University of Bisha Journal for Basic and Applied Sciences

Volume 1 | Issue 1 Article 5

2025

Machine Learning Approach for Fake Image Forensics

Abdullah Mujawib Alashjaee
Department of Computer Sciences, Faculty of Computing and Information Technology, Northern Border University, Rafha 91911, Kingdom of Saudi Arabia, abdullah.alashjaee@nbu.edu.sa

UNIVERSITY OF BISHA JOURNAL FOR BASIC AND APPLIED SCIENCES 2025;1:44–49

RESEARCH ARTICLE

Machine Learning Approach for Fake Image Forensics

Abdullah Mujawib Alashjaee

Department of Computer Sciences, Faculty of Computing and Information Technology, Northern Border University, Rafha 91911, Kingdom of Saudi Arabia

ABSTRACT

Digital forensics is crucial in investigating and analyzing digital evidence, including images, to uncover potential crimes and identify manipulated or fraudulent content. A significant challenge in digital forensics is distinguishing between authentic and manipulated images. This paper applies a popular pre-trained deep learning model, VGG16, for feature extraction in real vs. fake image detection and various machine learning (ML) models as classifiers. The feature extraction part until the flatten layer of the VGG16 model is utilized to extract discriminative features that capture high-level image representations comprising local and global patterns. ML models, including Random Forest, Logistic Regression, Decision Tree, and k-Nearest Neighbors, are explored to distinguish between real and fake images. ML models trained on a labeled dataset, encompassing a wide range of authentic and manipulated images, to learn the underlying patterns and correlations. The experimental results demonstrate that a combination of VGG16-based features and the powerful learning capabilities of the RF provides the highest detection accuracy of 78.56 %, thereby enhancing the efficacy of forensics investigations in detecting image forgeries and ensuring the integrity of digital evidence.

Keywords: Digital forensics, Machine learning, Cybercrime, Digital evidence

1. Introduction

Nowadays, data produced from different sources, such as medical sectors, businesses, and industries, exponentially increases with each passing minute, immersing us in a wealth of potentially untrue material [1]. Since the 1990s, a digital revolution has changed the globe and our way of life. The internet, mobile phones, and other digital services and tools have merged into our daily lives. Emails, digital photographs, and phone books are just a few examples of the vast volumes of data and essential information that have also increased. Law enforcement requirements have changed due to this situation [2].

Machine learning (ML) and artificial intelligence can be seen as the fundamental component of behavioral forensics [3, 4]. Software for pattern recognition can deal with enormous amounts of data through ML to reduce the success of any criminal activities. These actions may range from burglary and money laundering to incursion attacks. Using networked software and tools with ML can entail remote analysis more effectively. For instance, to determine whether video evidence is authentic, ML methods can recognize any instances of frame deletion by extracting discriminative features from the reconstructed video images and bitstream. Then, ML techniques can be employed to indicate the true positive or false negative rates.

Recently, ML methods have shown great potential in Digital forensics (DF) [5–8]. In [9], the author showed ML models' role and usage in DF and investigations. In another work [10], the authors claimed that cyber infrastructures are extremely vulnerable to intrusions. On the other hand, human intervention and physical devices are insufficient for monitoring and protecting these infrastructures. Therefore, there is a need for more sophisticated cyber-defense systems that need to be flexible, adaptable, and able

Received xx xx 2025; revised xx xx 2025; accepted xx xx 2025.
Available online 1 January 2025

E-mail address: abdullah.alashjaee@nbu.edu.sa (A. M. Alashjaee).

https://doi.org/xx.xxxxx/xxxx-xxxx.1005

to detect such types of threats and make intelligent decisions. Then, they highlighted the role of applying ML methods in combating cybercrimes and preventing cyber-attacks.

In [11], a generic framework for diverging deep learning cognitive computing techniques into cyber forensics is proposed. They used neural networks to simulate human decision-making. Based on these grounds, Deep learning (DL) holds the potential to dramatically change the domain of cyber forensics and provide solutions to forensic investigators. In [12], proposed a model that can detect copy-move and splice attacks in color images using Local Binary Pattern (LBP) and Discrete Cosine Transformation (DCT). The proposed model is evaluated using a Support Vector Machine (SVM) model. The DCT and LBP operators capture the changes in the local frequency distribution and detect micro-patterns. The resulting images are then classified into authentic and tampered ones using the SVM model. The results showed that the proposed method is suitable for image forgery detection.

In [13], a hybrid model is proposed to address the issues related to the IP reputation system. This approach combined the abilities of various DF techniques such as ML, Dynamic Malware Analysis, and Cyber Threat Intelligence. They evaluated the system against various existing systems using several ML techniques. The results showed that the Decision Tree (DT) performed better than other comparative models. In [14], a framework is used to detect distributed denial of service attacks using the K-Nearest Neighbor (KNN) and Nave Bayes algorithms. The method utilizes statistical techniques to enhance anomalous network traffic detection performance. The KNN method performed better than the other model when the authors used the Knowledge Discovery and Data Mining Cup 1999 (KDDCup99) competition dataset and network security laboratory (NSLKDD) dataset to evaluate the models.

In [15], a multi-label approach is used to organize and analyze emails. This approach employed data preprocessing, which removes the most repeated words in a given sentence. Then, various ML techniques are applied to extract normal email features from harmful emails. Results showed that the Logistic Regression (LR) model performed better than other techniques used in their work. In [16], the K-Means algorithm is used to cluster the data collected from various sources from multiple forms of cybercrime. In the Clustering stage, the K-means algorithm detects the interactions between features. The proposed method can then provide actionable steps to prevent these crimes from succeeding.

Convolutional Neural Networks (CNNs), Vision Transformers (ViT), and Generative Adversarial Networks (GANs) for fake images are used [17]. The results showed that the CNN model achieved better accuracy than other tested models. In [18], a fake image classification approach using neural networks (NN) is presented. The results showed that the presented scheme can effectively classify fake images from real ones. An improved Visual Geometry Group (VGG) network named NA-VGG is proposed to detect Fake face images [19]. The results showed that the developed NA-VGG gives significant improvements compared to other ML models in their work.

ML can provide valuable information for DF investigations in detecting fake images and ensuring the integrity of digital evidence. Therefore, more efforts are needed to detect fake images effectively. This paper uses a pre-trained VGG16 model to extract high-level features from the face images dataset. Then, LR, KNN, DT, Artificial NN (ANN), and Random Forest (RF) models are employed to distinguish fake images from real ones.

The remaining parts of this paper are structured as follows: Section 2 briefly describes the dataset and pre-trained model-based feature extraction. In Section 3, experimental results, evaluation metrics and results are presented. The concludes of this work with future research directions are provided in Section 4,

2. Methods and materials

2.1. Dataset

A “Real and Fake Face Detection” dataset by the Computational Intelligence and Photography Lab at Yonsei University detects authentic and modified images [20]. The dataset aims to address the issue of fake identities on social networks by providing expert-generated high-quality photoshopped face images. These images are composite images created by separating different facial parts like eyes, nose, mouth, or face. The dataset is valuable for training classifiers to distinguish between real and fake face images. The dataset consists of two classes: Real and Fake. The Real class comprises 1081 real face photos, and the Fake class has 960 expert-generated fake face photos. Each image is in RGB format, with 600×600 pixels in dimension. The fake photos are further categorized into three groups: easy, mid, and hard, although these categories are subjective and may not be explicitly reliable. The filenames of the fake images indicate which parts of the faces have been replaced. While generative models like GANs can generate fake face images, they may not capture the intricacies of human-generated fake images.

Using expert-level fake face photos, the dataset aims to train classifiers to detect these more realistic and deceptive counterfeits. This dataset allows one to explore the differences between machine-generated and expert-generated fake face images and develop robust detection techniques.

2.2. Feature extraction

VGG16 is a famous CNN architecture widely used for image classification tasks [21]. Feature extraction with VGG16 involves utilizing the pre-trained model to extract high-level features from images for real and fake face detection tasks. VGG16 has 16 layers, comprising 13 convolutional and three fully connected layers. The model is trained on a large-scale dataset called ImageNet, which contains millions of images from various object categories. Due to this extensive training, VGG16 has learned to recognize various visual patterns and features. The following steps are followed to perform feature extraction for real and fake face detection:

Preprocessing: The input image size expected by VGG16 must be consistent. In this work, several image sizes are investigated. All input images are initially 600×600 pixels with RGB channels. This high image size will increase feature dimensionality, so resizing is necessary. For smaller resize factors, the output of VGG16 remains very high and generates a higher feature vector dimension, which may not suit most ML models. Larger resize factors decrease feature dimensionality but may lose essential features. All images are resized to 150×150 pixels based on a trade-off between maximum accuracy and feature dimension. The pixel values are also normalized to a range [0, 1].

Loading VGG16: Load the open-source pre-trained VGG16 model. This model has been trained on vast data, allowing it to capture intricate image features. The classification layers of VGG16 are removed, leaving the convolutional layers intact. These convolutional layers serve as feature extractors. Each image is passed through the network, obtaining the output of the last convolutional layer.

Feature Extraction: The VGG16 can be divided into 'feature extraction' comprising 13 layers from input to the last convolutional layer and 'classifier' comprising three fully connected layers. The output of the feature extraction part (at the 13^{th} layer) is flattened to obtain a feature vector representation for each image. A vector of 73984 features represents each image. This feature vector captures high-level features extracted by VGG16, encapsulating patterns that can discriminate between real and fake faces. By leveraging the knowledge and learned features of VGG16, the feature extraction process captures discriminative information in the form of feature vectors. A classifier can then utilize these vectors to differentiate between real and fake faces. The success of VGG16 feature extraction hinges on having a diverse and well-labeled dataset, choosing an appropriate classifier, and employing effective training strategies. Fine-tuning the model or incorporating additional techniques like data augmentation may be necessary to achieve optimal results for real and fake face detection.

Training: The feature vectors and their corresponding labels (real/fake) are used to train various ML algorithms as classifiers. The performance of the ML models can be assessed on a separate validation or test set. The following section discusses the hyper-parameters tunning and evaluation measures.

3. Experimental results

3.1. Experimental setup

As described earlier, real and fake face detection is assessed by conducting experiments on the dataset. All the methods are executed per their implementations in their original work, and the hyper-parameters settings are given in Table 1. The grid-search method optimizes the hyper-parameters by maximizing accuracy. The preprocessing and ML model training and evaluations are conducted on a computer with 3.13 GHz PC with 32 GB RAM and Ubuntu 22.04.1 LTS operating system. The VGG16 is obtained from the Tensorflow module and ML models are implemented using Scikit-learn.

3.2. Evaluation measures

In the context of real vs fake face detection, the evaluation metrics Accuracy, Precision, Recall, and F1-score play an essential role in assessing the performance of the detectiong system.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FN + FP} \tag{1}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

Table 1. Hyper-parameters for ML models.

Method	Parameters
LR	Penalty = *L2*, fit_intercept = *true*, solver = *lbfgs*, max_iter = *150*
KNN	#Neighbors = *35*, weights = *uniform*, distance = *euclidean*
DT	Criterion = *gini*, min_samples_split = *6*, min_samples_leaf = *3*, max_features = *25*
ANN	Layers= [*256, 32,2*], Epochs = *100*, batch_size = *100,* validation_split = *0.2,* loss = *categorical_crossentropy*, optimizer = *ADAM*, learning rate = *0.001*
RF	#Estimators = *500*, bootstrap = *True* min_samples_split = *5*, min_samples_leaf = *10,*

$$\text{F1}-\text{score} = \frac{2PR}{P+R} \tag{4}$$

True Positive and (TP) and True Negative (TN) denote the samples of customers correctly detected as churner or not, while False Negative (FN) and False Positive (FP) represents the number of misclassified positive and negative cases, respectively.

Accuracy: Accuracy measures the overall correctness of the detection system's predictions. It is the ratio of the correctly classified instances (TP and TN) to the total number of instances. In real and fake face detection, accuracy indicates how well the system correctly identifies real and fake faces. A higher accuracy value signifies a more reliable detection system.

Precision: Precision, also known as positive predictive value, measures the system's ability to identify TP among all positive predictions (TP + FP). It is the ratio of true positives to the sum of TP and FP. In real and fake face detection, precision indicates the proportion of correctly identified fake faces among all the faces predicted as fake. A higher precision value indicates fewer false positives, implying that the system has a lower tendency to label real faces as fake.

Recall: Recall, also known as sensitivity or TP rate, measures the system's ability to identify TP among all actual positive instances (TP + FN). It is the ratio of true positives to the sum of true positives and false negatives. In real and fake face detection, recall indicates the proportion of correctly identified fake faces among all the actual fake faces. A higher recall value indicates fewer false negatives, implying that the system has a lower tendency to miss fake faces and can accurately identify them.

F1-score: The F1-score is the harmonic mean of precision and recall. It provides a balanced measure that considers precision and recall, making it useful when an imbalance between classes or FP and FN needs to be minimized. In real and fake face detection, the F1-score captures the balance between correctly identifying fake faces (precision) and not missing any fake faces (recall). A higher F1-score indicates a better overall performance of the system in detecting fake faces.

By considering these evaluation metrics together, a comprehensive understanding of the performance of the detection system can be obtained. High accuracy, precision, recall, and F1-score indicate a robust and reliable system that can effectively differentiate between real and fake faces, contributing to enhanced digital forensics and security measures.

3.3. Experimental results and discussion

Fig. 1 compares different machine learning models for real vs fake face image detection based on various evaluation metrics: Accuracy, Precision, Recall, and F1-score.

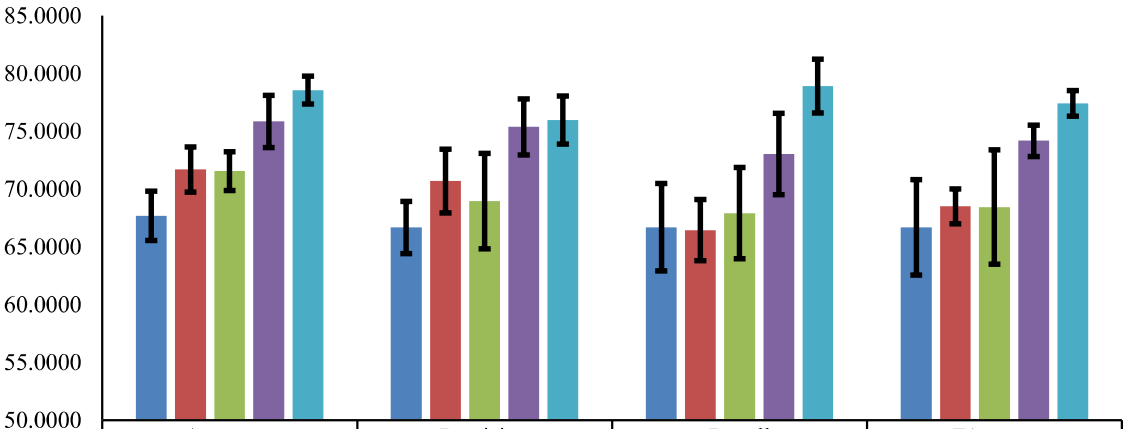
Accuracy represents the overall correctness of the model's predictions. It is the ratio of correctly identified real and fake face images to the total number of face images. The highest accuracy is observed for the RF model, with an average accuracy of 78.563 %, followed by ANN with 75.850 %. Lower values indicate less accurate models.

Precision measures the model's ability to identify TP among all positive predictions. It is the ratio of true positives to the sum of TP and FP. Higher precision indicates fewer false positives. The highest precision is observed in the table for the RF model, with an average precision of 75.97%, followed by the ANN with 75.37%.

Recall, also known as sensitivity or TP rate, measures the model's ability to correctly identify true positives among all actual positive instances. It is the ratio of true positives to the sum of true positives and false negatives. Higher recall indicates fewer false negatives. The table shows the highest recall for the RF model, with an average recall of 78.91 %, followed by the ANN with 73.02 %.

The F1-score is the harmonic mean of precision and recall and provides a balanced measure of a model's performance. It considers both precision and recall and is especially useful when an imbalance between classes exists. In the table, the highest F1-score is observed for the RF model with an average F1-score of 77.4121%, followed by the ANN with 74.1764%.

Based on the comparison in the table, the RF model shows the highest performance across all evaluation



	Accuracy	Precision	Recall	F1-score
LR	67.6800	66.6800	66.7000	66.6900
KNN	71.6900	70.6900	66.4500	68.5045
DT	71.5600	68.9600	67.9200	68.4360
ANN	75.8500	75.3700	73.0200	74.1764
RF	78.5630	75.9700	78.9100	77.4121

Fig. 1. Quantitative comparison of ML models for real and fake face image detection using VGG16-based features.

metrics, indicating its effectiveness in real vs fake face image detection. However, considering other factors, such as computational complexity, interpretability, and scalability, is important when choosing the most suitable model for a specific application.

4. Conclusion and future works

This research paper focused on real vs fake face detection using feature extraction with the pre-trained DL model and evaluated the performance of different machine learning models. The VGG16 architecture extracts high-level features from the face images. Several ML models were compared, including LR, KNN, DT, ANN, and RF. The results demonstrated that the RF model consistently outperformed the other models across all metrics, achieving the highest accuracy, precision, recall, and F1-score. The efficacy of the RF model in distinguishing between real and fake face images is highlighted. Our findings suggest that the combination of VGG16 feature extraction and ML models holds promise for real vs fake face detection in digital forensics. The implications of this research extend to various domains, including social media, online identity verification, and security applications. By accurately detecting manipulated or forged face images, the risks associated with fake identities and fraudulent activities can be mitigated, making the online environment a safer place. Future research directions could explore advanced deep learning architectures, incorporating additional features such as texture analysis or facial landmarks and expanding the dataset to include a broader range of real and fake faces.

Conflict of interest

Author contribution

Data availability

References

1. Carvalho, T., Faria, F. A., Pedrini, H., Torres, R. D. S., and Rocha, A. (2015) Illuminant-based transformed spaces for image forensics. *IEEE Transactions on Information Forensics and Security*, 11(4), 720–733.
2. Ding, F., Zhu, G., Alazab, M., Li, X., and Yu, K. (2020) Deep-learning-empowered digital forensics for edge consumer electronics in 5G HetNets. *IEEE Consumer Electronics Magazine*, 11(2), 42–50.
3. Bhatt, P., and Rughani, P. H. (2017) Machine learning forensics: A new branch of digital forensics. *International Journal of Advanced Research in Computer Science*, 8(8).
4. Costantini, S., De Gasperis, G., and Olivieri, R. (2019) Digital forensics and investigations meet artificial intelligence. *Annals of Mathematics and Artificial Intelligence*, 86(1–3), 193–229.
5. Mitchell, F. (2010) The use of artificial intelligence in digital forensics: An introduction. *Digital Evidence & Elec. Signature L. Rev.*, 7, 35.
6. Ariu, D., Giacinto, G., and Roli, F. (2011, October) Machine learning in computer forensics (and the lessons learned from machine learning in computer security). In *Proceedings of the 4th ACM workshop on Security and artificial intelligence*, 99–104.

7. Hall, S. W., Sakzad, A., and Choo, K. K. R. (2022) Explainable artificial intelligence for digital forensics. *Wiley Interdisciplinary Reviews: Forensic Science*, 4(2), e1434.
8. Arshad, M. Z., Rahman, H., Tariq, J., Riaz, A., Imran, A., Yasin, A., and Ihsan, I. (2022) Digital Forensics Analysis of IoT Nodes using Machine Learning. *Journal of Computing & Biomedical Informatics*, 4(01), 1–12.
9. Bhatt, P., and Rughani, P. H. (2017) Machine learning forensics: A new branch of digital forensics. *International Journal of Advanced Research in Computer Science*, 8(8).
10. Dilek, S., Çakır, H., and Aydın, M. (2015) Applications of artificial intelligence techniques to combating cyber crimes: A review. *arXiv preprint arXiv:1502.03552*.
11. Karie, N. M., Kebande, V. R., and Venter, H. S. (2019) Diverging deep learning cognitive computing techniques into cyber forensics. *Forensic Science International: Synergy*, 1, 61–67.
12. Islam, M. M., Karmakar, G., Kamruzzaman, J., Murshed, M., Kahandawa, G., and Parvin, N. (2018, December) Detecting splicing and copy-move attacks in color images. In *2018 Digital Image Computing: Techniques and Applications (DICTA)*, IEEE. 1–7.
13. Usman, N., Usman, S., Khan, F., Jan, M. A., Sajid, A., Alazab, M., and Watters, P. (2021) Intelligent dynamic malware detection using machine learning in IP reputation for forensics data analytics. *Future Generation Computer Systems*, 118, 124–141.
14. Kachavimath, A. V., Nazare, S. V., and Akki, S. S. (2020, March) Distributed denial of service attack detection using naïve bayes and k-nearest neighbor for network forensics. In *2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA)*, IEEE. 711–717.
15. Hina, M., Ali, M., Javed, A. R., Srivastava, G., Gadekallu, T. R., and Jalil, Z. (2021, October) Email classification and forensics analysis using machine learning. In *2021 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/IOP/SCI)*, IEEE. 630–635.
16. Sudha, T. S., and Rupa, C. (2019, March) Analysis and Evaluation of Integrated Cyber Crime Offences. In *2019 Innovations in Power and Advanced Computing Technologies (i-PACT)*, IEEE. 1, 1–6.
17. See-To, J. (2023) Fake image detection through automatic fact checking. https://hdl.handle.net/10356/166830.
18. Lu, W., Sun, W., Huang, J. W., and Lu, H. T. (2008, July) Digital image forensics using statistical features and neural network classifier. In *2008 International Conference on Machine Learning and Cybernetics*, IEEE. 5, 2831–2834.
19. Chang, X., Wu, J., Yang, T., and Feng, G. (2020, July) Deepfake face image detection based on improved VGG convolutional neural network. In *2020 39th Chinese Control Conference (CCC)*, IEEE. 7252–7256.
20. Nam, Seonghyeon, Oh, Seoung Wug, Kang, Jae Yeon, Shin, Chang Ha, Jo, Younghyun, Kim, Young Hwi, Kim, Kyungmin, Shim, Minho, Lee, Sungho, Kim, Yunji, Han, Suho, Nam, Gunhee, Lee, Dasol, Jeon, Subin, Cho, In, Cho, Woongoh, Yang, Sejong, Kim, Dongyoung, Kang, Hyolim, Hwang, Sukjun, and Kim, Seon Joo. (2019, January) Real and Fake Face Detection, Version 1. Retrieved [10 Feb 2023] URL: https://www.kaggle.com/datasets/ciplab/real-and-fake-face-detection.
21. Simonyan, K., and Zisserman, A. (2014) Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.