A LIME-Explained VGG16 Model for Disguise and Makeup Face Recognition in Forensics

Abdelkarim Khedher[1] ; Jaouhar Fattahi[2] ; Mohamed Mejri[2] ; Ridha Ghayoula[3] and Lassaad Latrach[1]

Abstract— **This paper exploits the rise of artificial intelligence (AI) and deep learning (DL) to improve the use of digital forensic evidence analysis, specifically criminal identification from facial images despite disguise and makeup. Our approach leverages the VGG16 architecture for face recognition and identification, coupled with the LIME framework (Local Interpretable Model-Agnostic Explanations) to explain model recognition. This combination enables interpretation and verification of results with enhanced trust and confidence in forensic analysis. We follow a "watch and iterate" procedure, utilizing the insights generated from LIME to curate the training dataset, improving the model's performance iteratively. The efficacy of this procedure is reflected in the remarkable outcomes: our model has an accuracy of 98.10%, precision of 98.16%, recall of 98.10%, F1-score of 98.11%, AUC of 100%. This development in forensic technology has great potential to enhance the precision and speed of criminal identification, thus leading to safer and fairer societies.**

Index Terms— **Deep Learning, Digital Forensic Evidence, Face Recognition, VGG16, XAI.**

I. Introduction

Identification and recognition of criminals based on their faces, even under disguise or makeup, plays an important role as forensic evidence in the investigation and judgement process [1]–[4]. This evidence allows during the investigation to link suspects to crime scenes [5], [6] [7]. However, with the rise of technology and the multiplication of digital devices, the digitization of forensic evidence and its justification represent a challenge and have become more complicated to be acceptable as forensic evidence [8] and law enforcement. Traditional methods of face recognition find it difficult to support complex models for the complex patterns in facial data. Today, with the emergence of artificial intelligence (AI) and deep learning techniques [9], new perspectives have been opened and offer a new dimension to this challenge to improve facial identification [10], many sophisticated models such as convolutional neural networks (CNN) [11] [12] [13] have been used to perform visual recognition tasks and provide solid forensic evidence. Among these algorithms, VGG16 [14] has proven effective for the extraction of facial characteristics, but it is essential to understand the context of the application of this model, such as the identification of criminals based on their faces and the challenge to face, where the data set can be small and biased [15]. However, the black-box nature of deep-learning to make decision and facial recognition poses an obstacle to gaining trust for forensic evidence, which without this confidence it can not be admissible in court [16]. In order to deal with this constraint, Explainable AI (XAI) techniques such as Local Interpretable Model-agnostic Explanations (LIME) algorithm have been introduced to explain and understand decisions [17]–[21]. In our context of face recognition with VGG16 model, LIME provides information about the decision made for complex models with locally interpretation of prediction by highlighting the positive and negative facial feature that contributing or perturbing most into model decision thereby improving transparency and trustworthiness. In this paper, we explore the integration of the VGG16 deep learning algorithm with the LIME explainability framework in purpose to improve criminals face recognition and identification in digital forensic process. LIME will be used to explain and understand VGG16 decision, and improve iteratively dataset between two experiments by identify dataset gaps by focusing on poorly represented segments or problematic areas and highlighting the weakness in limited datasets. With our experiment we aim not only improve accuracy of criminals recognition but also improve confidence of model's decision by exposing interpretation of disguised and simple faces which crucial in legal and forensic context. The experimental results show that this approach significantly improves the accuracy of identification, even in complex scenarios. However, some limitations remain, including the difficulty of processing partially visible or poorly lit faces

II. Context and work strategy

The research context is the recognition and identification of criminals faces even under disguise or makeup and provide a solid and trustable digital forensic evidence. During our experiments, we use VGG16 model for training and images classification and LIME Framework to explain recognition results and understand model quality. As mentioned before, we conduct two distinct experiments: the first one, using the unmodified VGG16 model, and the second one will be built upon the results of the first experiment to improve model's metrics and the quality of our dataset.

A. VGG16 model

VGG16 is one of the many algorithms based on the Convolutional Neural Network (CNN) architecture, it is a deep

[1] Abdelkarim Khedher and Lassaad Latrach are with the Ecole Nationale des Sciences de l'informatique, Université de la Manouba, Tunis, Tunisia. Khedher_Karim@yahoo.fr; Lassaad.Latrach@ensi-uma.tn

[2] Jaouhar Fattahi and Mohamed Mejri are with the Department of Computer Science and Software Engineering, Laval University, Quebec, Canada. Jaouhar.Fattahi.1@ulaval.ca; Mohamed.Mejri@ift.ulaval.ca

[3] Ridha Ghayoula is with the Faculty of Engineering, University of Moncton, New Brunswick, Canada. Ridha.Ghayoula@umoncton.ca

and Transfer Learning model designed to process visual data making it effective for facial recognition tasks. It has been widely used in various studies, demonstrating high accuracy in identifying facial features and expressions. VGG16 is composed of 21 layers including 16 layers with weights, hence the 16 in its name come from, thirteen convolutional layers, five Max Pooling layers, and three Dense layers Fig. 1 [22]

.

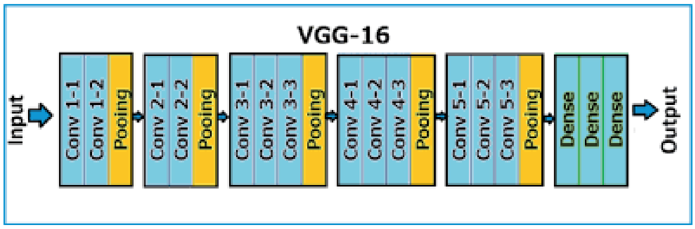


Fig. 1: VGG16 Architecture

B. Metrics

Performance metrics allow us to evaluate effectiveness and interpret the model [23]. Here are the main ones we used:

- Accuracy : is the ratio of the number of correct predictions to the total number of predictions 1. This is a global metric, but it can be misleading if classes are unbalanced.

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}} \tag{1}$$

- Precision : Measures the proportion of true positives among positive predictions 2.High precision indicates that the predicted class is generally correct.

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \tag{2}$$

- Recall : measures the proportion of true positives among samples that really belong to the class 3. A high recall means that the model usually captures samples from the positive class.

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \tag{3}$$

- F1-Score : is the harmonic mean of precision and recall. It helps detect imbalance between classes 4. A value close to 1 indicates a good balance between precision and recall.

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

- Confusion Matrix : It shows the number of true positives, true negatives, false positives, and false negatives of each class. Helps to understand where the model is wrong in prediction
- AUC : measures the model's ability to distinguish between classes. An AUC of 0.5 indicates that the model does not better than a random draw, while an AUC of 1.0 indicates perfect classification.

C. Local Interpretable Model-agnostic Explanations LIME

To interpret VGG16's facial recognition results, LIME (Local Interpretable Model-agnostic Explanations) represent a powerful framework that convert the black box of deep learning into understandable insights [24] [25]. LIME explanations integrated with VGG16 model can achieve high interpretability scores in facial identification and recognition tasks. The process consist to applying segmentation on facial images into ”superpixels” and interpreting how these segments influence the model's decisions. For understand a face classification, LIME highlights crucial facial features by generating variations of the images and turning specific regions on or off, effectively revealing which areas of the face most strongly influence the VGG16 model's predictions. This technique demonstrate valuable in facial recognition systems, where LIME can determinate whether the model is focusing on the right facial features like eyes, nose, and mouth contours, or if it's confused by irrelevant background elements such as disguise with accessories or makeup. Through this interpretation, user can validate if the VGG16 model make decisions based on meaningful facial characteristics or rather than arbitrary or biased patterns in the training data such as described in Fig. 2. In our case, other than its explainability function LIME will be used to understand the quality and weak point of our dataset between two experiments and iteratively improve it [26].

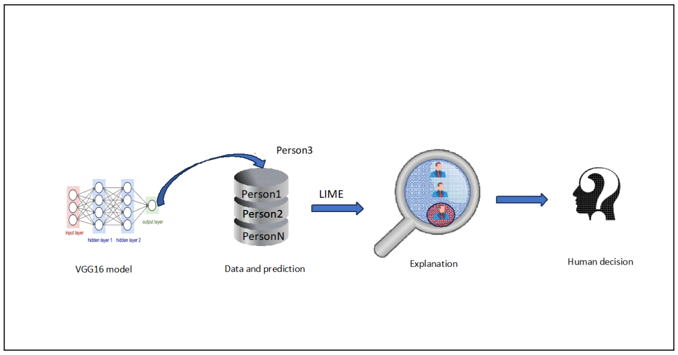


Fig. 2: Explaining person predictions

III. Experiment

As already mentioned, we will proceed with the method of Evaluate and Iterate. Continuously evaluate the model performance on a validation set. We use training and validation graph for loss and accuracy to understand model behavior. Metrics like accuracy, precision, recall, F1-Score, and AUC help to assess effectiveness and LIME to provide visual explaining by projecting superpixels and draw boundaries into test pictures to see the segment and features that help to make decision and help to identify necessary adjustments in model and dataset.

A. First experiment

1) Data set: We started with a dataset consisting of 1,400 images, equally distributed across seven classes representing seven celebrities named from celebrity one to seven, with 200 images per class downloaded from [27]. To realistically augment our dataset, we utilized the OpenCV library to apply optical distortion techniques, including deformations and warp transformations. As a result, the dataset was expanded to 600 images per celebrity. The expanded dataset was then automatically split using the ImageDataGenerator from the TensorFlow library into 70% for training, 20% for validation, and 10% for testing. The dataset includes various facial positions and expressions. In this study, we focus exclusively on simple facial images in different position without disguises such as glasses or hats, and without makeup, as illustrated in Fig. 3, to evaluate the model's capacity for face identification under conditions free of accessory disguises.

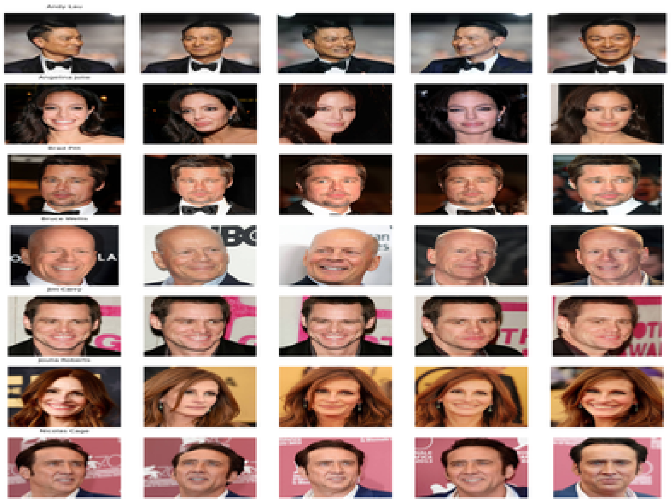


Fig. 3: Sample of the used dataset [27]

2) First experiment process: The experiment p is represented by Fig. 4. Whose first layer represents the Input in which we do image augmentation by transformation and resizing to the fixed size accepted by VGG16 algorithm (224,224). After loading pre-trained VGG16 model we freeze training layer to customize it, the output will be followed by Flatten layer and dense layers with Relu activation, a Dropout layer with a 0.5 rate to prevent over fitting and reducing co-adaptation is used before output when we use Softmax activation with a number of deduced classes. The default learning rate is used in model compilation with Adam optimizer. We train a model with 30 epochs and use EarlyStopping callback with a patience parameter fixed at 3 if the training does not improved. In the end we finish with a model evaluation and extraction of performance metrics and application of the LIME explainability algorithm to evaluate and understand prediction and data set in order to improve it in the second experiment and make decisions.

3) Results and discussion: The results of the first experiment present poor performances, suggesting that the model is not well perform in classifying faces, and it's unable to identify all faces belonging to a given class. The model is able to differentiate some classes but not all accurately.

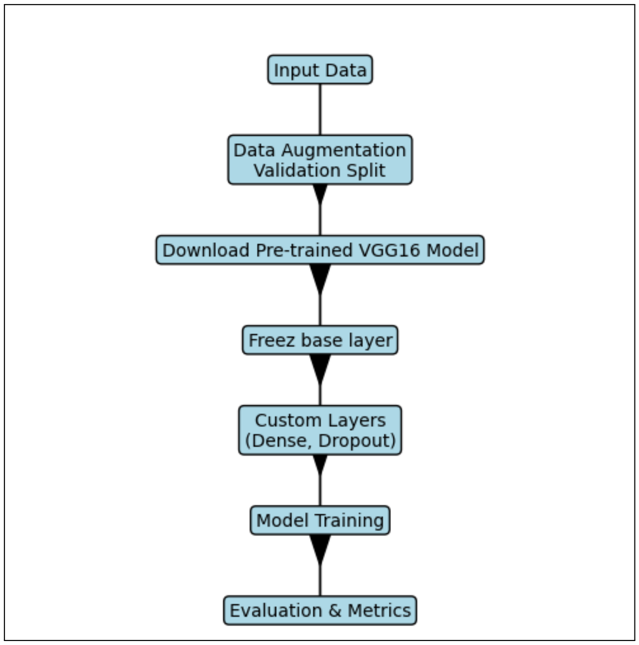


Fig. 4: Experiment Process

Fig. 5 and Fig. 6 show, respectively, matrix confusion, model accuracy and loss evolution over epochs, indicating that the model performance on a validation dataset has stopped improving and no significant improvement is observed in a set number of epochs. Like seen in both training and validation accuracy, the significant gap between performance on training and validation data explain that the model is too complex for the dataset and suggesting that the model is overfitting to the training data. The model learns details specific to the training data and does not generalize well to new examples. The LIME explanation is shown in data table of figures I where the yellow bound represents positive segment helping model making decision. Like seen in the first sample there is a problem with shadow and picture luminosity. Samples two, three and four shows that there are several off-face areas considered for decision making, so faces need to be cropped more in the preprocessing of dataset. Disguises and make-ups are ignored for decision making as is clear in the last four examples which represents a positive point of model that helps to predict and identify faces even under disguise and make-up.

B. Second experiment

Based on the performance metrics of the first experiment and the LIME explainability, we note that the data set needs to be reprocessed and modifications made to the general experiment process.

1) Preprocessing Dataset: : To increase data set diversity we use OpenCV library to add accessories like glasses and hat on some images. After that, we proceeded to severely crop the faces, detect blurry images and adjust the brightness

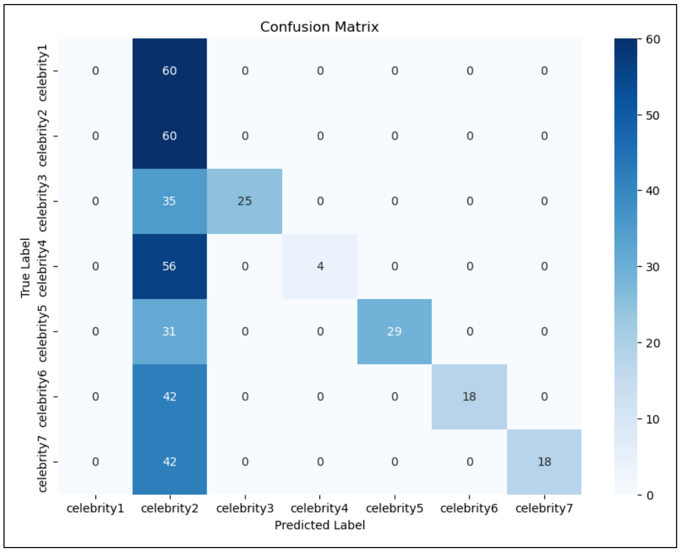


Fig. 5: Confusion matrix of first iteration

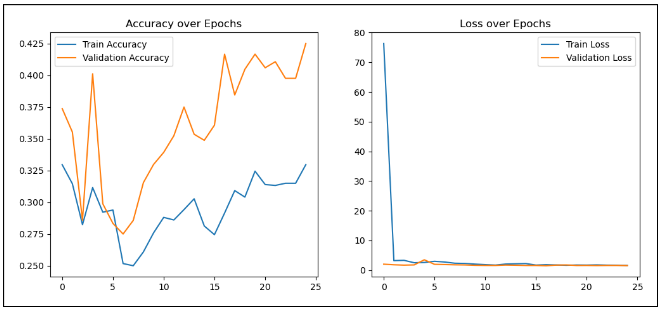


Fig. 6: First Iteration Accuracy and Loss evolution

and contrast of the images by performing a linear transformation on each pixel of the image [28] according to the following Equation 5

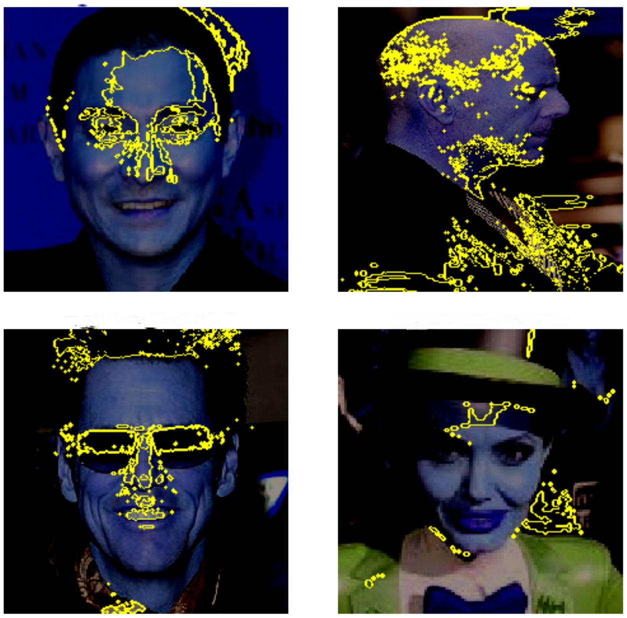
$$g(i,j) = \alpha \cdot f(i,j) + \beta \quad (5)$$

where (f(i, j)) is the value of the input pixel and (g(i, j)) is the value of the output pixel using the cv2.convertScaleAbs function of OpenCV

2) Second experiment process: To improve the model and process we combined several techniques

- Hyperparameter optimization : We decrease the learning rate to 0.0005 and we use a scheduler to dynamically adjust this rate during training progress.
- Regularization : We use L2 regularization to prevent overfitting and improve model generalization.
- Validation process : To mitigate the impact of imbalanced data and provide reliable validation results, we use K-Fold cross-validation [29] by evaluating the model learning performance by dividing the dataset into 5 equal-sized subsets called folds. Firstly,the dataset is randomly divided into 5 equal-sized subsets, then for each fold, the model is trained on k-1 folds and tested on the remaining fold. Finally, the average of the performance scores obtained in each fold is calculated and used to evaluate the performance of the model.

TABLE I: LIME explanability to understand prediction and dataset quality [27]



3) Results and discussion:

- Performance Metrics: The analysis of the results of the first experiment in terms of performance metrics, graphs and LIME's explanation of the quality of the data set allowed us to adjust the model and processed the images as mentioned in the previous section. This improvement of performance metrics is significantly clear as presented in the comparative table II. As shown also, the second experiment presents an improvement in evaluation time, a shorter time indicates a more efficient model, capable of performing predictions quickly. This is crucial for our context where processing speed is critical.
  Fig. 7 shows the evolution of accuracy and loss over training and validation suggest that no more overfitting and the model perform linearly. Classification report shown in Table.III, which summarizes model performance on the entire test data, and confusion matrix shown in Fig. 8, show that our model performs well on all classes, with excellent accuracy and recall compared to the first experiment.
- Faces recognition and LIME explainability: To test the ability of our model to provide reliable forensic evidence for the identification of criminals from their faces even under disguise and makeup, we used images of celebrities with accessories such as glasses, hats and caps or makeup as seen in the table of figures IV. Each pair of images presents the original image and the recognition result with the LIME explanation by super-

TABLE II: Comparison of Performance metrics between the two experiments

Metrics	First experiment	Second experiment
accuracy (%)	36.67	98.10
precision (%)	74.07	98.16
recall (%)	36.67	98.10
f1-score (%)	37.13	98.11
AUC	70.00	100
Evaluation time (sec)	57.97	45.58

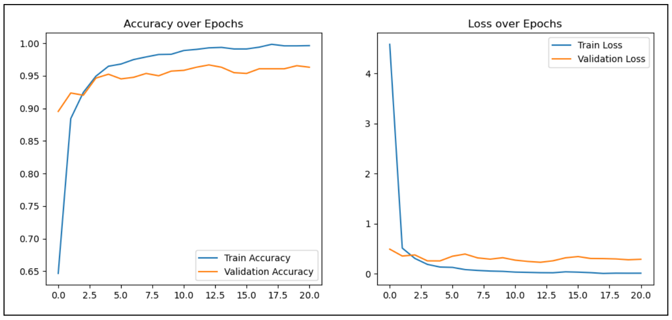
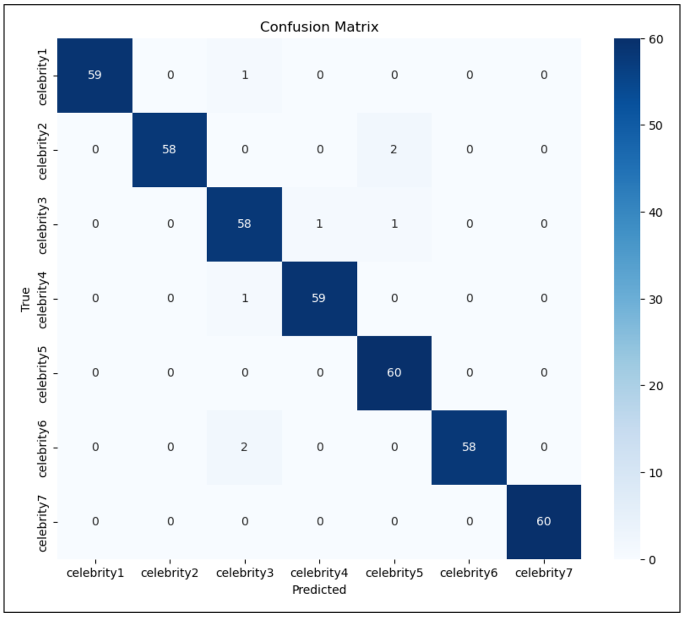


Fig. 7: Accuracy and Loss evolution for training and validation over epochs

TABLE III: Classification report for model performance for 7 celebrities

Celebrity Name	accuracy (%)	precision (%)	recall (%)	f1-score (%)
celebrity 1	98.33	100.00	98.33	99,15
celebrity 2	96.67	100.00	96.67	98.30
celebrity 3	96.67	93.54	96.66	95.07
celebrity 4	98.33	98.33	98.33	98.33
celebrity 5	100.00	95.23	100.00	97.55
celebrity 6	96.66	100.00	96.66	98.30
celebrity 7	100.00	100.00	100.00	100.00

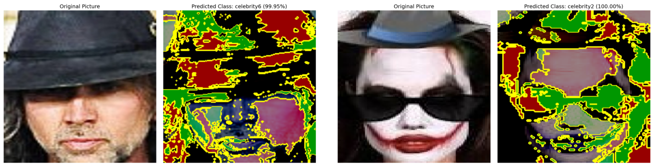


True \ Predicted	celebrity1	celebrity2	celebrity3	celebrity4	celebrity5	celebrity6	celebrity7
celebrity1	59	0	1	0	0	0	0
celebrity2	0	58	0	0	2	0	0
celebrity3	0	0	58	1	1	0	0
celebrity4	0	0	1	59	0	0	0
celebrity5	0	0	0	0	60	0	0
celebrity6	0	0	2	0	0	58	0
celebrity7	0	0	0	0	0	0	60

Fig. 8: Confusion matrix

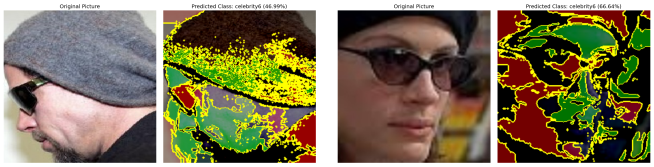
imposing colors indicating the most important areas of the image and the areas neglected or ignored for the identification. The yellow color represents the outline of the areas that most contributed to the recognition, the green color represents the areas that less important information contributed to the recognition and the red color represents the ignored areas in the recognition. As seen in the table of figures the results of identification is always correct with a high confidence rate sometimes reached 100%. In its face recognition the model focused on the shape of the eyes and the part of the face around the eyes such as the mouth and the nose which is presented by the yellow lines. We also notice the concentration of green and red colors on makeup and accessories such as glasses and hats, which explains that, our model neglects or completely ignores disguises in its identification.

TABLE IV: LIME explanability to understand prediction and dataset quality [27]


Original Picture
Predicted Class: celebrity6 (99.95%)
Original Picture
Predicted Class: celebrity2 (100.00%)

- Model limitation: Despite its high performance in recognizing most tested faces in in disguise and makeup, the model has difficulty recognizing some orientations of disguised faces. As shown in table of figures V, face recognition is wrong. LIME's color overlay shows that the identification is influenced by off-face areas as shown by the yellow outlines. This limitation can be corrected by adding more images with these orientations for training and validation, but as already mentioned before, in a face criminal recognition context and to be more realistic, it may be that we do not find enough images with such positions.

TABLE V: LIME explainability to understand model limit with turned face [27]


Original Picture
Predicted Class: celebrity6 (46.99%)
Original Picture
Predicted Class: celebrity6 (66.64%)

IV. Conclusion

To build solid forensic evidence that can be used by judges or lawyers in court, it must be trustworthy and easy to understand. The application of the LIME explainability algorithm allowed us to achieve this goal and consolidate the test results for our model. The performance metrics of our experiment reached values that strengthen the credibility of facial recognition results as digital forensic evidence. The limitations identified in our research will be the focus of future studies, exploring other AI and deep learning techniques, such as the GAN (Generative Adversarial Networks) algorithm, in combination with other CNN algorithms for face recognition.

References

[1] A. K. Jain, A. Ross, and K. Nandakumar, *Introduction to biometrics*. Springer Science & Business Media, 2011.

[2] J. Fattahi, B. E. Lakdher, M. Mejri, R. Ghayoula, E. Manai, and M. Ziadia, "Fingfor: a deep learning tool for biometric forensics," in *10th International Conference on Control, Decision and Information Technologies, CoDIT 2024, Vallette, Malta, July 1-4, 2024*, pp. 1667–1672, IEEE, 2024.

[3] J. Fattahi, F. Sghaier, M. Mejri, R. Ghayoula, E. Pricop, and B. E. Lakdher, "Handwritten signature recognition using parallel cnns and transfer learning for forensics," in *10th International Conference on Control, Decision and Information Technologies, CoDIT 2024, Vallette, Malta, July 1-4, 2024*, pp. 1697–1702, IEEE, 2024.

[4] J. Fattahi, O. Fkiri, M. Mejri, and R. Ghayoula, "Hands and palms recognition by transfer learning for forensics: A comparative study," in *New Trends in Intelligent Software Methodologies, Tools and Techniques - Proceedings of the 23rd International Conference on New Trends in Intelligent Software Methodologies, Tools and Techniques (SoMeT_24), Cancun, Mexico, September 24-26, 2024* (H. Fujita, H. M. P. Meana, and A. Hernandez-Matamoros, eds.), vol. 389 of *Frontiers in Artificial Intelligence and Applications*, pp. 213–225, IOS Press, 2024.

[5] Z. Zhao, Y. Zhang, and Y. Wang, "Face recognition under disguise: A survey," *Journal of Visual Communication and Image Representation*, vol. 58, pp. 1–12, 2019.

[6] J. Fattahi, B. E. Lakdher, M. Mejri, R. Ghayoula, F. Sghaier, and L. Boumlik, "The good and bad seeds of CNN parallelization in forensic facial recognition," in *10th International Conference on Control, Decision and Information Technologies, CoDIT 2024, Vallette, Malta, July 1-4, 2024*, pp. 1719–1724, IEEE, 2024.

[7] Y. Wang, Y. Zhang, and Z. Zhao, "Deep learning for face recognition: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 8, pp. 1947–1965, 2020.

[8] A. Kumar, R. Singh, and S. Gupta, "Challenges in forensic face recognition: A review," *Forensic Science International*, vol. 319, p. 110688, 2021.

[9] J. Fattahi, "Machine Learning and Deep Learning Techniques used in Cybersecurity and Digital Forensics: a Review," *arXiv e-prints*, p. arXiv:2501.03250, Dec. 2024.

[10] S. Kowshik and Y. Rama Devi, "A machine learning and deep learning integrated model to detect criminal activities," *International Research Journal of Engineering and Technology (IRJET)*, pp. 25–35, 2023.

[11] M. Sewak, M. R. Karim, and P. Pujari, *Practical Convolutional Neural Networks: Implement advanced deep learning models using Python*. Packt Publishing Ltd, 2018.

[12] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2015.

[13] Y. LeCun, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 2015.

[14] M. Abas, N. Ismail, I. Yassin, and M. N. Taib, "Vgg16 for plant image classification with transfer learning and data augmentation," *International Journal of Engineering and Technology(UAE)*, vol. 7, pp. 90–94, 10 2018.

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

[16] P. J. Phillips, W. T. Scruggs, A. J. O'Toole, P. J. Flynn, K. W. Bowyer, C. L. Schott, and M. Sharpe, "Face recognition vendor test 2006 and iris challenge evaluation 2006 large-scale results," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 5, pp. 831–846, 2010.

[17] S. Ali, T. Abuhmed, S. El-Sappagh, K. Muhammad, J. M. Alonso-Moral, R. Confalonieri, R. Guidotti, J. Del Ser, N. Díaz-Rodríguez, and F. Herrera, "Explainable artificial intelligence (xai): What we know and what is left to attain trustworthy artificial intelligence," *Information Fusion*, vol. 99, p. 101805, 2023.

[18] M. T. Ribeiro, S. Singh, and C. Guestrin, ""why should i trust you?": Explaining the predictions of any classifier," 2016.

[19] 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," *International Journal of Computer Vision*, pp. 1–12, 2019.

[20] C. C. Ukwuoma, D. Cai, E. Eziefuna, A. Oluwasanmi, S. F. Abdi, G. W. Muoka, D. Thomas, and K. Sarpong, "Enhancing histopathological medical image classification for early cancer diagnosis using deep learning and explainable AI - LIME & SHAP," *Biomed. Signal Process. Control.*, vol. 100, p. 107014, 2025.

[21] E. Manai, M. Mejri, and J. Fattahi, "Fingerprint fraud explainability using grad-cam for forensic procedures," in *New Trends in Intelligent Software Methodologies, Tools and Techniques - Proceedings of the 23rd International Conference on New Trends in Intelligent Software Methodologies, Tools and Techniques (SoMeT_24), Cancun, Mexico, September 24-26, 2024* (H. Fujita, H. M. P. Meana, and A. Hernandez-Matamoros, eds.), vol. 389 of *Frontiers in Artificial Intelligence and Applications*, pp. 457–470, IOS Press, 2024.

[22] Wisdomml, "Understanding vgg16: A powerful deep learning model for image recognition," 2023.

[23] J. Lee, S. Kim, and M. Park, "Performance metrics and evaluation of vgg16 model in image classification tasks," *Journal of Artificial Intelligence Research*, pp. 101–115, 2023.

[24] P. Biecek and T. Burzykowski, *Explanatory Model Analysis: Explore, Explain, and Examine Predictive Models*. Chapman and Hall/CRC, 2021.

[25] A. Rajpal, K. Sehra, R. Bagri, and P. Sikka, "Xai-fr: Explainable ai-based face recognition using deep neural networks," *Journal of Artificial Intelligence Research*, vol. 67, pp. 123–145, 2023.

[26] N. Nayyem, A. Rakin, and L. Wang, "Bridging interpretability and robustness using lime-guided model refinement," *arXiv preprint arXiv:2412.18952*, 2024.

[27] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[28] C.-M. Tsai and Z.-M. Yeh, "Contrast enhancement by automatic and parameter-free piecewise linear transformation for color images," *IEEE Transactions on Consumer Electronics*, vol. 54, no. 2, pp. 213–219, 2008.

[29] K. Pal and B. V. Patel, "Data classification with k-fold cross validation and holdout accuracy estimation methods with 5 different machine learning techniques," in *2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC)*, pp. 83–87, 2020.