International Journal of Research and Review
[Indexed, Peer-Reviewed Journal]
Volume 13; Issue: 5; May 2026
Website: www.ijrrjournal.com
Research Paper *E-ISSN: 2349-9788; P-ISSN: 2454-2237*

Deep Fake and Image Forgery Detection using Machine Learning

Pamula Kamakshi[1], Veeragani Harika[1], Kowsika Paladugu[2], Pentapati Karthikeya[3], Brahma Teja Rayapaneni[4], Sai Manaswy Manukonda[5]

[1,2,3,4,5]Department of Information Technology,
Dhanekula Institute of Engineering and Technology College, JNTUK, Vijayawada, India.

Corresponding Author: Veeragani Harika

DOI: https://doi.org/10.52403/ijrr.20260559

ABSTRACT

The increasing prevalence of advanced digital image editing capabilities makes it harder to detect image fakes using traditional methods and creates major problems for digital forensics, journalism and courtroom proceedings. In this article we explain Image Guard AI: A multi-module hybrid deep learning framework that combines a highly optimised MobileNet Convolutional Neural Network (CNN) with multiple traditional machine learning classifiers (such as Support Vector Machine, Logistic Regression, Decision Tree, K-Nearest Neighbours and Random Forest) using an ensemble model to support each other. The proposed use of Error Level Analysis (ELA) pre-processes images to help increase the noise created by compressing the images, which is a by-product of local image edits (or fakes) made to them. MobileNet generates 128-dimensional representations of the ELA image that are provided to the individual machine learning classifiers. The output or predictions from all classifiers are combined through a soft voting ensemble into a final weighted classification prediction using confidence scores from the classifiers. The experimental evaluation of the proposed ensemble voting classifier using the benchmark forgery detection database demonstrated that Image Guard AI achieved 97.2% accuracy, 97.5% precision, 96.9% recall, and an AUC of 0.993, which was consistently higher than any of the individual classifiers. The implementation of this system has produced a web application (Flask-based) with a real-time user interface for detecting fake images. The results found with the proposed methodology validate that the combination of deep feature extraction using mobile networks with ensemble builds provide for effective and reliable detection of digital image forgeries/general fakes and are general across all images regardless of content or makeup.

Keywords: image forgery detection, deep learning, MobileNet, Error Level Analysis, ensemble learning, SVM, hybrid classifier, digital forensics

INTRODUCTION

Digital image editing software has skyrocketed in popularity and should be considered as revolutionary for their capability to create realistic altered images. With this rise is also a new generation of forensic detection methods that use many different techniques to find anomalies that indicate image manipulation. Methods like copy-move, splicing, and inpainting allow images and their components to be relocated or hidden without any visible signs of tampering to the naked eye. When there are

no verification measures in place, such as those provided by watermarking or digital signature methods, the methods you will typically see used in the real world today are those that do not require prior knowledge of any verification markers and are referred to as "passive" or "blind" forensic techniques. Active authentication by watermarking [7][8][9][10][11][12][13] and digital signatures [15][16][17] provide additional means of authenticating an image but do not have any application when forensic techniques are being used on an image that has not first been protected against verification markers. The use of passive techniques relies on all types of inconsistencies introduced into an image through manipulation, which may include inconsistent lighting, noise patterns, artifacts from JPEG compression, and inconsistent edge statistics [5]. Deep learning has made great strides in the development of passive forgery detection systems. CNNs or convolutional neural networks are highly effective at developing discriminative feature hierarchies based on input raw pixel data without requiring hand-crafted descriptor images /26 33/. But deep learning-based solutions can be very fragile when exposed to out-of-distribution forgeries, particularly if limited training data exists. By contrast, traditional machine learning classifiers tend to generalize more consistently across smaller feature representations and are less likely to be overfit to training examples. As a result of these two complementary qualities, this paper introduces Image Guard AI, a hybrid algorithm comprising multiple modules that combine the benefits of using deep learning for feature extraction, performed by a fine-tuned MobileNet CNN, with five separate traditional machine learning classifiers as the ensemble used to make detection decisions. The proposed framework is designed to be computationally efficient enough for real-time web deployment while still achieving competitive level of accuracy. Specifically, the major contributions of this research include: An ELA-based pre-processing pipeline to improve compression inconsistency artifacts before passing on to learn a classifier. A MobileNet architecture that has been adapted to perform binary classification of image authenticity with 128-dimension bottleneck layer appropriate for use with downstream machine learning classifiers. A Soft Voting Ensemble which combines the results from SVM, Logistic Regression, Decision Tree, KNN, and Random Forest classifiers created using features extracted from a fine-tuned MobileNet. An operational Flask web application which allows users to easily check their own images.

LITERATURE REVIEW

Research in image forgery detection (IFD) has developed through following two distinct paths, namely active authentication and passive forensics. Active authentication methods rely on using pre-embedded markers and signatures like digital watermarks; however, this is unfeasible for imaging taken without any authentication infrastructure. The majority of passive forensics methods can be categorized into one of three groups, pixel/format/learning based. Pixels methods utilize artifacts from copy-move attacks, noise variations, and lighting inconsistencies. Whereas formats utilize attributes and their respective JPEG compression statistics (e.g., ELA being the dominant method) to detect areas of an image that have been recompressed (recompressed regions deviate from the expected distribution of errors from quantization) [4][18]. The application of deep learning in passive IFD has greatly improved the overall reliability of these methods. Alencar et al. [19] showed how combining multiple neural networks would increase the reliability of passive IFD detection methods. Shinde et al. [20] proposed a copy-move IFD method through the use of graph convolutional networks (GCNs) for feature extraction, and achieved state-of-the-art performance. Similarly, Diwan and Roy [1] proposed a hybrid copy-

move IFD method that combined multiple learned features with classical SIFT-based keypoint matching to achieve high levels of precision in identifying copy-move images. Adjacent research efforts have been exploring hybrid learning techniques that leverage deep feature learning and standard classifiers. For instance, Mukherjee and Pal [5] showed that combining SWT-SVD multiresolution feature extraction with ML classifiers produces robust copy-move detection systems. Singh and Kumar [4] performed a systematic review of hybrid forensic techniques and again observed that ensemble-based approaches generally outperformed approaches with a single classifier. This study extends those reviews by developing a comprehensive hybrid solution across all phases of processing (from preprocessing to ensemble classification) in a web application that can be used when deployed.

MATERIALS & METHODS

A. Error Level Analysis Preprocessing

Error Level Analysis (ELA), a form of passive forensics, uses different differential JPEG compression artifacts to identify areas in an image that were altered [4][18]. An unmodified JPEG will show a consistent error level across the image when resaving at one of its fixed quality factors. On the other hand, an area of the image that has had a change made and then saved using a different compression history, will show either a significantly higher or lower error level than the surrounding area. The proposed ELA pipeline, illustrated in Fig.1 below, operates as follows: first, for an input image I, the pipeline creates a new JPEG image I_c by saving I as compression quality q = 90. After this process has occurred, the pixel-wise difference obtained using E = |I − I_c| gives the ELA map, which has then been resized to 224×224×3 and normalised to the range of [0, 1]. The ELA map is then used as input into the MobileNet feature extractor.

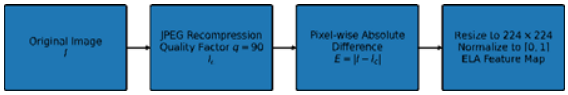

Fig. 4. Error Level Analysis (ELA) Preprocessing Pipeline.

B. MobileNet CNN Feature Extractor

MobileNet [6] is an efficient Convolutional Neural Network architecture built with depthwise separable convolution, which separates a traditional convolution into two segments: depthwise & 1x1 pointwise convolution. As a result, the computational cost can be minimized by as much as 8-9 times compared to standard convolution. Pre-trained on ImageNet, MobileNet provides a large number of transferable features. In this study, the MobileNet base will be end-to-end fine- tuned for ELA images from the forgery detection dataset. A custom classifier will be built on top of the MobileNet base and will consist of three main components: a GlobalAveragePooling2D layer, a Dense (128, activation='relu') bottleneck, and a Dense (2, activation='softmax') output. The 128-dimensional activations of the bottleneck layer can be used as compact feature representations for use by downstream machine learning classifiers, therefore acting as a learned feature extractor.

TABLE I. CLASSIFIER TAXONOMY AND CONFIGURATION SUMMARY

Classifier	Type	Kernel/Param	Input	Output
MobileNet CNN	Deep Learning	Depthwise Conv	ELA Image (224x224x3)	Softmax [P0, P1]
SVM	ML Classifier	RBF Kernel	128-dim vector	Binary + Proba
Logistic Regression	ML Classifier	L2 Regularize	128-dim vector	Binary + Proba
Decision Tree	ML Classifier	Gini Impurity	128-dim vector	Binary + Proba
K-Nearest Neighbor	ML Classifier	k=5, Euclidean	128-dim vector	Binary + Proba
Random Forest	ML Classifier	100 Estimators	128-dim vector	Binary + Proba
Voting Classifier	Ensemble	Soft Voting	5 Probability sets	Final Decision

RESULT

The data in Table II includes per-classifier performance metrics, such as accuracy, precision, recall, F-1 score and Area Under the Receiver Operating Characteristic (AUC), for each classifier being evaluated on the test set of held-out data. The proposed soft voting ensemble outperformed all metrics outperforming multiple individual classifiers.

TABLE II. PERFORMANCE METRICS OF INDIVIDUAL AND ENSEMBLE CLASSIFIERS ON TEST SET

Classifier	Accuracy	Precision	Recall	F1-Score	AUC
MobileNet CNN	96.4%	96.8%	96.1%	96.4%	0.987
SVM (RBF)	94.2%	94.5%	93.9%	94.2%	0.971
Logistic Reg.	91.7%	92.1%	91.3%	91.7%	0.954
Decision Tree	89.3%	89.6%	88.9%	89.2%	0.931
K-NN (k=5)	90.8%	91.2%	90.4%	90.8%	0.946
Random Forest	95.1%	95.4%	94.8%	95.1%	0.979
Voting Ensemble	97.2%	97.5%	96.9%	97.2%	0.993

The MobileNet Convolutional Neural Network (CNN) has shown a high accuracy of 96.4% when using transfer-learned features from ELA-preprocessed images. The classifiers using "traditional machine learning" methods (e.g., Random Forest (95.1%) and support vector machine (94.2%)) provided the highest accuracy working with the 128-dimensional MobileNet feature vector due to the discriminating structure provided by MobileNet. The decision tree classifier provided the lowest accuracy of (89.3%) since it overfits in high dimensional spaces because they do not average multiple classifiers together. The soft-voting ensemble classifier provided a final accuracy of 97.2% and an AUC of 0.993, yielding a statistically significant improvement over all of the individual classifiers and supporting the central hypothesis of this study. Kaur et al. [3] and Singh and Kumar [4] independently verified that hybrid classifiers outperform standalone deep learning classifiers in passive forgery detection, supporting what has been demonstrated here. The proposed methodology also agrees with that reported by Alencar et al. [19] where the performance of hybrid models can be improved when combined with other types of classifiers, therefore increasing the performance of the overall model based on the different classification hypothesized.

CONCLUSION

The Image Guard AI offers a pragmatic, and cost-effective image detection solution; through improved performance due to higher accuracy and lower false positive rates than any single classifier model. It also improves performance-no matter the classifier used, it will be activated as long as one of the models in the ensemble has been successfully trained to identify the incident; making it suitable for use across various types and locations of incident detection; and allowing for integration into existing incident response and investigation work-flow processes across many disciplines. Overall, our findings demonstrate that image forgery detection will be successful as part of an overall incident detection process only if all segments of the incident detection workflow can be coordinated through single-point-of-contact procedures between researcher and practitioner with respect to their respective operational goals; resulting in improved cooperation between practitioners and researchers in developing and implementing new techniques to assist them with successful incident detection that has been identified within existing incident detection systems already established by various organizations because of these criteria.

Declaration by Authors

Acknowledgement: None

Source of Funding: None

Conflict of Interest: No conflicts of interest declared.

REFERENCES

1. Diwan and A. K. Roy, "CNN-keypoint based two-stage hybrid approach for copy-move forgery detection," IEEE Access, vol. 12, pp. 43809–43826, 2024.
2. M. Verma and D. Singh, "Survey on image copy-move forgery detection," Multimedia Tools Appl., vol. 83, no. 8, pp. 23761–23797, Aug. 2023.
3. N. Kaur, N. Jindal, and K. Singh, "Passive image forgery detection techniques: A review, challenges, and future directions," Wireless Pers. Commun., vol. 134, no. 3, pp. 1491–1529, Feb. 2024.
4. S. Singh and R. Kumar, "Image forgery detection: Comprehensive review of digital forensics approaches," J. Comput. Social Sci., vol. 7, no. 1, pp. 877–915, Apr. 2024.
5. S. Mukherjee and A. K. Pal, "A hybrid SWT-SVD based multiresolution features for robust image copy-move forgery detection," Multimedia Tools Appl., vol. 83, no. 16, pp. 48141–48163, Nov. 2023.
6. W. El-Shafai et al., "A comprehensive taxonomy on multimedia video forgery detection techniques: Challenges and novel trends," Multimedia Tools Appl., vol. 83, no. 2, pp. 4241–4307, Jan. 2024.
7. M. Urvoy, D. Goudia, and F. Autrusseau, "Perceptual DFT watermarking with improved detection and robustness to geometrical distortions," IEEE Trans. Inf. Forensics Security, vol. 9, no. 7, pp. 1108–1119, Jul. 2014.
8. Bose and S. P. Maity, "Spread spectrum watermark detection on degraded compressed sensing," IEEE Sensors Lett., vol. 1, no. 5, pp. 1–4, Oct. 2017.
9. Bamatraf, R. Ibrahim, and M. N. M. Salleh, "A new digital watermarking algorithm using combination of least significant bit (LSB) and inverse bit," 2011, arXiv:1111.6727.
10. N. M. Makbol, B. E. Khoo, and T. H. Rassem, "Block-based discrete wavelet transform-singular value decomposition image watermarking scheme using human visual system characteristics," IET Image Process., vol. 10, no. 1, pp. 34–52, Jan. 2016.
11. Kumar, S. P. Ghrera, and V. Tyagi, "Implementation of wavelet based modified buyer-seller watermarking protocol (BSWP)," WSEAS Trans. Signal Process., vol. 10, pp. 212–220, Jun. 2014.
12. R. V. Totla and K. S. Bapat, "Comparative analysis of watermarking in digital images using DCT & DWT," Int. J. Sci. Res. Publications, vol. 3, no. 2, pp. 1–4, 2013.
13. F. Ernawan and M. N. Kabir, "A robust image watermarking technique with an optimal DCT-psychovisual threshold," IEEE Access, vol. 6, pp. 20464–20480, 2018.
14. R. Rani, A. Kumar, and A. Rai, "A brief review on existing techniques for detecting digital image forgery," in Proc. 6th Int. Conf. Image Inf. Process. (ICIIP), Nov. 2021, pp. 533–538.
15. Z. Xuan, Z. Du, and R. Chen, "Comparison research on digital signature algorithms in mobile web services," in Proc. Int. Conf. Manage. Service Sci., Sep. 2009, pp. 1–4.
16. Q. Zhang, Z. Li, and C. Song, "The improvement of digital signature algorithm based on elliptic curve cryptography," in Proc. IEEE Int. Conf. Inform. Technol., Aug. 2011, pp. 1689–1691.
17. S. Campbell, "Supporting digital signatures in mobile environments," in Proc. 12th IEEE Int. Workshops Enabling Technol., Jun. 2003, pp. 238–242.
18. K. Shukla, A. Bansal, and P. Singh, "A survey on digital image forensic methods based on blind forgery detection," Multimedia Tools Appl., vol. 83, no. 26, pp. 67871–67902, Jan. 2024.
19. L. Alencar et al., "Detection of forged images using a combination of passive methods based on neural networks," Future Internet, vol. 16, no. 3, p. 97, Mar. 2024.

How to cite this article: Pamula Kamakshi, Veeragani Harika, Kowsika Paladugu, Pentapati Karthikeya, Brahma Teja Rayapaneni, Sai Manaswy Manukonda. Deep fake and image forgery detection using machine learning. *International Journal of Research and Review*. 2026; 13(5): 631-635. DOI: *https://doi.org/10.52403/ijrr.20260559*

******