination of GradCAM and LIME techniques is then used to explain the model's predictions and verify the selection of high-quality face images. For transfer learning, a MobileNet model with pre-trained ImageNet weights was loaded using TensorFlow, and the model was modified to predict five output classes. The model was trained on five classes until convergence. After this, a combination of Grad CAM and LIME was used for explainability. The system and hyperparameters used in this work include an RTX 3090 GPU and an Intel Core i9 CPU with 32 GB of RAM, providing the computational power necessary for training and analysis. The environment is set up using Anaconda Spyder with Python 3.7, and the TensorFlow 2.0 library is employed for deep learning tasks. The model uses transfer learning with 8 output layers configured as softmax, and is trained over 30 epochs with a learning rate of 0.0001.

For model interpretability, Grad-CAM and LIME are applied with specific parameters: 1,000 sample images, a kernel size of 4, a maximum distance of 200, a kernel width of 0.25, and the top 5 features are identified for explanation purposes.

3. Results and Discussion

We collected a custom dataset with five different classes: blur, glasses, good, mask, and wrong objects for image classification. In the first step, a MobileNet model with pre-trained weights for the ImageNet dataset was used for transfer learning. To verify the model and have trustworthy predictions, we used exploitability techniques. LIME was used to explain the model's behavior to each class in the dataset. To have better explanations, both LIME and GradCAM were applied to see the similarity in results. Using the predicted images from this model to show the application, we decide if the eyes are open or closed. The transfer learned MobileNet Model converged within 30 epochs. Fig. 4(a) shows the training accuracy and Fig 4(b) shows the loss for training and validation.



(a) (b)

Fig. 4. (a) Loss curve for training and validation dataset and (b) confusion matrix for test set predictions by the transfer learned model. Here, class 1 = blur, class 2 = eyeglasses, class 3 = good, class 4 = covered face, and class 5 = wrong.

As the confusion matrix shows, the model can successfully differentiate between the good face images that apply to facial feature analysis and security applications. To verify the performance of the used transfer learning approach, we have compared the results for our model with simple CNN architecture. Table 1 shows the results for the transfer learned MobileNet architecture, and a simple CNN architecture on the same dataset. As the dataset contains very similar images, a highly optimized model is required to accurately predict image classes. This comparison between the two models validates that MobileNet has significantly better results for classification in this dataset.

Although the predictions are accurate for this dataset with the transfer learned MobileNet architecture, we need to confirm that the model uses the correct features for prediction. Fig. 5 shows the LIME results for all five categories of images.

We can see from the masked images that the model looks at the right superpixels for the prediction of facial image categories and the model looks at random superpixels for the wrong objects. This explains the usability of this model for facial image classification for skin feature extraction.

Table 1. Comparison of a simple CNN model with transfer learned MobileNet architecture for classification in 5 classes

Class	Simple CNN			MobileNet architecture		
	Precision	Recall	F1_score	Precision	Recall	F1_score
Blur	0.54	0.68	0.60	0.92	0.96	0.94
Eyeglasses	0.36	0.43	0.39	0.91	0.86	0.88
Good images	0.91	0.57	0.70	0.89	0.91	0.90
Covered face	0.54	0.80	0.64	0.91	0.98	0.94
Wrong object	0.89	0.46	0.61	1.00	0.92	0.96
Macro average	0.65	0.59	0.59	0.93	0.93	0.93
Weighted average	0.66	0.59	0.59	0.93	0.93	0.93

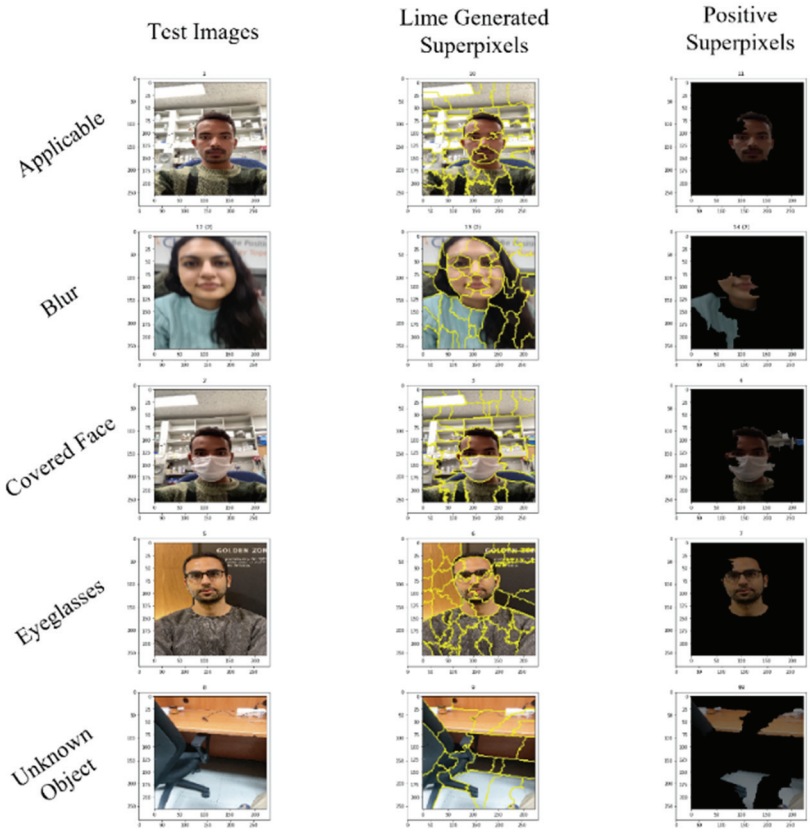
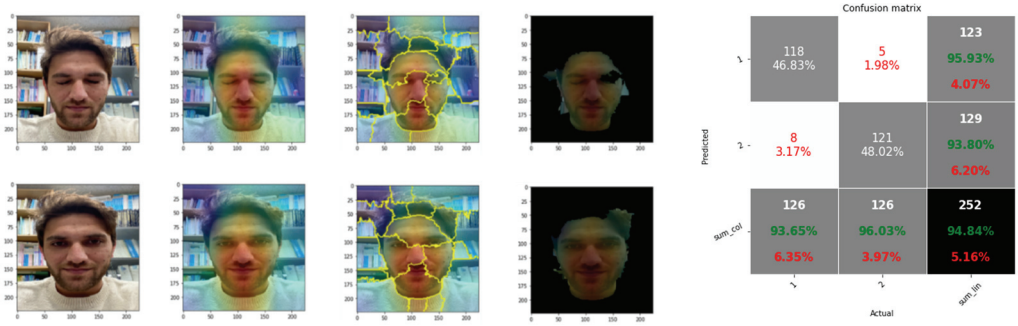


Fig. 5. Local interpretable model-agnostic explanations for each predicted class.

LIME is a useful technique for model interpretability, but the explanations produced by LIME have a limitation based on the parameters used for creating superpixels, and the number of superpixels. In some cases, LIME ignores pixels that have been segmented into smaller regions that have no clear facial

features present in them. Also, background plays a vital role in creating segments within the image for LIME. Certain background conditions in images will lead to wrong regions in the interpretability analysis [42]. For these reasons, and to further explain the model and build trust in the predictions, we combined GradCAM and LIME for better visualization and explanation of the model's prediction behavior. This combination of two techniques has better visualization and reduces the chances for wrong interpretations of the model. To do so, the heatmap was first generated by calculating the gradients for the predicted class in the last convolution layer of the model. Then, by using these backpropagating gradients, neuron importance weights were obtained. ReLU activation was applied to obtain a course heatmap, and then the LIME technique was applied. Both techniques confirm the model's usefulness, and it has been noted that the model looks at the right feature within the image for classification. Fig. 6(a) shows the results of the explainability using a combination of GradCAM and LIME. Once the image has been classified as a good face image, in which the face is visible, and nothing is covering the face. We check if the eyes are closed or open using the cascade classifier available in OpenCV. Fig. 6(b) shows the confusion matrix for the classification of open or closed eyes by the cascade classifier. The use of preselection and classification by the eye provides only good-quality images to reach the final stage of cascade classification. This makes the system more accurate.



(a) (b)

Fig. 6. (a) Combination of GradCAM and LIME applied to open and closed-eye face images and (b) confusion matrix for open eyes and closed eyes classification. Here class 1 = face with open eyes and class 2 = face with closed eyes.

To compare this work with the related work we looked at different performance factors that are useful for practical applications of a computer vision model. Here we found that our model performs very well on the given dataset also a general comparison of the three main computer techniques is given in Table 2 [9,31,32,43–46].

Table 2. Comparison of the techniques used for image classification and quality assessment

Factor	Image processing and ML-based technique	Deep learning	Explainable transfer learning (this work)	Reference
Accuracy	High	Medium	High	[43]
Dataset size	Small	Large	Small	[44]
Pre-processing time	High	Low	Low	[45]
Training time	Low	High	Low	[46]
Memory efficiency	Low	Low	High	[9]
Trustworthiness	High	Low	High	[31], [32]

4. Conclusion

Deep learning is associated with critical decision-making in many fields, including automotive, telecommunications, security, and healthcare. The models used in deep learning were considered black boxes for a long time. Now, these models are being explored with emerging explainability techniques. In this work, we employed transfer learning on the MobileNet architecture for classifying facial images for security and healthcare purposes and verified its performance with explainability techniques. This model was trained using face images collected by a smartphone camera. The model can predict with an accuracy of 92%. The application area for these images involves critical decision-making in security and healthcare, which makes the explainability of the model crucial here. We used LIME and GradCAM to explain the model. Initially, LIME was used for a visual explanation of the model's prediction for each class. The results explain why our model looks at the correct features while classifying face images. For the three categories “good image, eyeglasses, and covered face,” this model predicts using superpixels mostly centered around the face region, and for the two categories "blur and wrong object," the model selects random superpixels. To have better explanations, we combined GradCAM and LIME; here we give a new strategy to apply a combination of both techniques to have better explanations and learn more about the model. LIME, for instance, chooses different superpixels while making interpretations, and this has been reported multiple times. So, a combination of GradCAM and LIME ensures more trust in the decisions we make using deep learning models.

Conflict of Interest

The authors declare that they have no competing interests.

Funding

None.

References

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” *Communications of the ACM*, vol. 60, no. 6, pp. 84-90, 2017. https://doi.org/10.1145/3065386

[2] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, et al., “ImageNet large scale visual recognition challenge,” *International Journal of Computer Vision*, vol. 115, pp. 211-252, 2015. https://doi.org/10.1007/s11263-015-0816-y

[3] K. Simonyan, “Very deep convolutional networks for large-scale image recognition,” 2014 [Online]. Available: https://arxiv.org/abs/1409.1556.

[4] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the inception architecture for computer vision,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, 2016, pp. 2818-2826. https://doi.org/10.1109/CVPR.2016.308

[5] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, “Inception-v4, Inception-ResNet and the impact of residual connections on learning,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, no. 1, pp. 4278-4284, 2017. https://doi.org/10.1609/aaai.v31i1.11231

[6] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, 2016, pp. 770-778. https://doi.org/10.1109/CVPR.2016.90

[7] J. Jin, A. Dundar, and E. Culurciello, "Flattened convolutional neural networks for feedforward acceleration," 2014 [Online]. Available: https://arxiv.org/abs/1412.5474.

[8] M. Wang, B. Liu, and H. Foroosh, "Factorized convolutional neural networks," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, Venice, Italy, 2017, pp. 545-553. https://doi.org/10.1109/ICCVW.2017.71

[9] A. G. Howard, "MobileNets: efficient convolutional neural networks for mobile vision applications," 2017 [Online]. Available: https://arxiv.org/abs/1704.04861.

[10] A. Shoeibi, M. Khodatars, M. Jafari, N. Ghassemi, D. Sadeghi, P. Moridian, et al., "Automated detection and forecasting of covid-19 using deep learning techniques: a review," 2020 [Online]. Available: https://arxiv.org/abs/2007.10785v1.

[11] I. R. I. Haque and J. Neubert, "Deep learning approaches to biomedical image segmentation," *Informatics in Medicine Unlocked*, vol. 18, article no. 100297, 2020. https://doi.org/10.1016/j.imu.2020.100297

[12] H. Guan and M. Liu, "Domain adaptation for medical image analysis: a survey," *IEEE Transactions on Biomedical Engineering*, vol. 69, no. 3, pp. 1173-1185, 2022. https://doi.org/10.1109/TBME.2021.3117407

[13] W. A. Hassan, Y. H. Ali, and N. J. Ibrahim, "A survey of latest techniques in medical image classification," in *Proceedings of 2021 International Conference on Communication & Information Technology (ICICT)*, Basrah, Iraq, 2021, pp. 68-73. https://doi.org/10.1109/ICICT52195.2021.9568454

[14] K. Kojima, K. Shido, G. Tamiya, K. Yamasaki, K. Kinoshita, and S. Aiba, "Facial UV photo imaging for skin pigmentation assessment using conditional generative adversarial networks," *Scientific Reports*, vol. 11, article no. 1213, 2021. https://doi.org/10.1038/s41598-020-79995-4

[15] R. Kips, L. Tran, E. Malherbe, and M. Perrot, "Beyond color correction: Skin color estimation in the wild through deep learning," *Electronic Imaging*, vol. 32, article no. 082, 2020. https://doi.org/10.2352/ISSN.2470-1173.2020.5.MAAP-082

[16] E. Borsting, R. DeSimone, M. Ascha, and M. Ascha, "Applied deep learning in plastic surgery: classifying rhinoplasty with a mobile app," *Journal of Craniofacial Surgery*, vol. 31, no. 1, pp. 102-106, 2020. https://doi.org/10.1097/SCS.0000000000005905

[17] M. A. Taufiq, N. Hameed, A. Anjum, and F. Hameed, "m-Skin Doctor: a mobile enabled system for early melanoma skin cancer detection using support vector machine," in *eHealth 360°*. Cham, Switzerland: Springer, 2017, pp. 468-475. https://doi.org/10.1007/978-3-319-49655-9_57

[18] C. I. Moon and O. Lee, "Age-dependent skin texture analysis and evaluation using mobile camera image," *Skin Research and Technology*, vol. 24, no. 3, pp. 490-498, 2018. https://doi.org/10.1111/srt.12459

[19] K. Ramlakhan and Y. Shang, "A mobile automated skin lesion classification system," in *Proceedings* of *2011 IEEE 23rd International Conference on Tools with Artificial Intelligence*, Boca Raton, FL, USA, 2011, pp. 138-141. https://doi.org/10.1109/ICTAI.2011.29

[20] S. Kim, D. Cho, J. Kim, M. Kim, S. Youn, J. E. Jang, et al., "Smartphone-based multispectral imaging: system development and potential for mobile skin diagnosis," *Biomedical Optics Express*, vol. 7, no. 12, pp. 5294-5307, 2016. https://doi.org/10.1364/BOE.7.005294

[21] M. Anggo and L. Arapu, "Face recognition using fisherface method," *Journal of Physics: Conference Series*, vol. 1028, no. 1, article no. 012119, 2018. https://doi.org/10.1088/1742-6596/1028/1/012119

[22] E. Winarno, I. H. Al Amin, H. Februariyanti, P. W. Adi, W. Hadikurniawati, and M. T. Anwar, "Attendance system based on face recognition system using cnn-pca method and real-time camera," in *Proceedings of 2019 International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*,

Yogyakarta, Indonesia, 2019, pp. 301-304. https://doi.org/10.1109/ISRITI48646.2019.9034596

[23] W. Y. Min, E. Romanova, Y. Lisovec, and A. M. San, "Application of statistical data processing for solving the problem of face recognition by using principal components analysis method," in *Proceedings of 2019 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus)*, Saint Petersburg and Moscow, Russia, 2019, pp. 2208-2212. https://doi.org/10.1109/EIConRus.2019.8657240

[24] A. Priadana and M. Habibi, "Face detection using Haar cascades to filter selfie face image on Instagram," in *Proceedings of 2019 International Conference of Artificial Intelligence and Information Technology (ICAIIT)*, Yogyakarta, Indonesia, 2019, pp. 6-9. https://doi.org/10.1109/ICAIIT.2019.8834526

[25] K. Cao, K. N. Choi, H. Jung, and L. Duan, "Deep learning for facial beauty prediction," *Information*, vol. 11, no. 8, article no. 391, 2020. https://doi.org/10.3390/info11080391

[26] B. H. Van der Velden, H. J. Kuijf, K. G. Gilhuijs, and M. A. Viergever, "Explainable artificial intelligence (XAI) in deep learning-based medical image analysis," *Medical Image Analysis*, vol. 79, article no. 102470, 2022. https://doi.org/10.1016/j.media.2022.102470

[27] D. Jin, E. Sergeeva, W. H. Weng, G. Chauhan, and P. Szolovits, "Explainable deep learning in healthcare: a methodological survey from an attribution view," *WIREs Mechanisms of Disease*, vol. 14, no. 3, article no. e1548, 2022. https://doi.org/10.1002/wsbm.1548

[28] E. Zablocki, H. Ben-Younes, P. Perez, and M. Cord, "Explainability of deep vision-based autonomous driving systems: review and challenges," *International Journal of Computer Vision*, vol. 130, article no. 2425-2452, 2022. https://doi.org/10.1007/s11263-022-01657-x

[29] G. Ras, N. Xie, M. Van Gerven, and D. Doran, "Explainable deep learning: a field guide for the uninitiated," *Journal of Artificial Intelligence Research*, vol. 73, pp. 329-396, 2022. https://doi.org/10.1613/jair.1.13200

[30] G. Novakovsky, N. Dexter, M. W. Libbrecht, W. W. Wasserman, and S. Mostafavi, "Obtaining genetics insights from deep learning via explainable artificial intelligence," *Nature Reviews Genetics*, vol. 24, pp. 125-137, 2023. https://doi.org/10.1038/s41576-022-00532-2

[31] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you?" Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, CA, USA, 2016, pp. 1135-1144. https://doi.org/10.1145/2939672.2939778

[32] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE International Conference on Computer Vision*, Venice, Italy, 2017, pp. 618-626. https://doi.org/10.1109/ICCV.2017.74

[33] Q. Liu and P. Hu, "Extendable and explainable deep learning for pan-cancer radiogenomics research," *Current Opinion in Chemical Biology*, vol. 66, article no. 102111, 2022. https://doi.org/10.1016/j.cbpa.2021.102111

[34] M. I. Patel, S. Singla, R. A. A. Mattathodi, S. Sharma, D. Gautam, and S. R. Kundeti, "Simulating realistic MRI variations to improve deep learning model and visual explanations using GradCAM," in *Proceedings of 2021 IEEE International Conference on Cloud Computing in Emerging Markets (CCEM)*, Virtual Event, USA, 2021, pp. 1-8. https://doi.org/10.1109/CCEM53267.2021.00011

[35] H. Panwar, P. K. Gupta, M. K. Siddiqui, R. Morales-Menendez, P. Bhardwaj, and V. Singh, "A deep learning and grad-CAM based color visualization approach for fast detection of COVID-19 cases using chest X-ray and CT-Scan images," *Chaos, Solitons & Fractals*, vol. 140, article no. 110190, 2020. https://doi.org/10.1016/j.chaos.2020.110190

[36] M. Xiao, L. Zhang, W. Shi, J. Liu, W. He, and Z. Jiang, "A visualization method based on the Grad-CAM for medical image segmentation model," in *Proceedings of 2021 International Conference on Electronic Information Engineering and Computer Science (EIECS)*, Changchun, China, 2021, pp. 242-247. https://doi.org/10.1109/EIECS53707.2021.9587953

[37] T. He, J. Guo, N. Chen, X. Xu, Z. Wang, K. Fu, L. Liu, and Z. Yi, "MediMLP: using grad-cam to extract crucial variables for lung cancer postoperative complication prediction," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 6, pp. 1762-1771, 2020. https://doi.org/10.1109/JBHI.2019.2949601

[38] S. Dey, P. Chakraborty, B. C. Kwon, A. Dhurandhar, M. Ghalwash, F. J. S. Saiz, et al., "Human-centered explainability for life sciences, healthcare, and medical informatics," *Patterns*, vol. 3, no. 5, article no. 100493, 2022. https://doi.org/10.1016/j.patter.2022.100493

[39] I. Palatnik de Sousa, M. Maria Bernardes Rebuzzi Vellasco, and E. Costa da Silva, "Local interpretable model-agnostic explanations for classification of lymph node metastases," *Sensors*, vol. 19, no. 13, article no. 2969, 2019. https://doi.org/10.3390/s19132969

[40] T. Schlett, C. Rathgeb, O. Henniger, J. Galbally, J. Fierrez, and C. Busch, "Face image quality assessment: a literature survey," *ACM Computing Surveys (CSUR)*, vol. 54, no. 10s, article no. 210, 2022. https://doi.org/10.1145/3507901

[41] G. Bradski, "The OpenCV Library," *Dr. Dobb's Journal of Software Tools*, vol. 25, no. 11, pp. 120-125, 2000.

[42] M. Zhu, B. Zang, L. Ding, T. Lei, Z. Feng, and J. Fan, "LIME-based data selection method for SAR images generation using GAN," *Remote Sensing*, vol. 14, no. 1, article no. 204, 2022. https://doi.org/10.3390/rs14010204

[43] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He, "A comprehensive survey on transfer learning," *Proceedings of the IEEE*, vol. 109, no. 1, pp. 43-76, 2021. https://doi.org/10.1109/JPROC.2020.3004555

[44] N. O'Mahony, S. Campbell, A. Carvalho, S. Harapanahalli, G. V. Hernandez, L. Krpalkova, D. Riordan, and J. Walsh, "Deep learning vs. traditional computer vision," in *Advances in Computer Vision*. Cham, Switzerland: Springer, 2020, pp. 128-144. https://doi.org/10.1007/978-3-030-17795-9_10

[45] A. Yilmaz, A. A. Demircali, S. Kocaman, and H. Uvet, "Comparison of deep learning and traditional machine learning techniques for classification of pap smear images," 2020 [Online]. Available: https://arxiv.org/abs/2009.06366.

[46] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345-1359, 2010. https://doi.org/10.1109/TKDE.2009.191

Kuldeep Gurjar **https://orcid.org/0000-0002-0800-6410**

He received a B.S. (2006) and M.S. degrees (2010) in Computer Science, from Department of Computer Science and Information Technology, University of Rajasthan, Jaipur. From the year 2010 to 2011, he worked for a website development company (Octal info. Solutions). From March 2012 to 2018 he was with the Department of Computer Science and Engineering from Kangwon National University as a Ph.D. candidate. He is an assistant professor at the University of Suwon and his current research areas are digital health and computer vision.

Surjeet Kumar **https://orcid.org/0000-0002-4129-7213**

He received a B.S. in Electronic Engineering from BUITEMS, Pakistan. Since March 2019, he is with the School of Advanced Materials Science and Engineering at Sungkyunkwan University as a Ph.D. candidate. His current research interests are explainable artificial intelligence and computer vision.

Arnav Bhavsar **https://orcid.org/0000-0003-2849-4375**

He received his Ph.D. from IIT Madras in 2011. He then worked as a postdoc at GE Global Research and the University of North Carolina, Chapel Hill in 2011 and 2012, respectively. He joined IIT Mandi in 2013 as an assistant professor and is presently working there as an associate professor since 2019.

Kotiba Hamad **https://orcid.org/0000-0001-5306-2932**

He is an assistant professor at the Department of Advanced Materials Science and Engineering at Sungkyunkwan University. He has vast experience in machine learning and deep learning for interdisciplinary research.

Yang-Sae Moon **https://orcid.org/0000-0002-2396-0405**

He received a B.S. (1991), M.S. (1993), and Ph.D. (2001) degrees in Computer Science from the Korea Advanced Institute of Science and Technology (KAIST). From 1993 to 1997, he was a research engineer at Hyundai Syscomm Inc., where he participated in developing 2G and 3G mobile communication systems. He is currently a professor in the computer science department at Kangwon National University.

Dae Ho Yoon **https://orcid.org/0000-0003-0720-6928**

He is a professor at the Department of Advanced Materials Science and Engineering, at Sungkyunkwan University, South Korea. He has worked on diverse research topics. He focuses on interdisciplinary research ideas.