Wireless Personal Communications (2023) 129:663–680
https://doi.org/10.1007/s11277-022-10127-z

# XAI-FR: Explainable AI-Based Face Recognition Using Deep Neural Networks

Ankit Rajpal[1] · Khushwant Sehra[2] · Rashika Bagri[1] · Pooja Sikka[1]

Accepted: 30 November 2022 / Published online: 13 December 2022

**Abstract**
Face Recognition aims at identifying or confirming an individual's identity in a still image or video. Towards this end, machine learning and deep learning techniques have been successfully employed for face recognition. However, the response of the face recognition system often remains mysterious to the end-user. This paper aims to fill this gap by letting an end user know which features of the face has the model relied upon in recognizing a subject's face. In this context, we evaluate the interpretability of several face recognizers employing deep neural networks namely, LeNet-5, AlexNet, Inception-V3, and VGG16. For this purpose, a recently proposed explainable AI tool–Local Interpretable Model-Agnostic Explanations (LIME) is used. Benchmark datasets such as Yale, AT &T dataset, and Labeled Faces in the Wild (LFW) are utilized for this purpose. We are able to demonstrate that LIME indeed marks the features that are visually significant features for face recognition.

**Keywords** Explainable AI · Face Recognition · Deep Neural Network · LeNet-5 · AlexNet · Inception-V3 · VGG16

## 1 Introduction

A face recognition system provides a means for the automatic recognition of the various subjects against the already stored datasets. The applications of face recognition include unlocking smartphones, searching missing persons, etc. With advancement in digital

✉ Khushwant Sehra
sehrakhushwant@gmail.com

Ankit Rajpal
arajpal@cs.du.ac.in

Rashika Bagri
rashika.mcs19.du@gmail.com

Pooja Sikka
pooja.mcs19.du@gmail.com

[1] Department of Computer Science, University of Delhi, New Delhi 110007, India

[2] Department of Electronic Science, University of Delhi, South Campus, New Delhi 110021, India

technology, face recognition is also being used in various cyber investigations [1, 2]. This complements well with the aftermath of the COVID-19 outbreak, which has forced the world to adopt face recognition technology with a primary focus on the contact-less operation [3, 4]. The most prominent issue that affects the outcome of face recognition systems is related to the illumination variation, which may be due to varying lighting conditions [5]. Aside from this, concerns with posture variation or camera angles can cause significant changes in facial appearance and/or form, as well as intra-subject face variations [2, 4–6]. Also, the occlusion of a face by other objects or varying levels of emotions may impede the performance of face recognition systems [7].

To deal with uncontrolled environments which may lead to false positives and negatives during classification, and to improve the overall performance of the face recognition systems, various techniques have been proposed by research groups all over the world. Tang et al. [8] employed a novel Distance Weighted Linear Regression Classifier (DWLRC) to overcome the problem of faces being misclassified in the systems using linear regression. The distance between each sample point and the original linear space is utilized as an adjustment parameter to optimize the regression line in order to produce a better result under varied scenarios. The proposed methodology outperforms the traditional Linear Regression Classifier (LRC), Nearest-Farthest Subspace (NFS), Kernel Linear Regression Classifier (KLRC), and Center-based Weighted Kernel Linear Regression Classifier (CWKLR). The method reported recognition rates of 96% on the AT &T dataset.

In addition to the above, other Machine Learning (ML) algorithms have also been utilized for the robust classification of face samples. Damale et al. [9] have presented three different methods based on ML paradigms, based on Support Vector Machine (SVM), Perceptron Multilayer (MLP), and Convolutional Neural Network (CNN). SVM and MLP approaches rely on features extracted through Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA), whereas in CNN the images are directly used as a feature vector. The proposed systems, as reported, demonstrated test accuracies of nearly 87%, 86.5%, and 98% for SVM, MLP, and CNN respectively on self-generated databases. Abuzneid et al. [10] proposed an improved face recognition system using Back-Propagation Neural Network (BPNN) supported by a pre-processing through a Haar-Cascade detection, Histogram Equalization (HE), and local feature extraction through Local Binary Patterns (LBP) descriptor. The system, as reported achieves an accuracy of $\sim$ 98% on both Yale and ORL datasets with a significant reduction in computational time. In addition to this, an Ensemble-aided face recognition approach proposed by Venkateswar et al. [11], demonstrated good performance in rough environments by relying on Image Frontalization and pre-processing through different enhancement methods. The feature extraction is based on several descriptors including histograms of gradients (HOG), improved center-symmetric local binary patterns (ICSLBP), SIFT descriptors, and dominant color structure descriptors for final classification through SVM. This approach combines the utility of robust preprocessing with a good classification accuracy of 99% and 94% for the data samples from FERET and LFW databases respectively.

Qu et al. [12] demonstrated another face recognition system based on CNN and FPGA. This is important primarily because the FPGA is able to implement parallel computing and can be used to design exotic logic circuits, which help in achieving higher processing speed in comparison to standard CPU, GPUT, and TPU processors. The network is reported to work at the clock frequency of 50MHz achieving the recognition speeds of up to 400FPS and a recognition rate of 99.25%. A modified Deep Neural Network (DNN) system was reported by Aiman et al. [13], which consists of CNNs, RELUs, and fully connected layers to improve recognition rates when the training dataset is limited. This is done by using the data augmentation technique which helps in increasing the number of training face samples. Further, as reported, this also improves the generalization capabilities of the employed CNN systems. The group reports, the accuracy of 95.21% the AT &T face database for 4 training samples whereas 99.92% for 5 training samples. A new eight-layered CNN architecture was proposed by Coskun et al. [14], which relies on the batch Normalization process to improve the accuracy of the proposed system and a Softmax classifier to classify the face samples. Görgel et al. [15] proposed another face recognition system that uses deep-stacked denoising sparse autoencoders (DSDSA) for the identification of face areas and/or distinctive landmark features. The classification methodology relies on multi-class support vector machines (SVM) and SoftMax classifiers. A novel deep neural network presented by Zhao et al. [16], makes use of CNN to realize a feature vector for human face representation. This is followed by PCA for dimension reduction to remove the redundant and contaminated visual features. The authors report a recognition rate of 98.52% on the CAS-PEAL dataset, and the system as reported is robust under face recognition attacks. FaceNet [17], introduced by Google researchers, proposed a face recognizer based on machine learning. The group makes use of two pre-trained models from CASIA-WebFace and VGGFace2 for testing the system performance. The proposed system is robust and can achieve recognition rates of 100%. This is because FaceNet relies on comparing each face sample, one after the other with the pre-trained Tensorflow model. The pre-trained data model has a considerable influence on the accuracy of the FaceNet approach, with VGGFace2 producing better average recognition accuracy.

Face recognition has also found its way towards robust and intelligent video surveillance. In this regard, Wang et al. [18] have put forward a brute force detection method for violence detection based on CNN and trajectory features. The authors have proposed two methods to deal with face images extracted at lower resolutions from a surveillance video by using multi-foot input and SPP-based CNN models. The accuracy as reported on Crow and Hockey datasets is 92% and 97.6%, respectively. To improve the performance of face recognition systems in adverse conditions (such as blurred low - resolution samples, improper illumination, etc), Li et al. [19] have proposed a new technique called as **L**earning the **C**ovariance **M**atrix **O**f **G**abor Wavelet (GW) (LCMoG). The Covariance Matrix, however, is disjoint from the Euclidean space, and therefore, Euclidean-based measures cannot be adopted directly. To address this, the authors propose two methods, one based on shallow CNN (called LCMoG-CNN) to project covariance matrix of GW into a feature vector of euclidean space, and the other based on matrix-logarithm (called LCMoG-LWPZ) which uses Whitening PCA to learn face features from the embedded covariance matrix. The recognition accuracies for LCMoG-CNN and LCMoG-LWPZ methods on Feret and Extended Yale-B datasets as reported is above 95% even under noisy environments. Further, the proposed models demonstrated a higher recognition accuracy of 96% through hybrid LCMoG-(LWPZ + CNN) on CMU MoBo and YouTube datasets. Fredj et al. [20] have developed a CNN framework based on aggressive data augmentation for face recognition in unconstrained environments. The authors have reported the robust performance of the proposed system in classifying noisy (face samples captured with higher noise content) and occluded face samples by using a deep face representation. The proposed model as reported demonstrates accuracies of 99.2% and 96.63% for LFW and

YTF datasets, respectively. Xie et al. [21] have reported a novel face recognition model that targets images having narrow spectral bands, often called hyperspectral face recognition. The authors put forward a modified version of the light CNN framework that is supported by transfer learning methodology. With this, the hyperspectral face samples could be projected into another subspace that has the capability to improve the classification accuracy of the proposed system. The proposed system as reported, sports classification accuracies of 92.83% (for PolyU), 95.12% (for CMU), and 99.73% (for UWA).

A CNN-based model for 3D face recognition was put forward by Dutta et al. [22]. The model works on 40 component faces generated by a combination of a mathematical model (4 components) and a data-level fusion technique (36 components) to project samples into a new space called ’complement component face space’. The model relies on extracting relevant features through a combination of SVD and fused through a crossover operation of a genetic algorithm based on hamming distance. Particle Swarm Optimization (PSO) is then used for discarding redundant features so that only the relevant features are selected, thereby improving the system performance. The proposed system, as reported demonstrates classification accuracy of 97.86%, 98.25%, and 99.89% for Frav3D, Bosphorus, and Texas3D datasets, respectively. Variability in the captured face samples degrades the performance of a face recognition system. In this regard, Meng et al. [23] have proposed a system called ’*MagFace*’ that works on an adaptive mechanism by sifting through easy and hard samples to avoid overfitting on noisy low-resolution samples. This consequently improves the face recognition in wild environments, and the proposed system sports verification accuracies of 92–99% on easy benchmarks, and 90–96% on difficult benchmarks. Qui et al. [24] on the other hand, focus on the generalization of face recognition systems in presence of real-world occluded face images. In this regard, the authors have proposed a single end-to-end DNN called Face Recognition with Occlusion Masks (FROM) which learns to discover corrupted features from Deep CNNs and clean them from dynamically learned masks. The proposed system, as reported exhibits classification accuracies of 96.22% for RMF2 and 98.32% for LFW-SM (Simulated Masks). A detailed review on the low - resolution face recognition systems [25, 26] gives insights into the different aspects of the face recognition system. These, however miss out on the most crucial aspect, that is the model explainability. Although there has been some investigation into the Explainable Face Recognition (XFR) [27–29], the model explainabiltiy for face recognition system has not gained much traction. In this regard, the work done in this manuscript is one such attempt towards XFR using LIME.

A close observation towards all the methods proposed for improving the face recognition systems as discussed above, reveals that the major focus of all the research groups has been towards improving the statistics of the proposed system. All the face recognition systems listed above were traditionally deployed like black boxes and did not indicate to the end-user the rationale behind these decisions. Answering "why" and "how" predictions are made, assists in understanding the behavior of the model. To elucidate this, an AI tool–LIME has been utilized to investigate the superpixels that have contributed to the black box for the classification of subjects. To the best of our knowledge, it has not been explained yet as to what features drive the black box in classifying a particular subject.

This paper is structured as follows. Section 2 presents the dataset description followed by a preliminary description of different Deep Neural Networks (DNN) used in this paper. Experimental setup and results are discussed in Sect. 3 and the explainability of models is discussed in Sect. 4 Finally, the paper is concluded in Sect. 5

## 2 Materials and Methods

This section presents the datasets used during the experiments and the deep neural networks used in XAI-FR framework. A brief description of the working of different DNN models employed and the working of LIME has been explained with a specific focus towards the explainability of black boxes in classifying a face sample.

### 2.1 Datasets Used

This section briefly summarizes the datasets used for experimentation.

#### 2.1.1 The Yale Dataset

Yale face database comprises 165 grayscale images of 15 distinct subjects [30]. Each subject has 11 face samples, one for each face expression (happy, normal, sad, sleepy, surprised, and wink) and configuration (center-light, with glasses, left-light, without glasses, right-light). An example of different face samples available in the Yale face database is shown in Fig. 1(a).

#### 2.1.2 The AT & T Dataset

The AT &T database originally known as ’The ORL Database of Faces’ comprising 400 grayscale images of 40 distinct subjects [31]. For each subject, there are 10 images that capture every possible combination of features. The face samples for each subject are available in PGM format. An example of different face samples available in the AT &T Face Database is shown in Fig. 1(b).

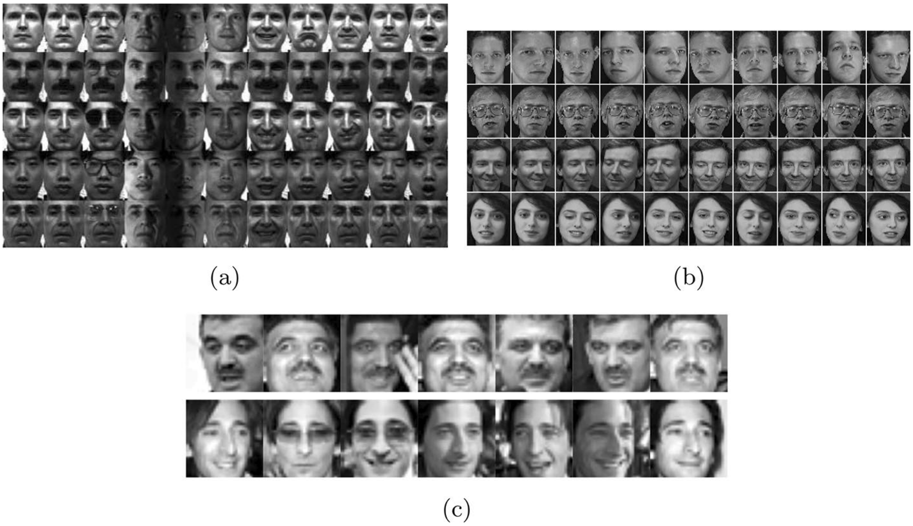

(a)
(b)
(c)

Fig. 1 Sample face images from **a** YALE, **b** AT &T, and **c** LFW Datasets

### 2.1.3 The LFW Dataset

LFW (Labeled Faces in the Wild) is a database of face images created to investigate the problem of unrestricted face recognition [32]. More than 13,000 photos of faces were gathered from the internet for the data collection. The collection contains 1680 subjects, each of which has two or more distinct photographs. In the present work, we considered only those subjects that have at least 70 face samples. An example of different face samples available in the LFW Face Database is shown in Fig. 1(c).

## 2.2 Methods

A Deep Neural Network [33, 34] is an artificial neural network [35] with several layers between the input and output layers. The subsections that follow briefly describe the variants of deep neural networks and explainable AI method – LIME used for the interpretability of the trained models..

### 2.2.1 LeNet-5 Model

The LeNet-5 model was proposed by LeCun et al. [36] for handwritten and machine-printed character recognition. This architecture is a simple multi-layer convolution neural network for the classification of images. A schematic of LeNet-5 model adapted from [36] is depicted in Fig. 2. Two convolutional and average pooling layers make up the LeNet-5 architecture. This is followed by two fully connected layers. Finally, a Softmax classifier is used which classifies images into respective classes.

### 2.2.2 AlexNet Model

The AlexNet model was proposed by Krizhevsky et al. [37] achieved a top-5 error rate of 15.3 % on the ImageNet LSVRC-2010 dataset comprising 1.2 million high-resolution images. The AlexNet is comparatively deeper as compared to its LeNet-5 counterpart. The schematic architecture of AlexNet is shown in Fig. 3. The AlexNet has 11 layers comprising five layers of convolutions layers and the subsequent three layers of max pooling. After convolution and max-pooling blocks, the architecture consists of 3 fully connected layers having RELU activation function, except in the last layer.

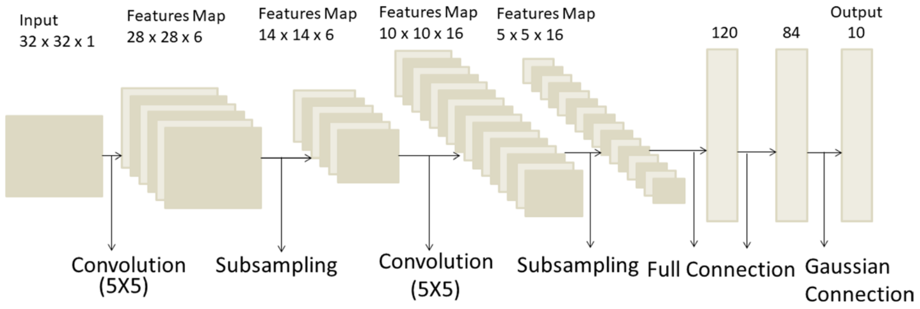


**Fig. 2** Depiction of LeNet-5 Architecture as adapted from [36]

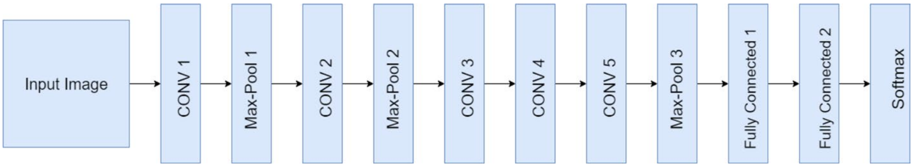


**Fig. 3** Depiction of AlexNet Architecture as adapted from [38]

### 2.2.3 Inception-V3 Model

Inception-V3 is the third variant of GoogLeNet used for image analysis and object detection. Inception-V3 scores over other CNN classifiers in terms of speed and accuracy. The previous models were just improving the performance and accuracy of their model but compromising the computational cost. To improve the system performance, the Inception-V3 relies on various tricks for optimizing its network. Szegedy et al. [39], had proposed several upgrades for the Inception-V3 model which increased the accuracy and reduced the computational complexity. These include optimizing the network, in order to loosen the constraints for easier remodeling by including factorized convolutions, regularization, dimension reduction, and parallelized computations. The architecture of an Inception-V3 network, as depicted in Fig. 4.

As can be observed from Fig. 4, the Inception-V3 architecture consists of a stem, comprising traditional pooling and convolutional layers. Subsequently, it comprises a pooling layer followed by fully connected and softmax layers. The Inception-V3 architecture also involves reduction modules that are designed for reducing the dimensions of the input. The architecture has about 24 million parameters and takes a default input of size $299 \times 299 \times 3$ .

### 2.2.4 VGG16 Model

Simonyan et al. [41] introduced VGG16, a CNN model that achieved 92.7 percent top-5 test accuracy in the ImageNet Dataset. The ImageNet comprises 14 million images belonging to 1000 different classes. It improves on AlexNet by successively replacing large kernel-sized filters of sizes 11 and 5 in the first and second convolutional layers with multiple kernel-sized filters of size 3.

In Fig. 5, the convolution layers using a non-linear activation function, known as rectified linear unit (ReLU), are represented by all the blue color rectangles. VGG16 comprises 13 convolution layers and 5 max-pooling layers. In addition to these, three green rectangles represent fully connected layers. Finally, there is an output layer which is a fully connected softmax output layer $\hat{y}$ with possible values corresponding to the number of classes.

### 2.2.5 LIME

Local Interpretable Model-agnostic Explanations, better known as LIME is an explainable AI method developed by Ribeiro et al. [42]. LIME can be used for a classifier model that classifies tabular data, pictures, or texts to better understand the behavior of the applied black-box classifier model. It is *'Local'*, meaning that LIME attempts to explain the

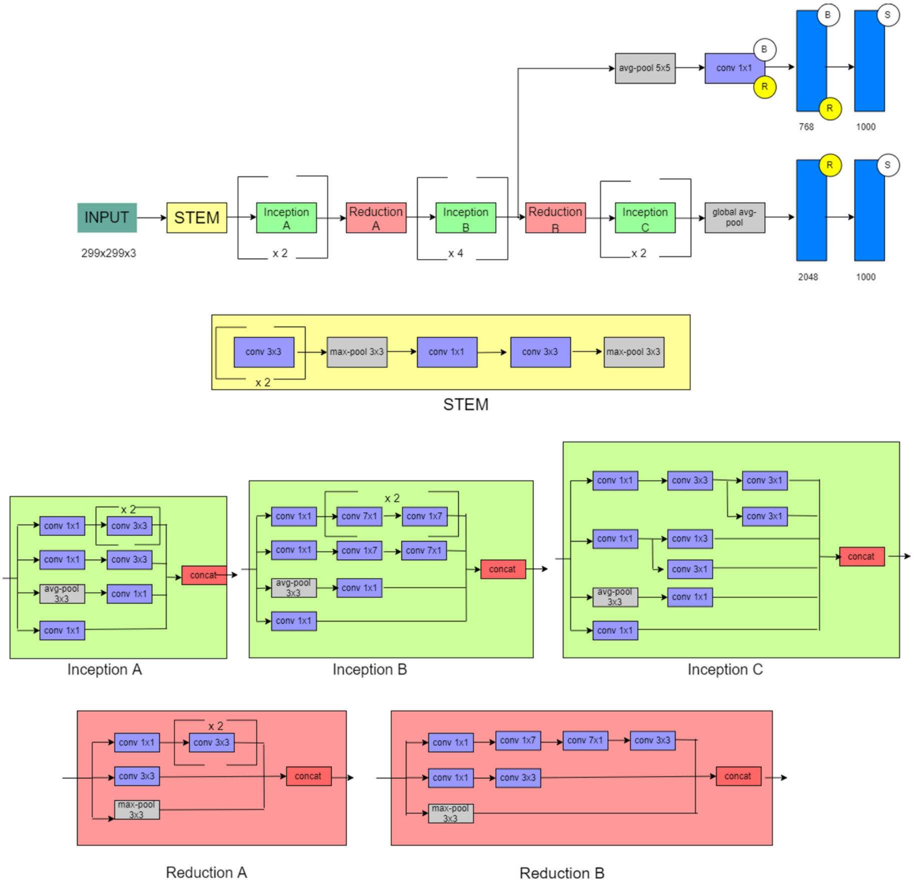


**Fig. 4** Schematic representation of a Inception -V3 Architecture as adapted from [40]

proposed black-box model by approximating the model's local linear behavior, and it is *'Interpretable'*, meaning that it provides a solution to understand why the model acts the way it does. The four steps involved in LIME:

1. **Input data permutation**: In this step, LIME generates several perturbed images similar to the input image by turning on and off some of the super-pixels of the image.
2. **Class prediction of each artificial image**: In this step, a class prediction for perturbed each artificially generated image is carried out using the trained model.
3. **Weight computation for each artificial image**: In this step, a weight is computed for each artificial image to measure its degree of importance. The distance is computed between every artificially generated image point and the corresponding points of the original input image. Using a kernel function, the distance metric value is mapped into a weight value between 0 to 1. The closer proximity of the perturbed instance to the instance being explained contributes to the higher associated weightage signifying its importance.

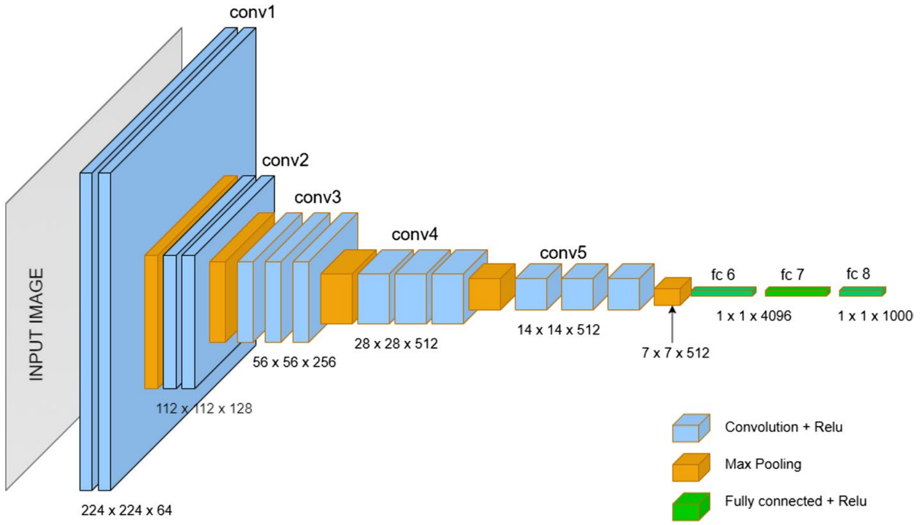


**Fig. 5** Schematic of a VGG16 Architecture as adapted from [41]

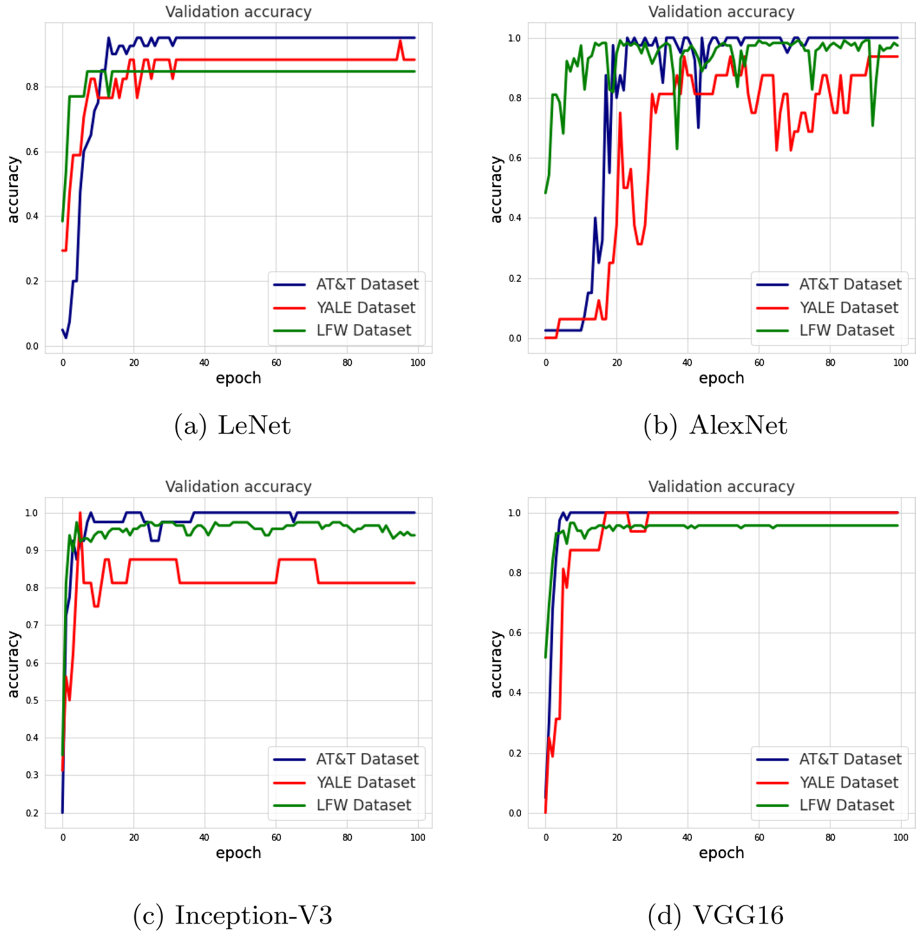
4. **Explaining important features by fitting a linear classifier**: This step involves fitting a linear regression model with the help of the weighted artificial data points. In this way, the fitted coefficient is obtained for each feature. On sorting based on coefficient values, the superpixels corresponding to higher coefficient values are the ones contributing significantly to the prediction of the black-box machine learning model.

## 3 Experimental Setup and Results

All experiments have been performed in Python 3.7 in the Google Colaboratory using runtime environment for NVIDIA Tesla K80 GPU. In order to test the applicability of deep neural networks (DNN) summarized in Sect. 2.2, the datasets mentioned in Section Sect. 2.1 were split in a ratio of 80:10:10 for realizing disjoint sets for training, validation, and test sets, respectively. The choice of hyperparameters for each DNN is based on the exploration of search space. The batch size is set to 32, the learning rate equals 0.001, and the optimizer employed is Adam optimizer [43, 44].

### 3.1 Results and Discussions

In this section, we present the results of employing LeNet, AlexNet, Inception-V3, and VGG16 on the three face datasets mentioned above. The plots depicted in Fig. 6 shows the variation of validation accuracy of the face recognizers based on different deep neural networks with respect to the number of epochs for each of the three datasets. The LeNet based face recognizer (Fig. 6(a)) shows fluctuations till $epoch = 30$ after which it almost stabilizes. The AlexNet based face recognizer (Fig. 6(b)) shows poor generalization performance on the unseen face samples. Likewise, Inception-V3 based face recognizer



(a) LeNet

(b) AlexNet

(c) Inception-V3

(d) VGG16

**Fig. 6** Plots depicting variation in validation accuracy w.r.t. number of epochs for three datasets namely, AT &T, Yale, and LFW for the face recognizers based on **a** LeNet, **b** AlexNet, **c** Inception-V3, and **d** VGG16

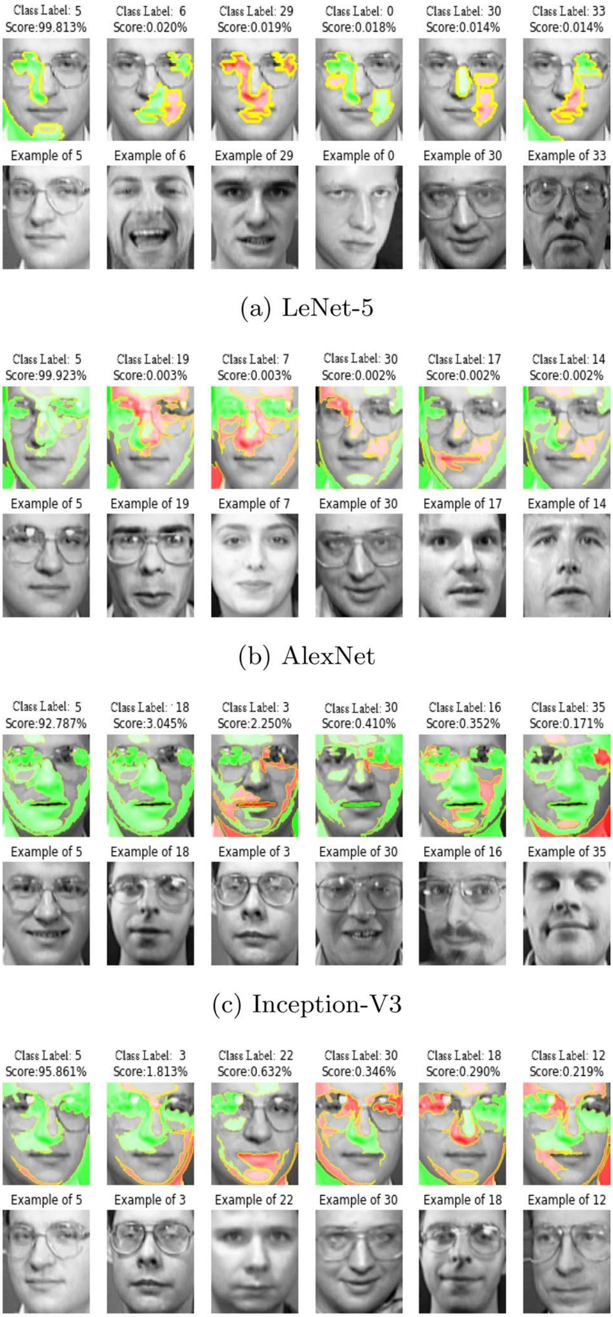
(Fig. 6(c)) is unable to stabilize with an increase of number of epochs. The VGG16 based face recognizer stabilizes after 15 epochs and (Fig. 6(d)) shows the best generalization capability. The classification performance of the above-mentioned four deep neural networks on different datasets is given in Table 1. We note that VGG16 yields consistently the best performance across the chosen datasets in terms of classification accuracy, recall, precision, and F1-Score.

## 4 Model Explainability

In this section, we evaluate the explainability of each of the four DNN-based face recognition models. Towards this end, LIME has been used to mark the superpixels that have contributed towards the classification label generated for a particular subject. The regions shown in green color have contributed positively for the predicted label and the regions

**Table 1** Comparison of Performance Metrics of DNN Models on Different Datasets

| Metrics | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Models | LeNet-5 | | | AlexNet | | | Inception-V3 | | | VGG-16 | | |
| | Yale | AT &T | LFW | Yale | AT &T | LFW | Yale | AT &T | LFW | Yale | AT &T | LFW |
| Accuracy (%) | 94.11 | 95.00 | 90.00 | 94.11 | 97.50 | 94.00 | 88.23 | 97.50 | 93.70 | 94.11 | 100.00 | 97.00 |
| Recall (%) | 95.00 | 90.00 | 89.00 | 95.00 | 92.00 | 92.00 | 82.00 | 99.00 | 92.00 | 95.00 | 100.00 | 94.00 |
| Precision (%) | 97.00 | 92.00 | 89.00 | 95.00 | 92.00 | 92.00 | 88.00 | 99.00 | 92.00 | 95.00 | 100.00 | 97.00 |
| F1-Score (%) | 95.00 | 91.00 | 89.00 | 93.00 | 92.00 | 89.00 | 83.00 | 99.00 | 92.00 | 93.00 | 100.00 | 95.00 |



(a) LeNet-5

(b) AlexNet

(c) Inception-V3

(d) VGG16

**Fig. 7** LIME-generated explanations for a correctly predicted face (True label: 5) using **a** LeNet-5, **b** AlexNet, **c** Inception-V3, and **d** VGG16. Each sub-figure shows explanations generated for the best six matches along with prediction score

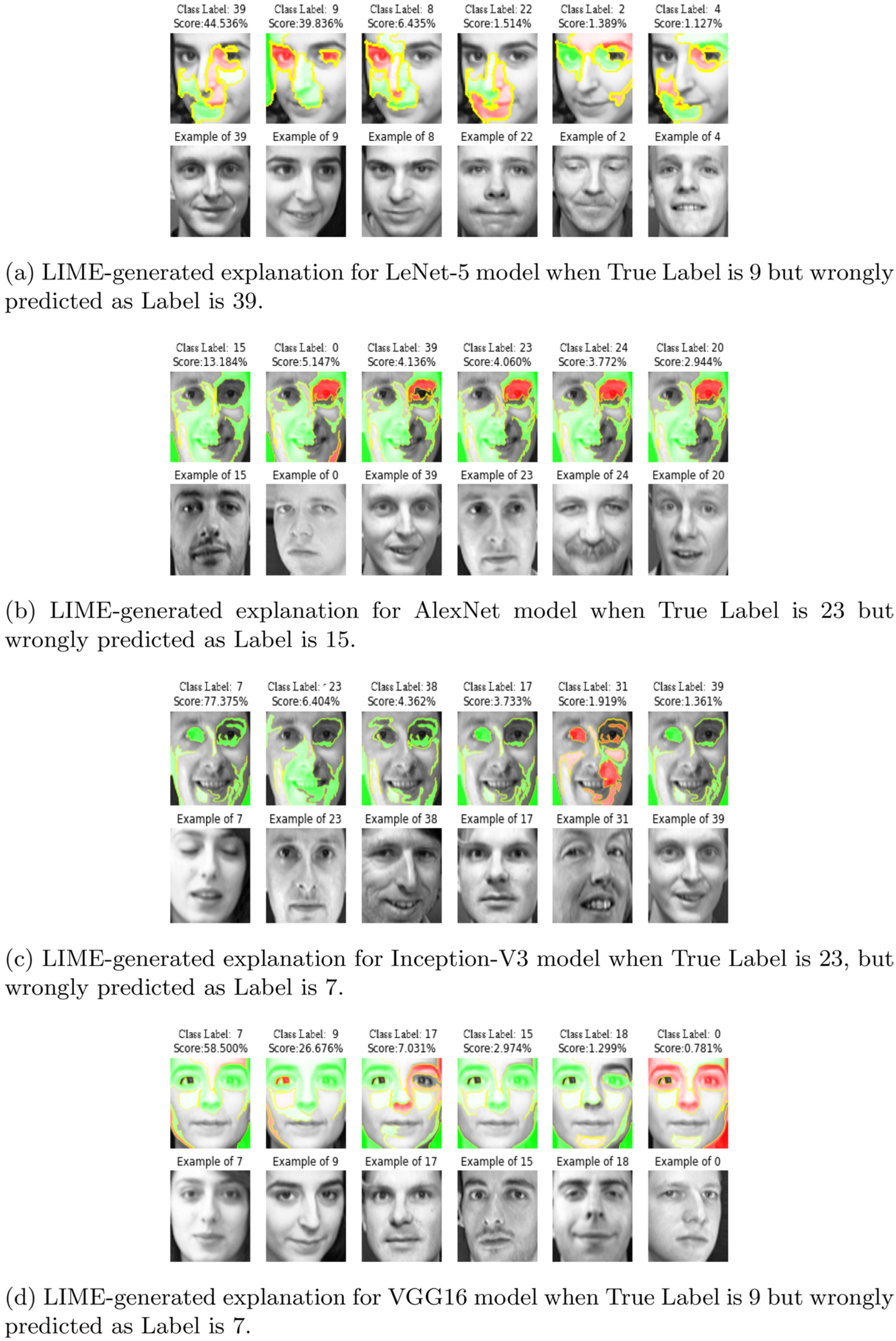
shown in red color have contributed negatively to the predicted label. Out of 40 test samples from the AT &T database, we randomly picked an image (true label 5) for which each of the four models predicted the correct label. Figure 7 shows the prediction scores of top six matches for each of the LeNet-5, Alex-Net, Inception-V3, and VGG16 models. As expected, for each model, the prediction score of the correctly predicted subject (99.81%, 99.92%, 92.78%, and 95.86% for the LeNet-5, AlexNet, Inception-V3, and VGG16 model respectively) that appears leftmost in a row is significantly higher than the other images that appear in the same row. Each sub-figure shows explanations generated for the best six matches. It may be noted that even though each of the models predicts the correct label, it focuses on somewhat different features for generating its prediction. As we could not find an image for which each of the models predicted a wrong label, Fig. 8 depicts different instances predicted wrongly corresponding to the models LeNet-5, ALexNet, Inception-V3, and VGG16 respectively. As the models output a wrong label, we note that in Fig. 8, the prediction score of the true label is lower than the best match that resulted in a wrong prediction.

### 4.1 Subjective Assessment of LIME generated Explanations

For subjective assessment of the LIME-generated explanations, we randomly selected 20 samples from the test dataset. Each image was passed to the LeNet, AlexNet, Inception-V3, and VGG16 models for recognizing the person in the image. For each image, the explanation generated by LIME for each of the four deep learning models was shown to a group of twenty volunteers. Each volunteer was asked to score the models on a scale of 4, Thus, given an image, a volunteer would assign a score of 4 to the model for which LIME generated the most comprehensible explanation as per his/her judgement and a score of 1 to the model for which the generated explanation was least comprehensible. The responses were collected using a Google Form. Based on the responses of 20 volunteers, LIME explanations using the VGG16 model ranked highest with an average score of 3.35, followed by Alex-Net, Inception-V3, and LeNet, having average scores 3.02, 2.87, and 1.75 respectively.

## 5 Conclusions and Future Scope

In this paper, we have examined the interpretability of four deep neural networks (DNN) models, namely, LeNet-5, AlexNet, Inception-V3, and VGG16 on the AT &T dataset. For this purpose, we used Local Intepretable Model-Agnostic Explanations (LIME) as the explanation model to mark the visually significant features in terms of the superpixels. Based on an experimental study involving twenty volunteers, we found that that the explanations generated for the classification performed by VGG16 were significantly more explainable than those produced for the other models. Furthermore, the LIME-generated superpixels on face images correspond to the region of non-interest (RONI) comprising background features. RONI can be segmented in future work so that these insignificant features do not influence the interpretability XAI methods.


Class Label: 39 Score:44.536%
Class Label: 9 Score:39.836%
Class Label: 8 Score:6.435%
Class Label: 22 Score:1.514%
Class Label: 2 Score:1.389%
Class Label: 4 Score:1.127%
Example of 39
Example of 9
Example of 8
Example of 22
Example of 2
Example of 4

(a) LIME-generated explanation for LeNet-5 model when True Label is 9 but wrongly predicted as Label is 39.

Class Label: 15 Score:13.184%
Class Label: 0 Score:5.147%
Class Label: 39 Score:4.136%
Class Label: 23 Score:4.060%
Class Label: 24 Score:3.772%
Class Label: 20 Score:2.944%
Example of 15
Example of 0
Example of 39
Example of 23
Example of 24
Example of 20

(b) LIME-generated explanation for AlexNet model when True Label is 23 but wrongly predicted as Label is 15.

Class Label: 7 Score:77.375%
Class Label: ' 23 Score:6.404%
Class Label: 38 Score:4.362%
Class Label: 17 Score:3.733%
Class Label: 31 Score:1.919%
Class Label: 39 Score:1.361%
Example of 7
Example of 23
Example of 38
Example of 17
Example of 31
Example of 39

(c) LIME-generated explanation for Inception-V3 model when True Label is 23, but wrongly predicted as Label is 7.

Class Label: 7 Score:58.500%
Class Label: 9 Score:26.676%
Class Label: 17 Score:7.031%
Class Label: 15 Score:2.974%
Class Label: 18 Score:1.299%
Class Label: 0 Score:0.781%
Example of 7
Example of 9
Example of 17
Example of 15
Example of 18
Example of 0

(d) LIME-generated explanation for VGG16 model when True Label is 9 but wrongly predicted as Label is 7.

**Fig. 8** LIME-generated explanations for the subjects with true labels 9, 23, 23, and 7 predicted wrongly by the models **a** LeNet-5, **b** AlexNet, **c** Inception-V3, and **d** VGG16 respectively

**Acknowledgements** The authors would like to thank University of Delhi for providing the tools for the completion of this work.

**Funding** This research received no specific grant from any funding agency.

**Data Availability** The authors confirm that the data and material supporting the finding of this study are available within the article.

**Code Availability** The codes that support the finding of this study are available from the corresponding author upon reasonable request.

## Declarations

**Conflict of interest** The authors declare that there is no conflict of interest.

## References

1. Dodson, C., Soldera, J., & Scharcanski, J. (2021). Some information geometric aspects of cyber security by face recognition. *Entropy, 23*(7), 878.
2. Juneja, K., & Rana, C. (2021). An extensive study on traditional-to-recent transformation on face recognition system. *Wireless Personal Communications, 118*(4), 3075–3128.
3. Loey, M., Manogaran, G., Taha, M. H. N., & Khalifa, N. E. M. (2021). A hybrid deep transfer learning model with machine learning methods for face mask detection in the era of the COVID-19 pandemic. *Measurement, 167*, 108288.
4. Petpairote, C., Madarasmi, S., & Chamnongthai, K. (2021). 2d pose-invariant face recognition using single frontal-view face database. *Wireless Personal Communications, 118*(3), 2015–2031.
5. Zou, X., Kittler, J. & Messer, K. (2007). "Illumination invariant face recognition: A survey." In *2007 first IEEE international conference on biometrics: theory, applications, and systems*, pp. 1–8, IEEE.
6. Sharma, S., & Kumar, V. (2021). Performance evaluation of machine learning based face recognition techniques. *Wireless Personal Communications, 118*(4), 3403–3433.
7. Azeem, A., Sharif, M., Raza, M., & Murtaza, M. (2014). A survey: Face recognition techniques under partial occlusion. *Int. Arab J. Inf. Technol., 11*(1), 1–10.
8. Tang, L., Lu, H., Pang, Z., Li, Z., & Su, J. (2019). A distance weighted linear regression classifier based on optimized distance calculating approach for face recognition. *Multimedia Tools and Applications, 78*(22), 32485–32501.
9. Damale, R.C. & Pathak, B.V. (2018). "Face recognition based attendance system using machine learning algorithms." In *2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS)*, pp. 414–419, IEEE.
10. Abuzneid, M. A., & Mahmood, A. (2018). Enhanced human face recognition using LBPH descriptor, multi-KNN, and back-propagation neural network. *IEEE access, 6*, 20641–20651.
11. VenkateswarLal, P., Nitta, G.R. & Prasad, A. (2019). "Ensemble of texture and shape descriptors using support vector machine classification for face recognition." *Journal of Ambient Intelligence and Humanized Computing*, pp. 1–8.
12. Qu, X., Wei, T., Peng, C. & Du, P. (2018). "A fast face recognition system based on deep learning." In *2018 11th International Symposium on Computational Intelligence and Design (ISCID)*, vol. 1, pp. 289–292, IEEE.
13. Aiman, U. & Vishwakarma, V.P. (2017). "Face recognition using modified deep learning neural network." In *2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, pp. 1–5, IEEE.
14. Coşkun, M., Uçar, A., Yildirim, Ö. & Demir, Y. (2017). "Face recognition based on convolutional neural network." In *2017 International Conference on Modern Electrical and Energy Systems (MEES)*, pp. 376–379, IEEE.
15. Görgel, P., & Simsek, A. (2019). Face recognition via deep stacked denoising sparse autoencoders (DSDSA). *Applied Mathematics and Computation, 355*, 325–342.
16. Zhao, F., Li, J., Zhang, L., Li, Z., & Na, S.-G. (2020). Multi-view face recognition using deep neural networks. *Future Generation Computer Systems, 111*, 375–380.

17. William, I., Rachmawanto, E.H., Santoso, H.A., Sari, C.A., *et al.* (2019). "Face recognition using FaceNet (survey, performance test, and comparison)." In *2019 Fourth International Conference on Informatics and Computing (ICIC)*, pp. 1–6, IEEE.
18. Wang, P., Wang, P., & Fan, E. (2021). Violence detection and face recognition based on deep learning. *Pattern Recognition Letters, 142*, 20–24.
19. Li, C., Huang, Y., Huang, W., & Qin, F. (2021). Learning features from covariance matrix of Gabor wavelet for face recognition under adverse conditions. *Pattern Recognition, 119,* 108085.
20. Ben Fredj, H., Bouguezzi, S., & Souani, C. (2021). Face recognition in unconstrained environment with CNN. *The Visual Computer, 37*(2), 217–226.
21. Xie, Z., Niu, J., Yi, L., & Lu, G. (2021). Regularization and attention feature distillation base on light CNN for hyperspectral face recognition. *Multimedia Tools and Applications, 81*(14), 1–17.
22. Dutta, K., Bhattacharjee, D., Nasipuri, M., & Krejcar, O. (2021). Complement component face space for 3D face recognition from range images. *Applied Intelligence, 51*(4), 2500–2517.
23. Meng, Q., Zhao, S., Huang, Z. & Zhou, F. (2021). "Magface: A universal representation for face recognition and quality assessment." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14225–14234.
24. Qiu, H., Gong, D., Li, Z., Liu, W. & Tao, D. (2021). "End2End Occluded Face Recognition by Masking Corrupted Features." *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
25. Wang, Z., Miao, Z., Jonathan Wu, Q., Wan, Y., & Tang, Z. (2014). Low-resolution face recognition: A review. *The Visual Computer, 30*(4), 359–386.
26. Goh, K. M., Ng, C. H., Lim, L. L., & Sheikh, U. U. (2020). Micro-expression recognition: an updated review of current trends, challenges and solutions. *The Visual Computer, 36*(3), 445–468.
27. Williford, J.R., May, B.B. & Byrne, J. (2020). "Explainable face recognition." In *European Conference on Computer Vision*, pp. 248–263, Springer.
28. Franco, D., Navarin, N., Donini, M., Anguita, D., & Oneto, L. (2022). Deep fair models for complex data: Graphs labeling and explainable face recognition. *Neurocomputing, 470*, 318–334.
29. Fu, B. & Damer, N. (2022). "Explainability of the implications of supervised and unsupervised face image quality estimations through activation map variation analyses in face recognition models." In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 349–358.
30. "Yale Face Database." http://vision.ucsd.edu/content/yale-face-database.
31. "AT &T Face Database." http://www.cl.cam.ac.uk/Research/DTG/attarchive:pub/data/att.
32. Huang, G.B., Mattar, M., Berg, T. & Learned-Miller, E. (2008). "Labeled faces in the wild: A database forstudying face recognition in unconstrained environments." In *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*.
33. Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural networks, 2*(5), 359–366.
34. Miikkulainen, R., Liang, J., Meyerson, E., Rawal, A., Fink, D., Francon, O., Raju, B., Shahrzad, H., Navruzyan, A., Duffy, N., *et al.* (2019). "Evolving deep neural networks." In *Artificial intelligence in the age of neural networks and brain computing*, pp. 293–312, Elsevier.
35. Hopfield, J. J. (1988). Artificial neural networks. *IEEE Circuits and Devices Magazine, 4*(5), 3–10.
36. LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE, 86*(11), 2278–2324.
37. Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems, 25*, 1097–1105.
38. "Getting Staarted with AlexNet." https://www.geeksforgeeks.org/ml-getting-started-with-alexnet/.
39. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. & Wojna, Z. (2016). "Rethinking the inception architecture for computer vision." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826.
40. "Illustrated: 10 CNN Architectures." https://towardsdatascience.com/illustrated-10-cnn-architectures-95d78ace614d.
41. Simonyan, K. & Zisserman, A. (2014). "Very deep convolutional networks for large-scale image recognition," *arXiv preprint*arXiv:1409.1556.
42. Ribeiro, M.T., Singh, S. & Guestrin, C. (2016). "" 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*, pp. 1135–1144.
43. Zhang, Z. (2018). "Improved adam optimizer for deep neural networks." In *2018 IEEE/ACM 26th International Symposium on Quality of Service (IWQoS)*, pp. 1–2, IEEE.
44. Lydia, A., & Francis, S. (2019). Adagrad-An optimizer for stochastic gradient descent. *Int. J. Inf. Comput. Sci., 6*(5).

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Ankit Rajpal** is currently working as an Assistant Professor at Department of Computer Science, University of Delhi. His research interests include image and video watermarking, machine learning, and Data Mining. He has published several papers in reputed international journals and conferences.

**Khushwant Sehra** is currently working as a Research Scholar with the Department of Electronic Science, University of Delhi. His research interests include modeling, simulation and fabrication of GaN based HEMT devices. He has worked on image processing, including digital image watermarking and development of facial recognition systems for uncontrolled environments.

**Rashika Bagri** received her Master's Degree in Computer Science from University of Delhi in July, 2021 and Bachelors in Computer Science from University of Delhi. She is currently working as a research scholar in the Department of Computer Science, University of Delhi. Her areas of Interests include Machine Learning, Deep learning and Computer Vision.

**Pooja Sikka** received her Bachelor's in Computer Science from SGGSCC, University of Delhi followed by Masters in Computer Science from Department of Computer Science, University of Delhi. She is currently working in EXL services as a business analyst.