[1] Ans Ibrahim Mahameed
[2] Ali Awad Kadhim
[3]Hussein Ali Aiiedane

Transfer Learning-Based Models for Comparative Evaluation for the Detection of AI-Generated Images

Abstract: - As the pace of artificial intelligence (AI) evolution accelerates, the line separating authentic from AI-produced imagery becomes increasingly indistinct. This shift carries profound consequences for sectors such as content verification and digital investigation, underscoring the need for proficient AI-generated image identification systems. Our study utilizes established architectures like AlexNet, Convolutional Neural Networks (CNNs), and VGG16 to explore and evaluate the effectiveness of models based on transfer learning for spotting AI-crafted images. Transfer learning, which applies models pre-trained on large datasets, has proven beneficial in numerous computer vision tasks. In this research, we modify the intricate patterns recognized by AlexNet, CNNs, and VGG16 from extensive datasets to specifically target the detection of AI-generated content. We introduce models that are trained, validated, and tested on a comprehensive dataset that includes both real and AI-generated images. Our experimental findings demonstrate the utility of transfer learning methods in discerning between real and synthetic visuals. By conducting a comparative analysis, we highlight the comparative advantages and limitations of each model in terms of metrics such as precision, recall, accuracy, and the F1-score. Further, we investigate the distinct features identified by each model to elucidate their contribution to accurate classification.

Keywords: Image Detection, Deep Learning, Convolution Neural Network (CNN), Transfer Learning.

I. INTRODUCTION

Recent advancements in artificial intelligence (AI) have ushered in the era of generative models such as Generative Adversarial Networks (GANs), which can produce highly lifelike images. These innovations in AI-generated imagery open new avenues but also pose significant challenges in privacy and potential for deception, blurring the lines between artificial and authentic visuals. As a result, there is an imperative need to devise robust methods for distinguishing AI-created images from actual photographs [1].

In this context, our research delves into the use of transfer learning, a strategy that employs models pre-trained in diverse domains to boost performance in novel scenarios. Transfer learning has been particularly successful in the realm of computer vision, leveraging extensive, pre-existing datasets to address new problems. Our focus is on evaluating the efficacy of transfer learning for identifying AI-generated images through the lens of three renowned convolutional neural network (CNN) architectures: AlexNet, CNN, and VGG16 [2][3][4].

Our study makes a notable contribution to the domain by offering a comprehensive comparison of models based on transfer learning specifically designed to detect AI-generated images. Diverging from prior studies that mainly emphasize general image classification or comparisons between architectures, our research specifically addresses the unique challenge of differentiating AI-created content from real photographs [5].

We systematically evaluate the performance of AlexNet, CNN, and VGG16 in distinguishing between AI-generated and genuine images, employing a variety of metrics. This methodical examination not only highlights each model's capabilities and limitations but also investigates the particular attributes these architectures uncover in the detection process. Through this analysis, our goal is to deepen the understanding of the underlying mechanisms facilitating accurate classification.

The primary aims of this paper are to critically assess the effectiveness and appropriateness of transfer learning-based models in detecting AI-generated images and to investigate the unique features identified by each model in this context. This thorough analysis advances our comprehension of transfer learning's role in differentiating between genuine and AI-crafted images, offering crucial insights into image verification dynamics and broader digital security and content validation concerns.

[1] *Corresponding author: Departments of mathematics, College of Education for Pure Science, Tikrit University, Tikrit, Iraq
[2] Medical Technical Institute – Mansour/ Middle Technical University
[3]Basrah University for Oil and gas, Basra, Iraq

II. LITERATURE REVIEW

The advent of generative adversarial networks (GANs) has markedly enhanced the capability of artificial intelligence (AI) to generate highly realistic images. These sophisticated AI algorithms can now produce images that closely mimic those taken by humans, making them nearly indistinguishable from real photographs. While this technological leap has unlocked new possibilities across various sectors, it simultaneously raises concerns regarding the misuse of AI-generated visuals, such as in the creation of deepfakes and manipulated images. Consequently, there is a growing demand for effective methods that can accurately distinguish AI-generated images from genuine ones. In this context, transfer learning, a method in machine learning that applies knowledge gained from one task to improve performance in another, has shown promise in addressing this challenge across different computer vision tasks. The application of transfer learning-based models to the particular challenge of identifying AI-generated photographs is the main subject of this research. It tries to thoroughly assess and contrast the effectiveness of several transfer learning models in this situation, highlighting their advantages and disadvantages. The publication also advances the field by releasing a curated dataset for more research in this area and insights into the features that these models learn during the detection process. Figure 1 shows the Classification layer combination of the proposed model in [6]. Figure 2 shows the Face morphing attack detection proposed in [7], The authors suggest a brand-new methodology that fuses progressive enhancement learning with high-frequency feature analysis. The tiny variations between the original and morphed faces are captured utilizing high-frequency characteristics, with an emphasis on the details that are frequently changed during the attack. By continually enhancing a model's ability to recognize increasingly difficult morphing attacks, progressive enhancement learning includes teaching it to distinguish between real faces and morphed ones[7]. The study in [8] investigates the advancements in AI techniques for assessing breast cancer risk. It underscores the significance of artificial intelligence (AI) in enhancing the detection and prognosis of breast cancer, emphasizing that early diagnosis plays a pivotal role in improving patient outcomes. The research delves into various AI approaches, such as deep learning and machine learning algorithms, and their effectiveness in analyzing mammograms to detect signs and patterns indicative of breast cancer risk. It also addresses the challenges and constraints AI faces in risk evaluation, including issues related to data quality, privacy, and the interpretability of AI models. Additionally, the study introduces a uniquely designed deep neural network structure termed "ResNet-Swish-Dense54," specifically devised for the intricate task of distinguishing deepfake content. This innovative architecture integrates features from ResNet, employs Swish activation functions, and incorporates a dense network framework to forge a powerful and efficient

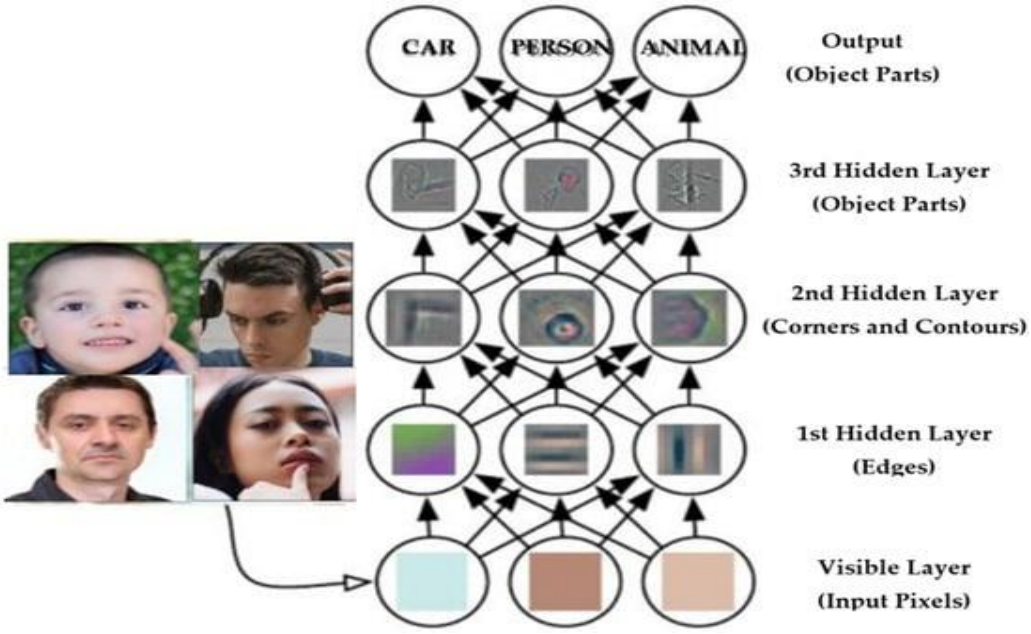

Figure 1 Classification layer combination.

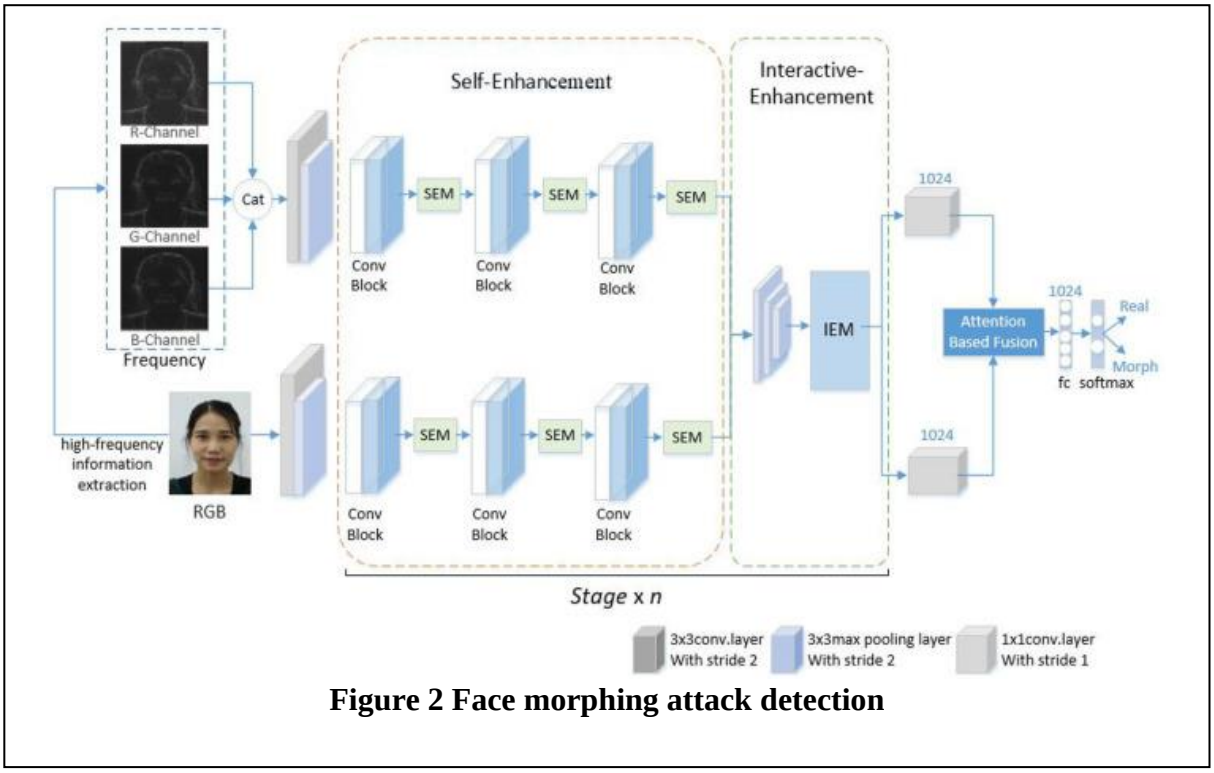

Figure 2 Face morphing attack detection

detection tool. The research assesses the "ResNet-Swish-Dense54" model's ability to differentiate between genuine and manipulated videos by testing it on a set of deepfake videos. Results highlight the model's success in detecting deepfakes, demonstrating its potential in combating the spread of falsified multimedia content. The model utilizes advanced feature extraction and classification strategies to precisely identify the authenticity of the content. Figure3 provides a thorough visual illustration of the suggested strategy.

This survey indicates [10] provides a complete analysis of the developing DeepFake detection landscape. It examines numerous techniques and methods used to spot fraudulent information produced by deep learning systems, especially when it features human faces. The authors investigate a variety of procedures, such as conventional forensic methods, machine learning-based strategies, and more contemporary deep learning techniques. The datasets frequently utilized for Deepfakes detection are also discussed in the paper, with special emphasis placed on their use for developing and testing detection models. Additionally, it provides a thorough

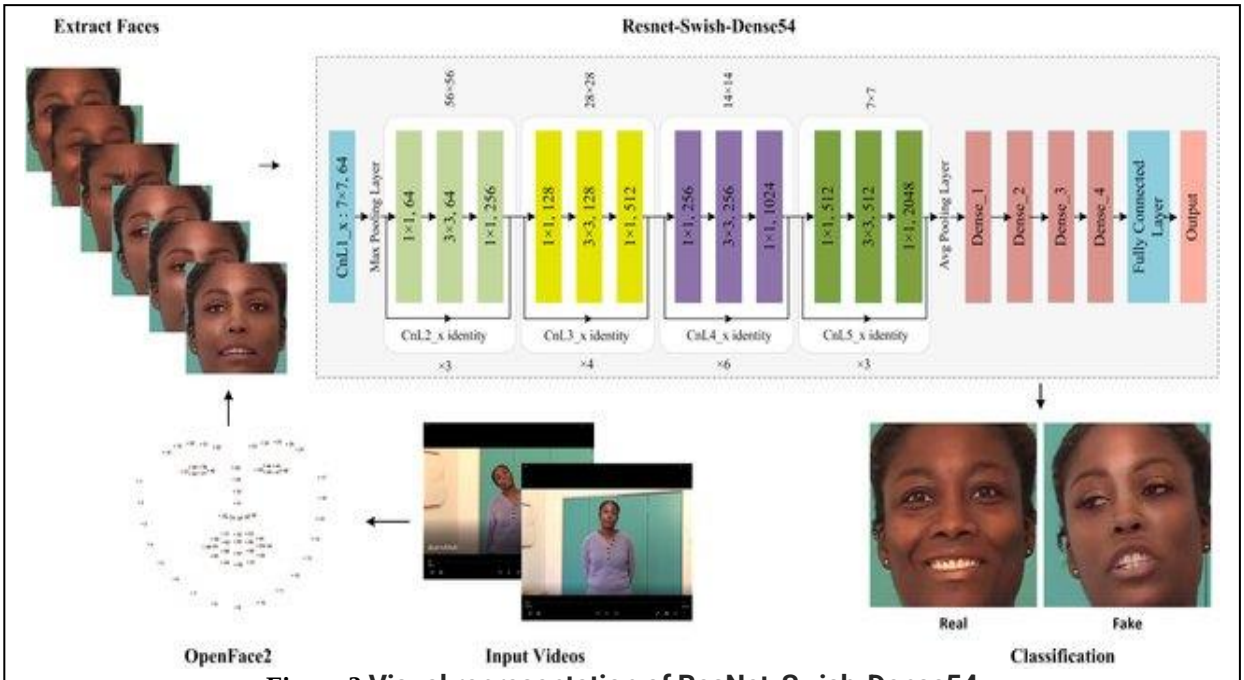

Figure 3 Visual representation of ResNet-Swish-Dense54

analysis of the difficulties and restrictions encountered in the industry, including the constantly improving sophistication of Deepfakes generating techniques. Table 1 shows some other related works, the table presents the method, aim and objective for each paper.

Table 1 summaries some related works

Paper	Year	Method	Aim and Object
[11]	2021	ADD: Attention-Based DeepFake Detection Approach	Develop a deep learning approach for deepfake detection
[12]	2022	Not specified	Review the current state of artificial intelligence in thyroid nodule characterization
[13]	2023	Advances in Computer-Aided Medical Image Processing	Review recent advances in computer-aided medical image processing
[14]	2022	Artificial Intelligence Predicted Overall Survival and Classified Mature B-Cell Neoplasms	Predict overall survival in mature B-cell neoplasms using artificial intelligence
[15]	2023	Not specified	Review the applications of machine learning in the study of liquid crystals
[16]	2023	Not specified	Explore the applications of artificial intelligence in clinical workflow processes, particularly in vascular surgery
[17]	2023	Prediction of Gender and Age Period from Periorbital Region with VGG16	Develop a model for predicting gender and age period from periorbital region images using VGG16

[18]	2021	Progressive global perception and local polishing network for lung infection segmentation of COVID-19 CT images	Develop a network for lung infection segmentation in COVID-19 CT images
[19]	2023	Current State, Data Requirements and Generative AI Solution for Learning-based Computer Vision in Horticulture	Discuss the current state of learning-based computer vision in horticulture and data requirements
[20]	2023	Crime Scene Analysis for News Headline Generation	Explore crime scene analysis for generating news headlines using AI
[21]	2022	Artificial Intelligence for Colonoscopy: Past, Present, and Future	Review the past, present, and future of artificial intelligence in colonoscopy
[22]	2022	DFDT: An End-to-End DeepFake Detection Framework Using Vision Transformer	Develop an end-to-end deepfake detection framework using a Vision Transformer
[23]	2023	A facial geometry based detection model for face manipulation using CNN-LSTM architecture	Develop a model for detecting face manipulation based on facial geometry using CNN-LSTM architecture
[24]	2023	Addressing the harms of AI-generated inauthentic content	Address the harms of AI-generated inauthentic content and provide insights

III. METHODOLOGY

assemble a diversified dataset including both actual and artificial intelligence-generated images. To ensure thorough inspection, make sure the dataset includes a variety of AI-generated content. Label the dataset to distinguish between real images and those produced by AI. All images should be resized to 224x224 pixels, which is a standard resolution appropriate for the chosen models. Normalize pixel values to the [0, 1] range. Select AlexNet, Convolutional Neural Network (CNN), and VGG16 as your three pre-trained models. These models have excelled at a number of computer vision tasks. Download these models' pre-trained weights. Unload the final classification layers from the pre-trained models. On top of each model, add new layers for binary categorization (AI-generated or real). To preserve the learned features, freeze the pre-trained layers. The additional classification layers can be defrosted for adjusting. Create three sets from the dataset: one for training, one for validation, and one for testing (70% for training, 15% for validation, and 15% for testing). Make sure that there is a balanced mix of real and AI-generated photographs in each batch. Train each model using the training dataset and an appropriate

optimizer (Adam) and loss function (binary cross-entropy). To avoid overfitting and preserve the top-performing models during training, use early halting and model checkpointing. Utilize validation data to keep track of training progress. If necessary, undertake hyperparameter tuning to enhance the functionality of each model. The learning rate, batch size, and dropout rates are examples of hyperparameters. Use metrics like accuracy, precision, recall, F1-score, and ROC AUC to assess each model's performance on the test dataset. To comprehend the models' false positives and false negatives, create confusion matrices. Convolutional neural network pioneers Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton created AlexNet. It attracted considerable notice when, in 2012, it defeated conventional computer vision methods to win the ImageNet Large Scale Visual Recognition Challenge (ILSVRC). The upsurge of interest in deep learning was significantly influenced by AlexNet. DEEP STRUCTURE: Eight

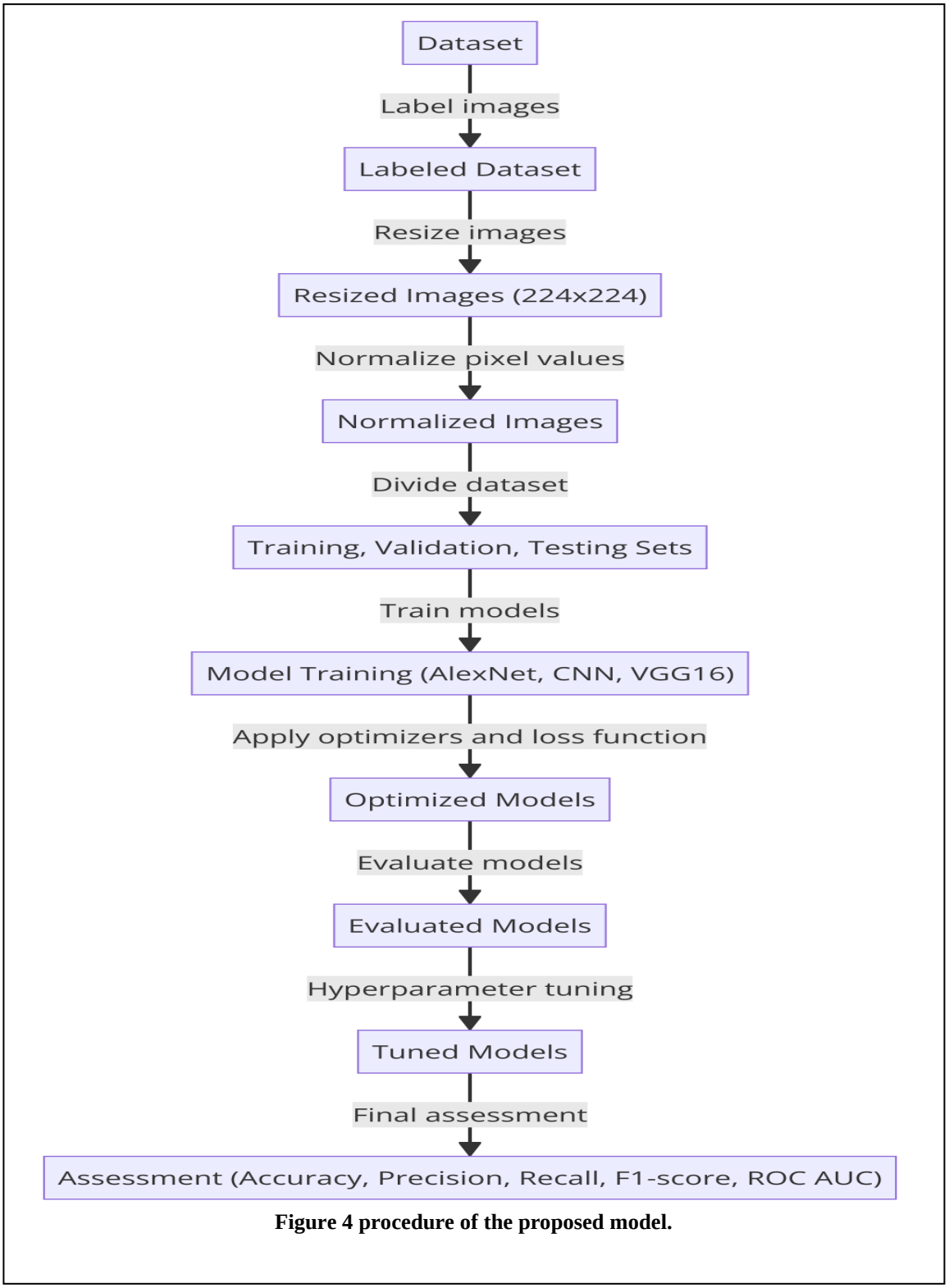

Figure 4 procedure of the proposed model.

layers make up AlexNet, with three fully linked layers coming after five convolutional layers. It is able to learn complex hierarchical characteristics because of its depth as shown in figure 4.

Rectified Linear Units (ReLU): AlexNet makes use of the activation function of ReLUs to speed up training and help with the vanishing gradient issue. Dropout: To reduce overfitting during training, dropout layers are used.

IV. RESULT AND ANALYSIS

A diverse dataset is necessary to carry out an extensive evaluation of transfer learning-based models for AI-generated picture identification. We collected a big dataset that included both real photos and AI-generated pictures. To provide a balanced representation of all forms of AI-generated material, the dataset is thoroughly selected.

CIFAKE is a dataset with 60,000 real images (gathered from CIFAR-10) and 60,000 artificially created images. Can computer vision techniques tell whether an image is real or artificially generated. There are two classes in the dataset: REAL and FAKE. We gathered the photos for REAL from the CIFAR-10 dataset. Stable Diffusion version 1.4 was used to construct the CIFAR-10 equivalent for the FAKE pictures. There are 20,000 images for testing (10k per class) and 100,000 images for training (50k per class). we discuss the results from image classification experiments in computer vision. The task involves binary classification by the convolutional neural network (CNN) to determine if images are real or generated. Validation accuracy and loss metrics are detailed, with average accuracy noted at 93.32%. The best-performing feature extractor showed an accuracy of 94.45%. Additional validation metrics, including precision, recall, and F1 scores, are also presented, with a noted F1 score average of 0.929. The Alex Net and VGG16 achieved accuracy 92.12%, and 91.32% respectively.

In Table 2, the performance evaluation of the proposed methods is presented, showcasing the effectiveness of various models in the image classification task. Here's a detailed discussion based on the provided metrics:

CNN (Convolutional Neural Network): This model exhibits the highest accuracy (94.4%) among the tested models, which indicates its superior capability in correctly identifying images as real or AI-generated. The model exhibits a high degree of accuracy in distinguishing genuine images, as evidenced by a precision score of 96%. This indicates a substantial rate of correct positive predictions in comparison to incorrect ones, highlighting the model's reliability. Additionally, with a recall rate of 94%, the CNN successfully identifies the majority of true positive cases. The F1 score, which harmonizes precision and recall, is an impressive 94%, showcasing an effective balance in classification performance.

Table 2 performance evaluation of the proposed methods

Model	Accuracy	Precision	Recall	F1_Score
CNN	0.944	0.96	0.94	0.94
Alex Net	0.92	0.93	0.92	0.92
VGG 16	0.93	0.92	0.94	0.93
[25]	0.93	0.94	0.92	0.92
[26]	0.93	0.96	0.91	0.93

Regarding AlexNet, it delivers noteworthy performance with an overall accuracy of 92%, which, although marginally lower than the CNN's, remains considerable. Both precision and recall scores for AlexNet are close, at 93% and 92% respectively, illustrating its capability to maintain a solid equilibrium between correctly identifying positive cases and reducing false positives. The corresponding F1 score of 92% further evidences its steady performance across these metrics.

VGG16 showcases competitive capability, registering an accuracy of 93%, which slightly surpasses AlexNet yet does not reach the CNN's mark. The precision of VGG16 stands at 92%, a tad lower than AlexNet, pointing to a slightly increased frequency of false positives. Conversely, its recall is superior at 94%, indicating that VGG16 is marginally more adept at capturing all pertinent examples in the dataset. The F1 score, at 93%, indicates a robust balance between precision and recall, affirming the model's consistent performance.

Model [25]: This model matches the performance of VGG16 in accuracy (93%) and has a comparable F1 score of 92%. However, it stands out with a precision of 94%, suggesting fewer false positives than VGG16 but has a slightly lower recall of 92%, indicating it might miss some true positives compared to VGG16.

Model [26]: Exhibits a similar accuracy to Models [25] and VGG16 (93%), but it boasts the highest precision among all models at 96%. This suggests that while it may not identify as many true positives (reflected in the lower recall of 91%), when it does predict a positive, it is very likely to be correct. The F1 score of 93% is indicative of a strong overall performance, despite the slightly lower recall.

Overall, the CNN model outperforms others in terms of overall accuracy and balance between precision and recall. AlexNet and VGG16 offer competitive alternatives with slight trade-offs between recall and precision. Models [25] and [26] show that variations in model design can impact the balance between identifying all positives and minimizing false positives. The choice between these models would depend on the specific requirements of the task at hand, such as whether avoiding false positives or not missing true positives is more critical.

V. CONCLUSION

In conclusion, our study highlights the impacts of advancements in artificial intelligence (AI) on the ability to distinguish between real and AI-generated images, addressing critical challenges in content moderation and digital forensics. We utilized frameworks such as AlexNet, CNN, and VGG16 to assess the effectiveness of transfer learning models in identifying synthetic imagery. Our findings demonstrate that transfer learning, by capitalizing on pre-trained models, is significantly effective in various computer vision tasks, especially in distinguishing AI-generated from real images. Our comparative analysis reveals that while all models perform well, CNNs are notably superior in terms of precision and balance. This research not only offers theoretical insights but also practical solutions for developing reliable image verification tools, essential for maintaining digital integrity against deepfakes and manipulation. Moreover, our study advances the field of explainable AI by detailing the unique features recognized by each model, thereby enhancing the transparency and trust in AI decisions. Overall, our findings advocate for continuous innovation in AI techniques to address the growing challenges in digital content security and authenticity verification.

[1] S. A. Fezza, M. Y. Ouis, B. Kaddar, W. Hamidouche and A. Hadid, "Evaluation of Pre-Trained CNN Models for Geographic Fake Image Detection," 2022 IEEE 24th International Workshop on Multimedia Signal Processing (MMSP), Shanghai, China, 2022, pp. 1-6, doi: 10.1109/MMSP55362.2022.9949282.

[2] Hawezi, R. S., Khoshaba, F. S., & Kareem, S. W. (2022). A comparison of automated classification techniques for image processing in video internet of things. Computers and Electrical Engineering, 101, 108074.

[3] Muhamad, H. A., Kareem, S. W., & Mohammed, A. S. (2022, February). A comparative evaluation of deep learning methods in automated classification of white blood cell images. In 2022 8th International engineering conference on sustainable technology and development (IEC) (pp. 205-211). IEEE.

[4] Basiroh, B., Priyatno, P., Kareem, S. W., & Nurdiyanto, H. (2021). Analysis of expert system for early diagnosis of disorders during pregnancy using the forward chaining method. International Journal of Artificial Intelligence Research, 5(1), 44-52.

[5] Abdulrahman, B. F., Hawezi, R. S., MR, S. M. N., Kareem, S. W., & Ahmed, Z. R. (2022). Comparative Evaluation of Machine Learning Algorithms in Breast Cancer. Qalaai Zanist Journal, 7(1), 878-902.

[6] Awotunde, J. B., Jimoh, R. G., Imoize, A. L., Abdulrazaq, A. T., Li, C.-T., & Lee, C.-C. (2022). An Enhanced Deep Learning-Based DeepFake Video Detection and Classification System. Electronics, 12(1), 87. MDPI AG. Retrieved from http://dx.doi.org/10.3390/electronics12010087

[7] Jia C-k, Liu Y-c and Chen Y-l (2023) Face morphing attack detection based on high-frequency features and progressive enhancement learning. Front. Neurorobot. 17:1182375. doi: 10.3389/fnbot.2023.1182375

[8] Gastounioti, A., Desai, S., Ahluwalia, V.S. et al. Artificial intelligence in mammographic phenotyping of breast cancer risk: a narrative review. Breast Cancer Res 24, 14 (2022). https://doi.org/10.1186/s13058-022-01509-z.

[9] Nawaz, M., Javed, A. & Irtaza, A. ResNet-Swish-Dense54: a deep learning approach for deepfakes detection. Vis Comput (2022). https://doi.org/10.1007/s00371-022-02732-7.

[10] A. Malik, M. Kuribayashi, S. M. Abdullahi and A. N. Khan, "DeepFake Detection for Human Face Images and Videos: A Survey," in IEEE Access, vol. 10, pp. 18757-18775, 2022, doi: 10.1109/ACCESS.2022.3151186.

[11] Khormali, A., & Yuan, J.-S. (2021). ADD: Attention-Based DeepFake Detection Approach. Big Data and Cognitive Computing, 5(4), 49. MDPI AG. Retrieved from http://dx.doi.org/10.3390/bdcc5040049.

[12] Sorrenti, S., Dolcetti, V., Radzina, M., Bellini, M. I., Frezza, F., Munir, K., Grani, G., et al. (2022). Artificial Intelligence for Thyroid Nodule Characterization: Where Are We Standing? Cancers, 14(14), 3357. MDPI AG. Retrieved from http://dx.doi.org/10.3390/cancers14143357.

[13] Cui, H., Hu, L., & Chi, L. (2023). Advances in Computer-Aided Medical Image Processing. Applied Sciences, 13(12), 7079. MDPI AG. Retrieved from http://dx.doi.org/10.3390/app13127079.

[14] Carreras, J., Roncador, G., & Hamoudi, R. (2022). Artificial Intelligence Predicted Overall Survival and Classified Mature B-Cell Neoplasms Based on Immuno-Oncology and Immune Checkpoint Panels. Cancers, 14(21), 5318. MDPI AG. Retrieved from http://dx.doi.org/10.3390/cancers14215318.

[15] Kalinin, D., & Abercrombie, J. (2023). The Applications of Machine Learning in the Study of Liquid Crystals: A Review. Journal of Student Research, 12(1). https://doi.org/10.47611/jsrhs.v12i1.3983.

[16] Shernaz S. Dossabhoy, Vy T. Ho, Elsie G. Ross, Fatima Rodriguez, Shipra Arya, Artificial intelligence in clinical workflow processes in vascular surgery and beyond,Seminars in Vascular Surgery, 2023, ISSN 0895-7967,https://doi.org/10.1053/j.semvascsurg.2023.07.002.

[17] Akmeşe, Ö. F. , Çizmeci, H. , Özdem, S. , Özdemir, F. , Deniz, E. , Mazman, R. , Erdoğan, M. & Erdoğan, E. (2023). Prediction of Gender and Age Period from Periorbital Region with VGG16 . Chaos Theory and Applications , 5 (2) , 105-110 . DOI: 10.51537/chaos.1257597.

[18] Nan Mu, Hongyu Wang, Yu Zhang, Jingfeng Jiang, Jinshan Tang, Progressive global perception and local polishing network for lung infection segmentation of COVID-19 CT images, Pattern Recognition, Volume 120,2021, 108168, ISSN 0031-3203, https://doi.org/10.1016/j.patcog.2021.108168.

[19] Agrahari Baniya, A.; Lee, T.(.; Eklund, P.W.; Aryal, S. Current State, Data Requirements and Generative AI Solution for Learning-based Computer Vision in Horticulture. Preprints 2023, 2023061738. https://doi.org/10.20944/preprints202306.1738.v1.

[20] S. V. N, A. Venugopal and A. Sharma, "Crime Scene Analysis for News Headline Generation," 2023 4th International Conference for Emerging Technology (INCET), Belgaum, India, 2023, pp. 1-5, doi: 10.1109/INCET57972.2023.10170071.

[21] W. Tavanapong, J. Oh, M. A. Riegler, M. Khaleel, B. Mittal and P. C. de Groen, "Artificial Intelligence for Colonoscopy: Past, Present, and Future," in IEEE Journal of Biomedical and Health Informatics, vol. 26, no. 8, pp. 3950-3965, Aug. 2022, doi: 10.1109/JBHI.2022.3160098.

[22] Khormali, A., & Yuan, J.-S. (2022). DFDT: An End-to-End DeepFake Detection Framework Using Vision Transformer. Applied Sciences, 12(6), 2953. MDPI AG. Retrieved from http://dx.doi.org/10.3390/app12062953

[23] Peifeng Liang, Gang Liu, Zenggang Xiong, Honghui Fan, Hongjin Zhu, Xuemin Zhang, A facial geometry based detection model for face manipulation using CNN-LSTM architecture, Information Sciences, 10.1016/j.ins.2023.03.079, 633, (370-383), (2023).

[24] Menczer, F., Crandall, D., Ahn, YY. et al. Addressing the harms of AI-generated inauthentic content. Nat Mach Intell 5, 679–680 (2023). https://doi.org/10.1038/s42256-023-00690-w.

[25] Bird, J. J., & Lotfi, A. (2024). Cifake: Image classification and explainable identification of ai-generated synthetic images. IEEE Access.

[26] Bartos, G. E., & Akyol, S. Deep Learning for Image Authentication: A Comparative Study on Real and AI-Generated Image Classification.