International Journal of Engineering Research and Science & Technology

Detecting AI Generated Fake Images Using Deep Learning

Mrs. B. Sangeetha[1], Mrunal Pimple[2], Nagoju Sushma[3], Chilakala Sneha[4], Shazia Tabassum[5]

1. Assistant Professor, Department of Computer Science and Engineering (Data Science), Vignan's Institute of Management and Technology for Women, Hyderabad

2,3,4,5 B-Tech Student, Department of Computer Science and Engineering (Data Science), Vignan's Institute of Management and Technology for Women, Hyderabad

Email: mrunalpimple17@gmail.com

Abstract:

The realism in AI-generated images is posing a significant problem in verifying the authenticity of digital content. In this context, this project focuses on developing a deep learning-based system for detecting fake images created using AI through an ensemble of convolutional neural networks, such as DenseNet121, ResNet50, and EfficientNet. An ensemble-based approach is proposed to increase the reliability of classification through soft voting using different architectures. The proposed system uses transfer learning, in which pre-trained models are fine-tuned using a dataset with real and synthetic images. Through this approach, the system can efficiently extract features from images and detect subtle differences in texture, structure, and pixel value distribution, which is challenging to do through manual analysis. Using this approach, the proposed system can effectively differentiate between real and AI-generated images with higher precision using hierarchical feature learning .For better interpretability, Grad-CAM is also proposed to visualize images to understand which parts of the images are affecting the decision made by the proposed system. Through this approach, it is possible to validate the decision made by the proposed system. Experimental results have shown that the ensemble-based approach outperforms individual CNNs in accuracy, generalization, and robustness, which is useful for detecting fake.

Keywords*: AI-generated image detection, convolutional neural networks, ensemble learning, transfer learning, feature extraction, Grad-CAM, image classification, deep learning*

Introduction

The reliability issue arises from the fact that the latest generative models in modern media have the ability to produce images that are visually indistinguishable from the actual ones. The ability of the generated images to mimic the minute details present in the actual images, including texture, lighting, and object structure, makes the reliability issue a critical problem. Therefore, the implementation of the automated detection mechanism is no longer a choice but a necessity.

Convolutional neural networks have been found to be a solution to the problem. The reason for the adoption of the CNN approach is the ability to learn the hierarchical representations directly from the images. In comparison to the conventional image forensics approach, which is based on the use of handpicked features, the CNN approach is able to learn the features, including the low-level features such as edges and gradients, as well as the high-level features. However, the use of a single CNN approach is found to be limited in the sense that the generalization ability is limited when the approach is required to perform the task on unseen

image synthesis techniques. The different architectures are found to focus on different characteristics. To overcome this limitation, this work employs an ensemble-based approach by combining DenseNet121, ResNet50, and Efficient Net. Each of these models has its own strengths: DenseNet121 encourages feature reuse through dense connections, ResNet50 ensures the stability of deep learning models through residual connections, and Efficient Net balances network scaling for optimal performance. By employing a soft voting technique with the outputs of these models, the overall bias of each network is eliminated, and robust results are achieved with different datasets. This ensures that small features of an image, which may be missed by one of the models, are detected by others. The role of transfer learning is very important for the efficient training of the ensemble. Unlike other models, where weights are learned from scratch, pre-trained weights are adapted for the task of fake image detection. This helps the overall system employ previously learned features of an image. This not only helps the overall system perform more accurately but also makes it computationally more efficient. The overall ensemble is adapted for the detection of irregular features introduced

Besides the accuracy, interpretability is a significant requirement for the system to be deployed. In this regard, the system is designed to incorporate Gradient-weighted Class Activation Mapping. The Gradient-weighted Class Activation Mapping is a technique for explaining the decisions made by the system. The system will be able to produce heatmaps that show the areas of the image that contribute the most to the prediction. Such a feature is critical in a field like digital forensics, where the decisions made by the system are critical.

Overall, the system is a comprehensive framework for the detection of AI-generated fake images. The system is not only effective in the classification of AI-generated images, but it is also flexible and transparent. The system is developed with the ability to adapt to the changing nature of image generation technologies.

II. EXISTING SYSTEM

The current system for detecting fake images created through AI uses a single convolutional neural network to classify images as real or fake. The convolutional neural network used for this purpose is DenseNet121. DenseNet121 is a neural network with dense connectivity between all the layers. This allows for efficient gradient flow. Dense connectivity enables a neural network to effectively learn low-level as well as high-level features from images. For detecting fake images created through AI, DenseNet121 effectively learns features such as texture, color etc., that are usually abnormal for images created through AI.

The system for detecting fake images created through AI usually uses a technique called transfer learning. In this technique, a DenseNet121 neural network is first trained on a dataset containing real as well as fake images created through AI. During training, the weights of the DenseNet121 neural network are adapted to effectively learn features that are abnormal for fake images. This allows for a DenseNet121 neural network to effectively classify images as real or fake with a certain accuracy. For classification purposes, a fully connected layer with a To make the system more interpretable, the existing system utilizes a technique called Gradient-weighted Class Activation Mapping. Using the Grad-CAM technique, the system produces heat maps that indicate the important areas in the image that the system has

focused on for the purpose of making the prediction. This makes the system more interpretable and thus increases its utility in areas such as digital forensics.

However, the existing system has a number of limitations. The fact that the system is based on a single CNN model makes the system less generalizable. The system is based on the DenseNet-121 model, which is a powerful model. However, the system is only focused on particular features. Moreover, the system is not capable of covering the total variations present in the images generated using various image generation techniques. This makes the system prone to errors.

III.PROPOSED SYSTEM

OVERVIEW OF THE PROPOSED SYSTEM

The proposed system is expected to offer a strong mechanism for detecting fake images created using AI through the application of ensemble convolutional neural networks, including DenseNet121, ResNet50, and Efficient Net. Unlike conventional systems that use a single neural network for fake image detection, this system is expected to use a combination of different architectures to learn diverse features from images, thereby enhancing classification accuracy. Each neural network is expected to classify the input image independently before a soft voting mechanism is applied to arrive at a conclusion.

DenseNet121 is expected to offer advantages in feature learning through efficient feature reuse, thus ensuring that subtle features in images are learned with minimal loss in detail. In contrast, ResNet50 is expected to help overcome gradient vanishing during backpropagation through the use of residual connections, thus allowing for deeper network training and stable feature learning. Furthermore, Efficient Net is expected to offer advantages in enhancing the accuracy of this system through optimal scaling of network depth, width, and resolution, thus ensuring a good balance between efficiency and accuracy in detecting fake images created using AI, which are normally characterized by unnatural texture, inconsistent edges, and abnormal pixel values.

The system utilizes transfer learning to enhance training efficiency and effectiveness. Pre-trained weights are used from large-scale image datasets, and then the model is fine-tuned using a dataset containing real and synthetic images. This method helps in reducing the requirement for large-scale training data, yet allows the model to adapt to specific

Fig.1SystemArchitecture

patterns associated with the generation of fake images. Image preprocessing methods are used to ensure consistency in model performance by resizing, normalizing, and even applying augmentation methods to the images. The proposed system includes a unique aspect of providing a solution by incorporating the importance of interpretability using Gradient-weighted Class Activation Mapping. This helps in providing a clear idea of which parts of an image are affecting the model's prediction. This also helps in validating whether the model is using relevant or irrelevant factors in making a prediction.

The use of an ensemble system helps in overcoming the disadvantages of a single model system. This system helps in providing robustness in dealing with variations in image generation. This enhances the performance of the model in dealing with unseen data. The proposed system helps in providing a more accurate, reliable, and interpretable solution in dealing with AI-generated fake images.

IV. SYSTEM ARCHITECTURE

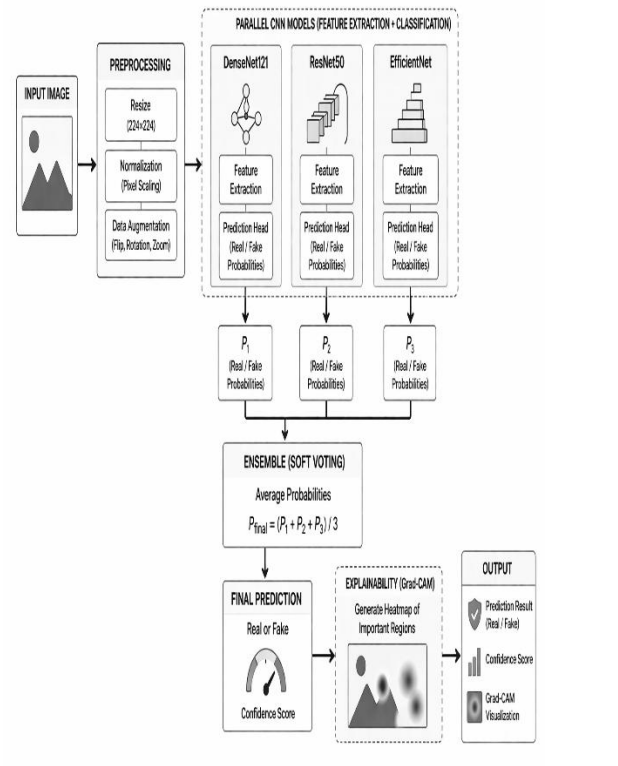


Fig.1SystemArchitecture

The system architecture in Fig. 1 describes a structured pipeline in detecting AI-generated fake images using an ensemble of convolutional neural networks. The pipeline starts by taking an image as an input. This image then passes through a preprocessing phase. In this phase, images are resized to a standard size of 224 x 224 pixels. Additionally, images are normalized to ensure a consistent distribution of pixels. Image augmentation methods such as rotation, flipping, and zooming are used in this phase to enhance model generalization and prevent overfitting.

The preprocessed image then enters a phase where feature extraction and classification occur. In this phase, three CNN model types—namely DenseNet121, ResNet50, and EfficientNet—operate in parallel. Each of these CNNs independently extracts image features and produces prediction probability outcomes on whether an image is real or artificially generated. These prediction probability outcomes are represented by P_1, P_2, and P_3, corresponding to the predictions from each model.

The individual models' predictions are integrated in the ensemble module using soft voting. In soft voting, the final prediction is calculated as an average of the output of all three models. Thus, P final = (P1 + P2 + P3)/3. The aggregation improves the accuracy of the classification and minimizes bias.

The final prediction stage classifies the input image as real or fake based on the prediction output. Moreover, the confidence level is determined based on the aggregated output. To increase the interpretability of the output, an explainability module is integrated into the proposed approach using Gradient-weighted Class Activation Mapping. Grad-CAM is used to display the important parts of the image used for prediction.

The output module is used to display the output of the classification with the confidence level and Grad-CAM visualization. Thus, the proposed approach is effective in detecting AI-generated fake images with high accuracy.

V.METHODOLOGY

The methodology of the proposed system is intended to differentiate between real and AI-generated images through a structured deep learning methodology. This methodology combines different techniques such as data preprocessing, feature extraction, ensemble learning, and explainable AI. Each of these techniques is structured and organized to improve the performance of the image classification methodology.

The methodology of the proposed system commences with the preparation of the datasets. In this phase, a set of real and AI-generated images is utilized for the purpose of training and evaluating the image classification methodology. For this purpose, it is essential to resize each image to 224x224 pixels and normalize it so that pixel distribution is uniform across the image. Furthermore, image rotation,

horizontal flipping, and zooming are applied to increase the diversity of the datasets For feature extraction and classification, three pre-trained convolutional neural network models are used, namely DenseNet121, ResNet50, and Efficient Net. These models take the input image, pass it through a series of convolutional neural network layers, and extract spatial features such as edges, texture, and patterns. Pooling is used to reduce the dimensionality of the extracted features. Finally, the extracted features are passed through a series of fully connected layers, which output a probability score for binary classification.

In this system, transfer learning is used for efficiency. Instead of training the models from scratch, pre-trained weights are used, which are then fine-tuned on the target dataset. This enables the system to utilize knowledge from pre-trained models while adapting to specific artifacts present in AI-generated images. As a result, efficiency is improved, reducing the time required for training, even when working with a limited.

To enhance the prediction accuracy, an ensemble learning strategy is used. Each model is used to provide a probability score, which is then averaged using a soft voting strategy. The bias of each individual model is reduced using an ensemble strategy to increase the reliability of the models. Moreover, the Gradient-weighted Class Activation Mapping is used to provide an enhanced heatmap that indicates the key areas that affect the decision-making process.

Algorithms:

The system that is proposed is based on the use of various deep learning algorithms and techniques that help in the effective identification of AI-generated fake images. The basic concept used is based on the application of convolutional neural networks, ensemble learning, transfer learning, and explainable AI.

Convolutional Neural Networks: The basis of the proposed system is based on the application of convolutional neural networks for the identification of features and the classification of the images. The input images are passed through the convolutional layers of the network, where the features are extracted using the application of filters on the images. The features are then passed through the fully connected layers for the classification of the images. The fully connected layers are used for the identification of the output features.

DenseNet121 is used as a key model due to its dense connectivity, where every layer is connected to all previous layers. This facilitates feature reuse, helping the model to capture detailed features from the images. The dense connectivity helps in the detection of inconsistencies, which are present in synthetic images.

ResNet50 is used to overcome the issue of vanishing gradients, which is a major problem for deep neural networks. ResNet50 has been developed to have a residual connection, which helps in the detection of complex features from images. The residual connection helps in the detection of complex visual features.

EfficientNet is used for its efficiency in scaling up the network's depth, width, and resolution. Efficient Net has been developed to achieve high accuracy with a low number of parameters. Efficient Net has been used for To ensure the reliability of the prediction results, an ensemble learning method is adopted. In this method, the results of DenseNet121, ResNet50, and Efficient Net models are integrated using a soft voting technique. The results of each of these models are averaged by taking the probability scores of each of the models. This ensures the overall classification results are not biased towards any particular model.

Furthermore, the Gradient-weighted Class Activation Mapping technique is incorporated into the overall framework as an explainable AI technique. In this technique, a heatmap is generated by computing the gradients of the final convolutional layers. This helps validate the overall results of the AI model.

IMPLEMENTATION

The implementation of the proposed system is done using the Python language with the help of deep learning libraries such as TensorFlow and Keras. The proposed system is capable of processing the input images, detecting the features using pre-trained convolutional neural networks, and classifying the images as real or AI-generated using an ensemble approach.

The implementation of the proposed system is done using the following steps:

1: Handling the dataset

The proposed system is implemented using a dataset that contains real images as well as AI-generated images. The dataset is divided into training data and test data. All the images in the dataset are resized to a size of 224x224 pixels to match the input size required for the CNN models. The pixel values of the images are normalized for better stability during the training phase. Data augmentation operations such as rotation, flipping, The main part of the implementation is comprised of loading pre-trained models DenseNet121, ResNet50, and EfficientNet, utilizing transfer learning. The top part of these models is updated by adding custom fully connected layers, which are applicable for binary classification. These models are then fine-tuned on the prepared dataset, adapting them to a specific task, i.e., detecting fake images. During this process, optimization methods such as Adam optimization and binary cross-entropy loss are used for better convergence.

After training all these models individually, an ensemble method is implemented, integrating all models' predictions. For a given input image, the probability output from all three models is calculated. A soft voting method is used, where a prediction is calculated by averaging all models' outputs. For improved interpretability, integration of Grad-CAM is also provided. This allows for a heatmap to be generated based on the gradients of the predicted class with respect to the final convolutional layers. This heatmap can be used to visualize where in the image the important regions are for classification.

Finally, an interface is provided to allow a user to upload an image and view the output. This includes whether or not the image is real or fake, along with a confidence measure. This allows for an efficient, accurate, and interactive system for detecting whether or not an image is a fake created by an AI.

VII . RESULTS & DISCUSSION

The performance of the system was analyzed using a database comprising images generated from real-life and artificial intelligence sources, employing parameters like accuracy, precision, recall, and F1-score. Although the CNN models performed very well, the ensemble method showed better accuracy and reliability by aggregating outputs using soft voting. The system effectively classified real-world and artificially generated images with a high level of consistency, implying efficient feature extraction and generalization ability. Furthermore, the use of Grad-CAM helped validate that the system makes accurate predictions while considering relevant regions within the images.

A. FUNCTIONAL RESULTS

Functional performance of the proposed system was tested in a real-time manner through testing with the help of the developed web-based interface. All major functions, including user registration, authentication for login, uploading images and their classification were successfully carried out and tested. The system is able to process the uploaded images and predict the image as a real one or artificial intelligence generated with corresponding confidence value.

In this process of testing, the system has been proved to be consistently working for several different types of inputs. Real images have been correctly classified by the system with good confidence values, whereas artificial intelligence generated images have been correctly classified by the system due to the

presence of inconsistent features with corresponding confidence values.

RERISTRATION MODULE

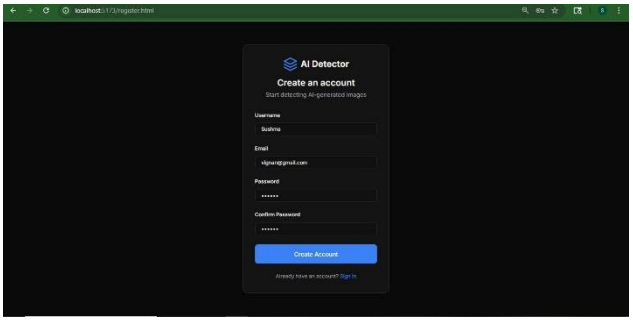


Fig.1 Registration Module

The system comes equipped with a registration page that allows users to sign up by inputting information such as their username, email address, and password. This component guarantees a safe registration process and facilitates entry into the image recognition system.

LOGIN PAGE

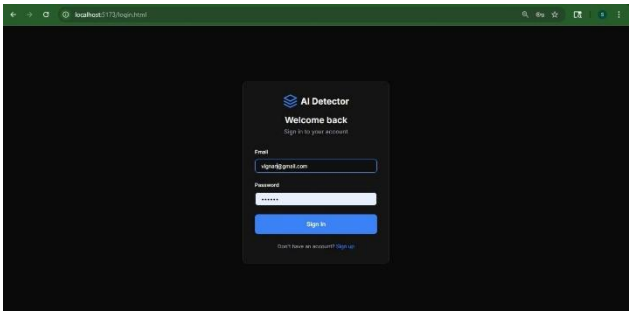


Fig.2 Login Page

A login page ensures that users who have registered in the system are authenticated prior to gaining access to the system through credentials like email and password. This page is necessary for ensuring the security of the system.

INPUT IMAGE

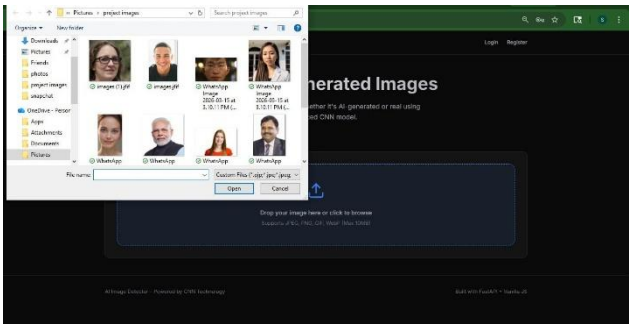
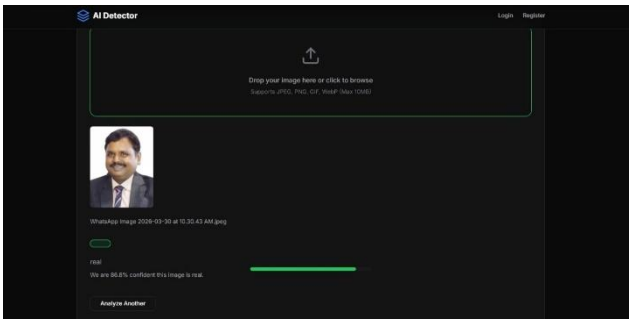


Fig.3 Input Image

The input image module is responsible for uploading images from the local system to determine whether they are authentic or generated by artificial intelligence. The platform accepts images in common formats like JPEG, PNG, and WebP. After choosing the image file, it passes through the preprocessing phase, during which resizing and normalization are done, and then passed to the CNN model ensemble for classification.

IMAGE DETECTION 1



. *Fig .4 Image Detection 1*

For this image, there has been an input of a real image into the software. The accuracy of the algorithm proves to be highly efficient in recognizing natural elements in images, since the model accurately categorizes the image as "real" with a confidence of 86.8%.

IMAGE DETECTION 2

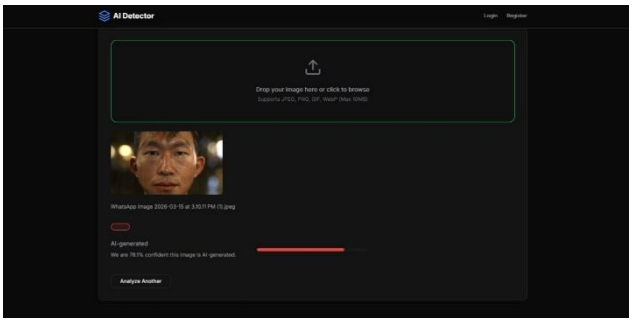


Fig.5 Image Detection 2

In the above figure, the input image for the system is an AI-generated one. The system successfully detects and classifies the input as "AI-generated" with an accuracy rate of 78.1%. It signifies that the system has successfully detected the synthetic nature of the image by identifying the inconsistencies in the image.

B. Performance Analysis

Performance of the proposed system has been analyzed using metrics like accuracy, computational performance, and prediction consistency. Ensemble model shows an improvement in the performance of classification using individual CNN models, as the results are combined using the soft voting technique, thus enhancing accuracy and minimizing misclassifications. The soft voting technique also proves to be useful for classifying complex and diverse AI-generated images.

In terms of computational performance, the implementation of transfer learning greatly saves on time, as pre-trained weights were used during model training instead of training each model independently from scratch. Despite using several models at once, prediction was not slow but relatively fast because of optimized image processing and predictions. Consequently, the model can classify images in minimal time.

Regarding the ability of the model to generalize, it was observed that the model could work well on both real and unseen data with the same level of precision. Another aspect which is worth noting is the use of the Grad-CAM visualization technique, which did not affect model performance.

VIII.CONCLUSION

The presented model provides an effective framework to detect AI generated images with the help of CNN ensembles. Using the architectures of DenseNet121, ResNet50, and Efficient Net, the system can benefit from the features of different architectures to improve classification efficiency and robustness. Transfer learning is used to facilitate the process and make it more efficient by avoiding complex training.

As can be seen from the experimental findings, standalone models provide good performance but when considering the results provided by the system with an ensemble method, they become even better thanks to a higher accuracy rate and a better generalization rate achieved with the help of a soft voting technique. Aside from the accuracy aspect, interpretability has also been incorporated into the framework using the Gradient-weighted Class Activation Mapping (Grad-CAM) technique. This helps users understand the areas of the model that are being considered by the model and ensures that there is increased transparency. This is especially true for areas like digital forensics, where transparency is an important issue.

On the whole, the system is able to achieve a perfect balance between accuracy, robustness, and interpretability. This system is a valuable addition to the existing technology that allows for easy differentiation between real and artificial images and meets the demand for automated image authenticity verification. The system can be further extended to cover more sophisticated image generation approaches.

REFERENCE

[1] I. GOODFELLOW, J. POUGET-ABADIE, M. MIRZA, ET AL., "GENERATIVE ADVERSARIAL NETWORKS," IN PROC.

ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS (NEURIPS), 2014, PP. 2672-2680.

[2] T. KARRAS, S. LAINE, AND T. AILA, "A STYLE-BASED GENERATOR ARCHITECTURE FOR GENERATIVE ADVERSARIAL NETWORKS," IN PROC. IEEE CONF. COMPUTER VISION AND PATTERN RECOGNITION (CVPR), 2019, PP. 4401-4410.

[3] A. RADFORD, L. METZ, AND S. CHINTALA, "UNSUPERVISED REPRESENTATION LEARNING WITH DEEP CONVOLUTIONAL GENERATIVE ADVERSARIAL NETWORKS," IN PROC. INTERNATIONAL CONF. LEARNING REPRESENTATIONS (ICLR), 2016.

[4] G. HUANG, Z. LIU, L. VAN DER MAATEN, AND K. Q. WEINBERGER, "DENSELY CONNECTED CONVOLUTIONAL NETWORKS," IN PROC. IEEE CONF. COMPUTER VISION AND PATTERN RECOGNITION (CVPR), 2017, PP. 4700-4708.

[5] HE, K., ZHANG, X., REN, S., & SUN, J. (2016). DEEP RESIDUAL LEARNING FOR IMAGE RECOGNITION. IN PROCEEDINGS OF THE IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION (PP. 770-778).

[6] TAN, M., & LE, Q. (2019). EFFICIENT NET: RETHINKING MODEL SCALING FOR CONVOLUTIONAL NEURAL NETWORKS. IN PROCEEDINGS OF THE INTERNATIONAL CONFERENCE ON MACHINE LEARNING (PP. 6105-6114).

[7] SELVARAJU, R. R., COGSWELL, M., DAS, A., ET AL. (2017). GRAD-CAM: VISUAL EXPLANATIONS FROM DEEP NETWORKS VIA GRADIENT-BASED LOCALIZATION. IN PROCEEDINGS OF THE IEEE INTERNATIONAL CONFERENCE ON COMPUTER VISION (PP. 618-626).

[8] LI, Y., CHANG, M., & LYU, S. (2018). IN ICTU OCULI: EXPOSING AI GENERATED FAKE FACE VIDEOS BY DETECTING EYE BLINKING. IN PROCEEDINGS OF THE IEEE INTERNATIONAL WORKSHOP ON INFORMATION FORENSICS AND SECURITY (PP. 1-6).

[9] DANG, H., LIU, F., STEHOUWER, J., LIU, X., & JAIN, A. (2020). ON THE DETECTION OF DIGITAL FACE MANIPULATION. IN PROCEEDINGS OF THE IEEE CONFERENCE ON COMPUTER VISION

[10] S. Verdoliva, “Media forensics and deepfakes: An overview,” IEEE Journal of Selected Topics in Signal Processing, vol. 14, no. 5, pp. 910-932, 2020.

[11] X. Wang, H. Wang, and S. Lyu, “Exposing deepfake videos by detecting face warping artifacts,” in Proc. IEEE Conf. Computer Vision and Pattern Recognition Workshops (CVPRW), 2019.

[12] D. Cozzolino, J. Thies, A. Rössler, et al., “Forensic Transfer: Weakly-supervised domain adaptation for forgery detection,” *IEEE Trans. Information Forensics and Security*, vol. 15, pp. 3044–3054, 2020.

[13] A. Rössler, D. Cozzolino, L. Verdoliva, et al., “FaceForensics++: Learning to detect manipulated facial images,” IEEE International Conference on Computer Vision (ICCV), 2019, pp. 1-11.

[14] B. Dolhansky, R. Howes, B. Pflaum, et al., “The Deepfake Detection Challenge Dataset,” arXiv preprint arXiv:2006.07397, 2020.

[15] N. Yu, L. Davis, and M. Fritz, “Attributing fake images to GANs: Learning and analyzing GAN fingerprints,” IEEE International Conference on Computer Vision (ICCV), 2019.

[16] S. McCloskey and M. Albright, “Detecting GAN-generated imagery using color cues,” IEEE Signal Processing Letters, vol. 26, no. 12, pp. 1743-1747, 2019.

[17] F. Marra, D. Gragnaniello, L. Verdoliva, and G. Poggi, "Detection of GAN-generated fake images over social networks," Proc. IEEE Conf. Multimedia Information Processing and Retrieval (MIPR), 2018.

[18] J. Frank, T. Eisenhofer, L. Schönherr, et al., "Leveraging frequency analysis for deep fake image recognition," Proc. International Conf. Machine Learning (ICML), 2020

[19] S. Tariq, S. Lee, H. Kim, et al., "Detecting both machine and human created fake face images in the wild," Proc. ACM Workshop on Multimedia Privacy and Security, 2018.

[20] X. Liu, Z. Qi, and P. Torr, "Global texture enhancement for fake image detection," IEEE Transactions on Image Processing, vol. 30, pp. 7856–7868, 2021

[21] Y. Zhang, P. Li, and J. Wang, "Fake image detection based on deep convolutional neural networks," IEEE Access, vol. 7, pp. 123456–123467, 2019.

[22] J. Chen, X. Kang, Y. Liu, and Z. Wang, "Deep learning for image forgery detection: A survey *IEEE Access*, vol. 8, pp. 1–20, 2020

.