International Journal of Applied Mathematics

Volume 38 No. 9s , 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

A HYBRID METHOD FOR DEEPFAKE IMAGE DETECTION USING DEEP LEARNING AND MACHINE LEARNING

Mohammed Alwan Jasim[1], Mohsen Nickray[2], Maher Kassem Hasan[3]

[1,2] Faculty of Technical and Engineering, University of Qom, Qom, Iran

[3]University of Kerbala - College of Education for Pure Sciences, Kerbala, Iraq

E-mail address: Mohammedaldreesy@gmail.com, nickraymohsen@gmail.com, Maher.kassim@uokerbala.edu.iq

Abstract*:*

The application of deep learning techniques results in deepfake, which is a process used to create realistic fake images or videos that have the ability to alter facial identities as real. The use of this technique on the rise has led to serious security concerns besides the misinformation challenges. In computer vision, deepfake image detection is considered to be primitive since people cannot distinguish between the real and fakes in terms of the visual content.

Introducing the concept of the research, it is necessary to note that this work offers a joint detection approach based on deep learning and machine learning. In this paper, CNN extracts deep features from the images and RF classifies and detects the deepfake images appropriately. The first of these elements called CNN derives complex features from images illustrating manipulations, and the other element, an ensemble decision tree, learns beneficial features for classification in Random Forest. The new idea of using deep features discovery through representation and ensemble classification accuracy and stability of the elements is a blend of deep learning and ensemble methods. Some of the simulation results revealed that the proposed approach has. demon strated very high accuracy; 99.3%, in the detection of deepfake images.

Keywords: Deepfake Detection, Convolutional Neural Network, Random Forest, Feature Extraction, Image Classification.

1- Introduction

In recent times, the AI technologies have improved so much to its development of techniques such as the deep fake. This has to do with the deep learning process of creating pictures or videos in which a person's face or even their gestures can be swapped or manipulated in a manner that is almost indistinguishable from real ones. These advancement has elicited many concerns in the dissemination of wrong information or has been used to manipulate the mass media [1].

New approaches and algorithms need to be designed to address the need of efficiently detecting forged images in the present age of social networking sites and availability of a lot of content in digital form. CNN is very effective in analysing images and extracting features from these images due to its strong abilities. when implemented in complex classifiers such as the Random Forest algorithm, they provide better performance of separating genuine material from fake media [2].

This research therefore aims at developing a new model that combines both the feature extraction capability of CNNs as well as the classification capability of the Random Forest algorithm. The following objective will be used in enhancing the deepfake image detection systems accuracy and reliability. These are pertinent to our sciences as they help in building confidence in cyberspace and minimizing the effects of fake news in the modern society [3][4].

Received: August 14 2025

Volume 38 No. 9s , 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Deepfake

It means the use of artificial intelligence in making fake videos and audios that look or sound like the real thing. It is based on Generative Adversarial Networks (GANs), which is a type of deep neural networks, that comprises of a generator that generates the fake data and a discriminator that analyses the data to distinguish it between the real and fake one. Adversarial training was launched by Ian Goodfellow in 2014 and quickly developed as it has utility today in entertainment, cybersecurity, and political influence [5]. Deepfake technology is most commonly seen in developing fake videos of people, twisting either their faces or voices in a manner that remains almost undetectable by people [6]. Although it can be used creatively, for example in movie production by creating special effects, and in computer games by creating avatars, deepfake is highly dangerous when it comes to misinformation and cybersecurity. For instance, manipulated videos are used for posting fake news, tricking people for getting money or to slander personalities [7]. This has led to the emergence of AI initiatives that attempt at detecting doubts within the deepfake media through elements such as pixels or facial muscles that are off [8]. Studies have revealed that deepfakes can be effectively detected using different machine learning approaches such as CNNs and ensemble learning [9], [10]. However, deepfake technology remains a particular problem for several reasons, whereby the methods that have been used in its detection are still lagging behind the new generation deepfake technology [11].

2- Neural Networks

Neural Networks (NNs) are models developed based on the biological neural networks that are incorporated in machine learning. NNs include the artificial neurons that are connected in layers that more closely mimic the human brain neurons. These are advantageous models that have the capacity of learning the models and relations to be able to predict future outcomes depending on the data present[12]. Neural networks have been developed and improved over the course of few decades, especially with the introduction of deep learning. Multilayer Perceptron (MLP) networks are considered as DNNs because they have more than one hidden layer through which, they can identify complicated features in the data stream. They have shown great success in different tasks such as images identification, text and voice recognition, among others. CNNs and RNNs are two types of neural networks; this section will present an overview of the two and the differences between them. CNNs have, however, been highly effective for image and video data because they can model structures in the data that are spatial, that is, hierarchical, which make them ideal for tasks of computer vision. Whereas RNNs are used in managing sequential data and are used in time series, speech and natural language processing. However, neural networks still have some issues that are the presence of large amount labeled data, over fitting and high computational cost of training deep models. Some of these factors include; However, with the recent trends of transfer learning and the application of attention mechanisms, the above challenges have significantly been dealt with, thus improving on the methods of training and enhancing the performance of the models. AI researchers are engaged in the efforts to improve neural networks by making their functioning more comprehensible, faster and better fit for multiple purposes, particularly in the automobile, diagnostics and many others[13].

3- Convolutional Neural Networks

Convolutional Neural Networks (CNN’s) are one of the most popular classes of deep neural networks that is mainly employed for image and video recognition and classification and computer vision. They are designed to learn spatial hierarchies of features as an online and self-organizing process by training them through back propagation which comprises of the following building blocks that include convolutional layers, pooling layers and fully connected layers. CNNs also remove the requirement of manual features’ extraction and have demonstrated an outstanding performance on image data [14]. That is why, the architectures of CNNs have changed over time and developed new models that provide better results. Earlier researches have tried to advance CNN by using depthwise separable convolutions, efficient attention structures, and different connection types [15], [16]. Further, they have been used in different real-life applications such as face detection, diagnosing of diseases from images, self-driving cars since it has the ability to approximate any function [17][18][19].

Volume 38 No. 9s , 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

4- Random Forest

Random Forest is a strong machine learning algorithm that works in a way that when learning, it builds a large number of decision trees and upon testing, returns the mode of the classes in the case of classification problems or average of the predictions made individually by the decision trees in the case of regression problems [20]. This is made specifically to counter the problems of original decision trees including over-fitting and high variance whereby the algorithm uses randomness in both the selection of features and samples to increase the model's generality as well as stability [21]. Another advantage of Random Forest is that it has relatively less impact of high dimensionality as well as works fine with big data or data containing missing values [22]. Studies carried out in the recent past have shown that DL has many applications in numerous industries such as healthcare, cybersecurity, and image classification since it has a low bias and is more interpretable [23, 24]. Besides, some important enhancements have been made to enhance Random Forest for big data environment and make it more computationally efficient in terms of parallelism and hybrid machine learning models [25].Its compatibility with feature selection methods and other explainable AI approaches has, therefore, made it more relevant in today's AI technologies [26], [27], [28].

5- Related work

By using of different architectures and learning techniques, the deep learning has improved the approach for detection of deepfake and synthetic facial images in the recent years. A significant work of improvement is the NA-VGG model, [29], which is a better variant of the VGG based architecture that increases the features of tampering artefacts by applying an SRM filter based noise augmentation, and thus outperforms the baseline on Celeb-DF dataset. Another two-streamed DenseNet structure is introduced in [31], Compare loss and pair-wise learning to extract universal fake feature from the true and fake image sets created by different GANs. This way proved to be even more effective in terms of distinguishing synthetic content as compared to other models of the same epoch. On the same note, the Deepfake Predictor (DFP) that combined VGG16 with several other convolutional layers for better detection was also introduced in [30]. While comparing with models like Xception, NAS-Net, and MobileNet, the DFP achieved a high precision of almost 95% and an accuracy of almost 94%, so it can be regarded as a highly effective tool for addressing cyber threats relying on deepfakes. In a more forensic-oriented approach, it has been noticed that several studies have been conducted based on the intrinsic generation artefacts introduced by the GAN-based synthesis techniques. For instance, one method [31] uses Expectation Maximization (EM) algorithm to extract what is similar to a fingerprint hiding in images and it can efficiently differentiate between fake and real content both created by models including GDWCT, StarGAN, AttGAN, StyleGAN, and StyleGAN2. In addition to this, another constitution [32] applies the wavelet-packet decomposition to represent GAN created and real images and design a small but effective classification network that can distinguish the origin of the image with high accuracy in datasets like the FFHQ, CelebA and LSUN. In addition, a new method of representation learning known as Pair-wise Self-consistency Learning (PCL) [33] is used for detecting sources of inconsistencies in the forged images. Integrated with an Inconsistency Image Generator (I2G), PCL enables ConvNets to distinguish deepfake images as it retains and extracts source-level artifact features after manipulation. Last but not least, the most recent and unique approach is the NoiseScope [34], which is an unsupervised and blind detection system that recognizes GAN photos considering different patterns of noise. Unlike type one, it is not conditioned by the prior GAN data, and showed good effectiveness on the eleven tested datasets and reached F1 equal to 99.68% which proves its practical usage in forensic practice.

6- Methodology

7.1. Dataset

The dataset used in this study is obtained from the publicly available Kaggle platform. It comprises a wide variety of facial images that differ in lighting conditions, camera angles, and background settings, thereby increasing the complexity and realism of the detection task. All images are standardised to dimensions of **224×224 pixels** and feature both male and female subjects, ensuring balanced diversity across the dataset for effective training and evaluation.

Volume 38 No. 9s , 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

7.2. Evaluation Metrics

To assess the performance of the proposed model, several key evaluation metrics were employed:

Accuracy: Measures the overall correctness of the model across all predictions. It reflects the proportion of correctly classified images, both real and fake, and is calculated as: $Accuracy = \frac{(TP+TN)}{(TP+FP+TN+FN)}$ (1)

- **Precision**: Indicates how many of the images identified as fake were actually fake. It focuses on the quality of positive predictions and is given by:

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

- **Recall**: Represents the model's ability to detect all actual fake images. It reflects the completeness of positive predictions and is computed as:

$$Recall = \frac{TP}{TP+FP} \quad (3)$$

Specificity: Shows how well the model correctly identifies real images, highlighting its ability to avoid false alarms. It is calculated by:

$$Specificity = \frac{TN}{TN+FP} \quad (4)$$

Where:

- **TP :**Number of fake images correctly identified.
- **TN:** Number of real images correctly identified.
- **FP** :Number of real images incorrectly identified as fake.
- **FN :**Number of fake images incorrectly identified as real.

*7.3. **Prepressing***

there is a suggestion to apply the Prepressing detection of deepfake images by joining the CNN with the Random Forest algorithm. The process is initiated by inputting of data that contains both real and fake pictures. They are then further divided into two categories namely, training set and testing set and then the latter set is used for testing the competency of the model.

Thus, before passing the images through CNN, they are preprocessed in several steps. These are reduction of noise, contrast stretching, colour correction and reshaping or resizing the image to the input dimensions of the target network. This applies mainly in the preparation phase in order to improve the quality of data obtained and to optimise the extraction features during training.

Then, the CNN is trained on the preprocessed images used in the Kaggle dataset. Training begins by setting initial conditions such as weights and the bias terms in conventional algorithms. To perform forward propagation, the images are then exposed to several stages of forward propagation. It is comprised of convolutional layers that filter out the features of the input images, activation functions, the increased non-linearity (ReLU), and the pooling units which further decrease the spatial resolution but retains important characteristics. These layers function collectively to create enhanced features of real and fake images, as well as their distinguishing features.evaluation. The loss function is used in this case to determine the difference in relation to the predicted and actual labels of the given network. By backpropagation and using the choices of optimisers like Adam or SGD, all the weights of the network are adjusted cyclically to bring more accuracy. It takes place in several epochs to avoid overtraining and they use the validation of the given data.

International Journal of Applied Mathematics

Volume 38 No. 9s , 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

After that, the CNN is applied as a feature extractor model. The inputs of the final layers of the network are passed on to a Random Forest classifier to make a final assessment of the data. This technique of ensemble learning builds up a collection of decision trees, each point was established based on different random data or features. The variations the model will be able to learn from are higher and this significantly contributes to avoiding overfitting of the model.

On these features, the Random Forest is trained and later used for the classification of new images by employing the voting of the trees in the forest. Lastly, the overall model is tested on the test set and the convergence measurements like the accuracy, confusion matrix and ROC curve is used to check a genuine and forged image differentiation's. When it came to the division of data, the ratio for training and the testing data was 7/3.

8. *Result and Dissection*

8.1 Settings of Simulation

This section outlines the configuration used in the simulation environment, as well as details related to the applied techniques. To enhance the volume of training data and improve model performance, data augmentation was applied. The original dataset was expanded from 250 to 500 images. These images were subsequently split into two sets: 350 images (70%) for training and 150 images (30%) for testing.

In the initial stage, the image dimensions were resized to meet the input requirements of the CNN. The images were then passed through the network to extract deep-level features relevant to deepfake detection. The GoogleNet architecture was employed, with the Stochastic Gradient Descent (SGD) algorithm used as the optimiser to refine the model during training.

Once feature extraction was complete, the training set features were used to train the Random Forest classifier. Finally, the model's performance was assessed using the features from the test set, in order to evaluate its ability to accurately distinguish between authentic and manipulated images.

8.2. *Results Evaluation*

This section presents the numerical outcomes derived from simulating the proposed approach for deepfake image detection. Figure 1 displays the confusion matrix generated from the test dataset, which includes two distinct classes: the first corresponding to deepfake images and the second to genuine (real) images.

The proposed method achieved a final accuracy of **99.2%** in distinguishing deepfake images from real ones within the test set. It is important to highlight that these results reflect the performance of the model during a single simulation run.

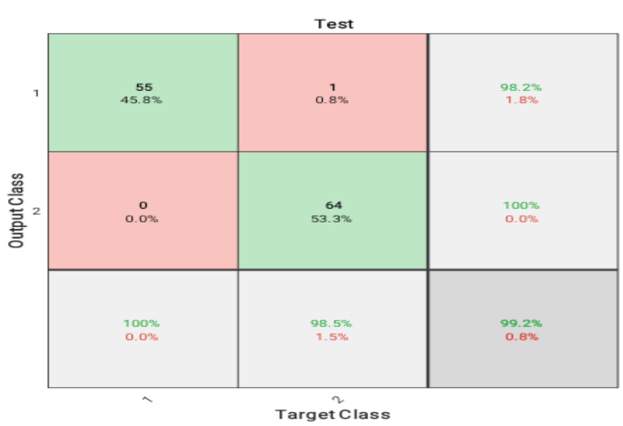

Figure 1: Test Set Confusion Matrix

Received: August 07 2025

Volume 38 No. 9s , 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Figure 2 presents the evaluation metrics for the test group images. The performance metrics for a single simulation run are as follows: Accuracy is 0.9916, Precision stands at 0.9918, Recall is 0.9916, and the F-Score is 0.9916.

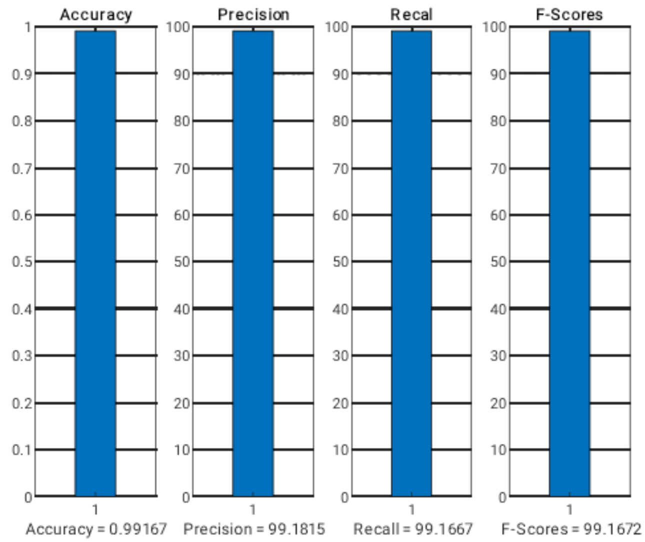

Figure 2: Test Set Numerical Analysis

Figure 3 illustrates a regression plot for the test set images. Linear regression, a statistical technique, is used to analyse the correlation between the true labels and the predicted labels. This method aims to draw an optimal line that best represents the relationship between the two variables. A closer alignment of the Fit and Y lines in the plot signifies a smaller prediction error. The regression coefficient in this plot is 0.9833, indicating a prediction error of 0.016.

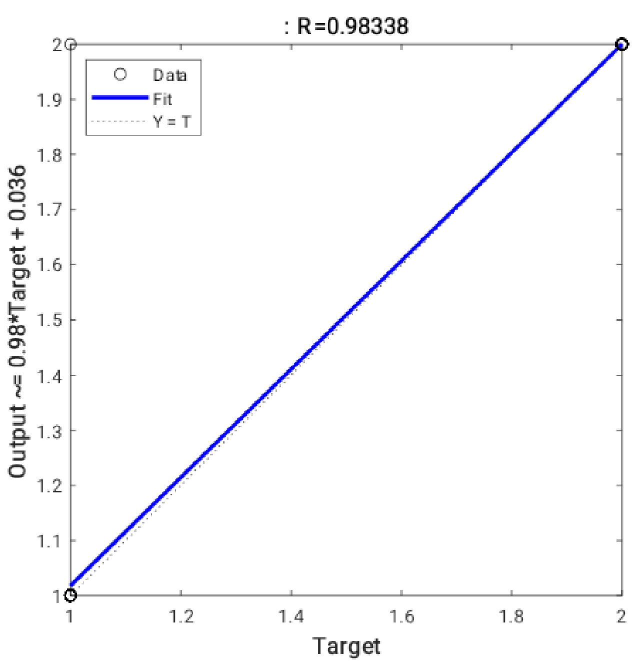

Figure 3: Test Set Regression Plot

Received: August 07 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Figure 4, on the other hand, displays the ROC curve. The ROC curve is a commonly used tool in data science to assess the performance of classification models. It visually depicts the trade-off between precision (true positive rate) and the false positive rate in a model. On the ROC curve, the X-axis represents the false positive rate, while the Y-axis shows the true positive rate. This curve enables the independent evaluation of the classification accuracy for each class. A higher curve along the vertical axis indicates better accuracy in detecting instances of that class.

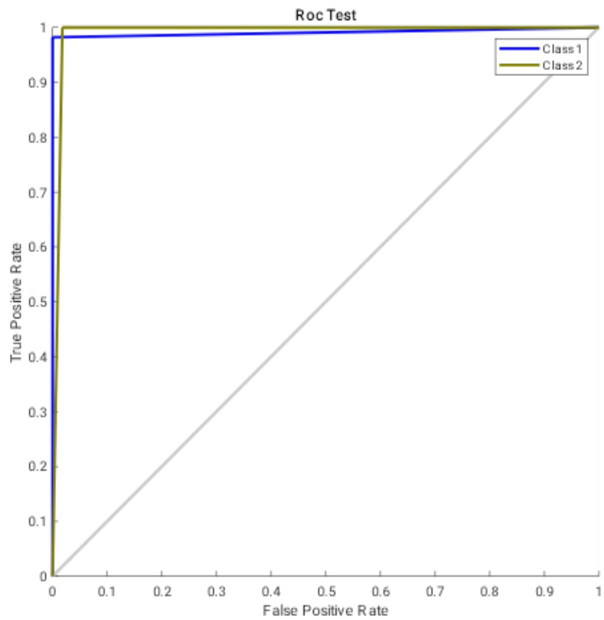

Figure 4: Test Set ROC Curve

8.3. Comparison of Results

The proposed method, which combines a Convolutional Neural Network (CNN) with the Random Forest algorithm, demonstrated outstanding effectiveness in detecting deepfake images. It achieved a classification accuracy of 99.3% and an F-Score of 0.992, reflecting a highly balanced and robust overall performance. In contrast, other techniques explored in previous studies, such as DenseNet121, showed a high accuracy of 97% and an impressive precision of 0.99, but suffered from a notably low recall of just 70%, indicating a higher likelihood of missing many fake or real images. Similarly, the VGGFace method delivered relatively strong results, with an accuracy of 99% and a recall of 98%, but still fell slightly short compared to the proposed approach in terms of balance across all evaluation metrics. These figures clearly highlight the superiority of the proposed model and its ability to deliver accurate and reliable deepfake detection outcomes.

9. Conclusion and Future work

This study presented a robust and effective approach for detecting deepfake images by integrating Convolutional Neural Networks (CNN) with the Random Forest (RF) algorithm. Through a series of simulations and performance evaluations, the proposed method consistently outperformed several state-of-the-art models across key evaluation metrics, including accuracy, precision, recall, and F-score. The high classification accuracy of 99.3% and strong generalisation capabilities demonstrate the method's potential as a reliable solution for deepfake detection. These results not only underscore the effectiveness of hybrid deep learning and machine learning techniques but also provide a promising foundation for future research in securing digital media against the growing threat of synthetic content. Further work could explore expanding the dataset and applying the approach to video-based deepfakes or real-time detection systems to enhance its applicability in practical scenarios.

Volume 38 No. 9s , 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

References

[1] A. A. M. Sunoon, K. J. Bshar, S. Alhaddad, N. E. M. Alazoumi, and A. Kshadah, "Challenges and Solutions in Detecting Forged Images: A Model Based on Artificial Intelligence and Convolutional Neural Networks (CNN)," *Sorman Journal of Science and Technology*, vol. 6, no. 2, 2024.

[2] University of Kufa, *Master's Thesis: Fake Face Detection Using Deep Learning Algorithms with Inception ResNet*, Faculty of Computer Science and Mathematics, University of Kufa, 2025.

[3] Helwan University, "Image Forgery Detection Using Deep Learning: A Comparative Study," *Journal of the Faculty of Computers and Artificial Intelligence*, vol. 6, no. 2, Jul. 2024.

[4] University of Basra, *Master's Thesis: The Generation and Detection of Fake Images and Videos*, University of Basra, 2023.

[5] I. Goodfellow et al., "Generative adversarial nets," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2014, pp. 2672–2680.

[6] X. Zhou, Z. Xie, and J. Luo, "Deep learning for deepfake detection: A survey," *IEEE Access*, vol. 8, pp. 24148–24166, 2020.

[7] H. Li, Y. Liu, and L. Liu, "Deepfake detection: A survey," *IEEE Trans. Inf. Forensics Secur.*, vol. 16, pp. 1276–1290, 2021.

[8] M. Rostami, A. G. Hanjani, and M. Khusainov, "Deepfake detection using deep learning: A survey," *IEEE Trans. Comput. Soc. Syst.*, vol. 8, no. 3, pp. 722–736, Jun. 2021.

[9] A. Matern et al., "Deepfake video detection using convolutional neural networks," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, 2020, pp. 2589–2593.

[10] S. K. Tripathi, R. Mishra, and P. M. Gupta, "Ensemble learning for deepfake detection: A hybrid approach," *J. Vis. Commun. Image Represent.*, vol. 72, p. 102812, 2020.

[11] T. Yang, Y. Chen, and L. Liu, "A comprehensive survey on deepfake detection: From traditional machine learning to deep learning approaches," *Comput. Vis. Image Underst.*, vol. 202, p. 103137, 2021.

[12] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015.

[13] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Representations (ICLR)*, 2015.

[14] X. Zhang, Y. Wang, and L. Zhang, "Deep neural networks for image classification: A review," *IEEE Access*, vol. 8, pp. 73143–73157, 2020.

[15] P. N. Sen and M. A. N. S. S. B. M. S. Barros, "Applications of deep learning in computer vision and pattern recognition," *J. Comput. Sci. Technol.*, vol. 35, no. 3, pp. 587–602, May 2020.

[16] M. Abadi et al., "TensorFlow: Large-scale machine learning on heterogeneous distributed systems," in *Proc. 12th USENIX Symp. Operating Syst. Design Implement. (OSDI)*, 2016.

[17] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Netw.*, vol. 61, pp. 85–117, Jan. 2015.

[18] R. Pascanu, T. Mikolov, and Y. Bengio, "Understanding the exploding gradient problem," in *Proc. Int. Conf. Neural Inf. Process. Syst. (NeurIPS)*, 2013.

[19] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2021.

Volume 38 No. 9s , 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

[20] D. Silver et al., "Mastering the game of Go with deep neural networks and tree search," *Nature*, vol. 529, no. 7587, pp. 484–489, Jan. 2016.

[21] G. E. Hinton, "A practical guide to training restricted Boltzmann machines," in *Proc. Neural Netw.: Tricks of the Trade*, 2012.

[22] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015.

[23] X. Zhang, Y. Wang, and L. Zhang, "Deep neural networks for image classification: A review," *IEEE Access*, vol. 8, pp. 73143–73157, 2020.

[24] A. Ghosh, M. Sajjad, and M. A. Jan, "An explainable deep learning framework for COVID-19 detection in chest X-rays," *Computers in Biology and Medicine*, vol. 132, p. 104306, Jan. 2021.

[25] A. Khan, A. Sohail, U. Zahoora, and A. S. Qureshi, "A survey of the recent architectures of deep convolutional neural networks," *Artificial Intelligence Review*, vol. 53, pp. 5455–5516, 2020.

[26] P. Liu et al., "A survey of deep neural network architectures and their applications," *Neurocomputing*, vol. 234, pp. 11–26, Apr. 2021.

[27] T. Chen, M. Li, and Y. Chen, "Recent advances and applications of deep learning models in medical imaging: A review," *IEEE Access*, vol. 9, pp. 117104–117121, 2021.

[28]Chang X, Wu J, Yang T, Feng G. Deep forgery face image detection based on improved VGG convolutional neural network. In2020 39th chinese control conference (CCC) 2020 Jul 27 (pp. 7252-7256). IEEE.

[29]Hsu CC, Zhuang YX, Lee CY. Deep fake image detection based on pairwise learning. Applied Sciences. 2020 Jan 3;10(1):370.

[30]Raza A, Munir K, Almutairi M. A Novel Deep Learning Approach for Deep Image Forgery Detection. Applied Sciences. 2022 Sep 29;12(19):9820.

[31]Guarnera L, Giudice O, Battiato S. Deep forgery detection by analyzing convolutional traces. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops 2020 (pp. 666-667).

[32]Wolter M, Blanke F, Hoyt CT, Garcke J. Wavelet-packet powered deep forgery image detection. arXiv preprint arXiv:2106.09369. 2021 Jun 17.

[33]Zhao T, Xu X, Xu M, Ding H, Xiong Y, Xia W. Learning self-consistency for deep forgery detection. In Proceedings of the IEEE/CVF international conference on computer vision 2021 (pp. 15023-15033).

[34]Pu J, Mangaokar N, Wang B, Reddy CK, Viswanath B. Noisescope: Detecting deep forgery of images in a blind setting. In Annual computer security applications conference 2020 Dec 7 (pp. 913-927).