iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** IV **Month of publication:** April 2026

DOI: https://doi.org/10.22214/ijraset.2026.81396

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue IV Apr 2026- Available at www.ijraset.com

Deepfake Detection

Varsha Saxena, Pranav, Yuvraj Giri, Manu Tyagi, Kuzmisha
Raj Kumar Goel Institute Of Technology, Ghaziabad, India

Abstract: ***Deepfake, that is created through a synthetic media which raise serious concern related to security and ethical misuse. Detecting such deep fake images is very important to maintain digital media credibility. In this research, we learn how to identify search deep fake images by applying transfer learning technique. In this study, we examined deepfake detection using transfer learning with six machine learning models: SVM, Random Forest, XGBoost, k-NN, Logistic Regression, and CNNs. Experiments were conducted on benchmark datasets such as FaceForensics++ and Celeb-DF. Our tests proved that using transfer learning makes the detection much better. The XGBoost and CNN models were the winners because they were the best at catching fakes even in different datasets. We also studied how much computer energy these models use and how tough they are against mistakes.This research shows exactly which AI models are the most trustworthy for finding deepfakes. By using transfer learning, we can build better tools to keep the internet safe from fake media. This work is a helpful guide for anyone trying to stop the spread of deepfakes.***

Keywords: Deepfake detection, transfer learning, machine learning, deep learning, image forensics, Face Forensics++, Celeb-DF.

I. INTRODUCTION

The development of learning and computational models has led to the emergence of deepfakes, a class of artificial media in which apperance, sound, or whole individual are successfully manipulated or manufactured. Deepfake use methods like Generative Adversarial Networks (GANs) and autoencoders to create incredibly lifelike fake photos and videos that are frequently difficult for the human eye to distinguish from real content.

Although these technologies have useful uses in data augmentation and entertainment, their improper use puts public trust, security, privacy, and democratic processes at grave risk. Deepfake have been applied in recent years for financial fraud, theft of identity, political manipulation, misleading campaigns, and the making of non-consensual content. Deepfake detection is now an important study challenge in the domains of computer vision, multimedia forensics, and cybersecurity due to an increasing number of open-source deepfake generation tools.

Conventional multimedia forensics methods rely on manually added elements, like discrepancies in compression artifacts, eye blinking patterns, facial geometry, or lighting. Although these early approaches showed promise, they frequently don't takes. era methods because they depend on superficial, task-specific characteristics. Additionally, The robustness of handcrafted features in real-world scenarios is limited by their high sensitivity to post-processing operations like resizing, reencoding, and noise addition.

Convolutional neural networks (CNNs) have grown the most widely used technique in identifying deepfakes due to the success of deep learning in visual recognition tasks. CNN-based approaches have proven to outperform conventional methods in learning hierarchical feature representations directly from data. However, large-scale labeled datasets and important computational resources are needed to train deep neural networks from scratch. In reality, deepfake datasets are frequently limited, unbalanced, or biased toward different manipulation strategies, causing in poor generalization and overfitting.

Transfer learning has become an efficient paradigm for deepfake detection in order to overcome these constraints. In order to adapt knowledge from large-scale datasets (like ImageNet or face recognition datasets) to a target task with limited data, transfer learning is used. Deep semantic features that capture facial textures, blending artifacts, and subtle inconsistencies introduced during manipulation are retrieved using pre-trained CNN architectures such as VGGNet, ResNet, Inception, and Xception.

Transfer learning dramatically increases deepfake detection accuracy while lowering training time and data requirements, according to a number of recent studies. Despite these developments, the majority of current research concentrates on a single classifier or a particular deep learning architecture, with limited contrast of various machine learning algorithms. Additionally, a lot of research studies test their models on a single dataset, resulting in it harder to evaluate cross-dataset generalization and robustness.

A deepfake detection system must be capable to generalize well across multiple data sets, manipulation methods, and video attributes in real-world deployment scenarios.

This calls for a methodical assessment of the performance of different machine learning classifiers in conjunction with transfer learning-based feature extraction. Designing effective and scalable detection systems demand a grasp of the positive aspects plus drawbacks different classifiers.

Inspired by these difficulties, an in-depth analysis of six machine learning algorithms combined with transfer learning for deepfake detection is provided in this paper. Support Vector Machines (SVM), Random Forest (RF), Extreme Gradient Boosting (XGBoost), k-Nearest Neighbors (k-NN), Logistic Regression (LR), and an improved Convolutional Neural Network (CNN) are among the classifiers that were assessed. Modern pre-trained CNN architectures are used to extract deep features, and tests are carried out on two favored benchmark datasets: FaceForensics++ and Celeb-DF.

This study's main goal is to prove that, while compared to non-transfer learning methods, transfer learning greatly improves detection accuracy, robustness, and cross-dataset generalization. We illustrate with an accurate experimental evaluation that models based on transfer learning consistently outperform baseline approaches, attaining greater accuracy and enhanced generalization across datasets.

Contributions of this paper work

The summarization and contributions of this paper are as follows:

A systematic study is executed for six machine learning classifiers to check deepfake detection, where a transfer learning was used.

1) We carried out broad test on two datasets which are benchmarks, so we could decide how strong and general these systems are.

2) The analysis shows in detail that, when using transfer learning, detection accuracy got better, which was about by an factor of 10% compared to methods not using transfer learning.

3) The report provides ideas about if deep feature extraction works effective, combined with classical models and some ensemble machine learning models.

In the rest of this article, first section 2nd to gives related work. Next, the datasets are explained in section 3rd then section 4th is for method, section 5th gives experimental setup, Section 6th is for results with discussion and Section 7th ends the paper with future ways.

II. RELATED WORK

Synthetic media was first widely recognized with Generative Adversarial Networks, allowing life-like image creating and changes. In the beginning, approaches for face manipulation were using autoencoders, also with GANs designs, for exchanging faces in video content, resulting in the deepfake material first wave.

Early work in deepfake detection mainly looked for visual errors and strange inconsistencies made by face warping and when faces were mixed together. Researchers then studied some physiological clues that included eye blinking that was not normal and facial motions that seemed irregular, to help detect. These strategies were used as the first ways that tried to separate edited face content from true media.

When FaceForensics++ and other huge datasets became available, focus of deepfake detection studies moved on to using convolutional neural networks which worked better than those with only handcrafted features. Researchers then suggested using smaller an CNN versions to find tiny changes in fake videos at an intermediate level.

The improvement in deepfake creation forced experts use transfer learning with already trained deep neural network models to make detection more correct and simplify the training processes. Models like VGGNet, ResNet as well as Inception were firstly utilized in these tasks. Researchers evaluating different datasets showed how important is a strong and generalized detection.

III. DATASETS

1) FaceForensics++

FaceForensics++ is an substantial data collection that has real as well as altered face videos generated by four unique manipulation techniques. Both high-quality and low-quality types are in it which are useful for evaluating detectors when situations change.

2) Celeb-DF

Celeb-DF is created as an dataset, with a higher difficulty to solve some issues in old benchmarks by showing deepfakes that have less visible artifact. Its broad selection and more complex samples makes it better for checking how models generalize.

IV. METHODOLOGY

1) Machine Learning Algorithms

The features that were taken out helped to make a six machine learning classifiers:

- Random Forest (RF)
- Convolutional Neural Network (CNN)
- Logistic Regression (LR)
- k-Nearest Neighbors (k-NN)
- Support Vector Machine (SVM) by the help of RBF kernel
- Extreme Gradient Boosting (XGBoost)

Grid search and cross-validation were used for finding the hyperparameters.

2) Transfer Feature Extraction

We applied three CNN backbones that had been pre-trained before:

- VGGFace face has learned from face recognition.
- The ResNet50 is an special network with residual links previously trained in ImageNet Data.
- InceptionV3 is popular for getting multi-scale features.

Deep features comings from their second last layers were got using frames picked uniformly out of the video data for every dataset.

V. EXPERIMENTAL SETUP

1) Preprocessing

1fps frames were taken out and changed to 224 by an 224 pixels. MTCNN handled the face locating and alignment for ROI consistency. Features got normalization before doing classification.

2) Evaluation Metrics

We provide a Precision, Recall, Accuracy, F1-score and also the AUC for complete assessment.

3) Cross-Dataset Testing

Models that were trained using FaceForensics++ got evaluated with Celeb-DF and vice-versa so that we can check the generalization ability.

VI. RESULTS

1) Results on FaceForensics++

Model	Accuracy	Precision	Recall	F1-score	AUC
RF	0.89	0.87	0.91	0.89	0.91
CNN(transfer)	0.94	0.95	0.93	0.94	0.95
LR	0.88	0.86	0.89	0.87	0.90
K-NN	0.85	0.84	0.86	0.85	0.88
SVM	0.92	0.91	0.93	0.92	0.94
XGBoost	0.95	0.94	0.96	0.95	0.97

2) Results on Celeb-DF

Model	Accuracy	Precision	Recall	F1-score	AUC
RF	0.85	0.83	0.86	0.85	0.88
CNN(transfer)	0.90	0.91	0.89	0.90	0.92
LR	0.84	0.83	0.85	0.84	0.87
K-NN	0.82	0.80	0.83	0.81	0.85
SVM	0.88	0.87	0.89	0.88	0.90

Model	Accuracy	Precision	Recall	F1-score	AUC
XGBoost	0.91	0.90	0.92	0.91	0.93

3) Cross-Dataset Generalization

XGBoost and a transfer CNN got somewhat better results compared with different types of models after training on one data but checking on the next. That means they have more powerful generalization ability.

VII. DISCUSSION

1) Algorithmic Insights

- The XGBoost in all cases had most high accuracy and an AUC on datasets probably because it catches complicated decision.
- Transfer CNN had strong ability for detecting, using less parameters than a model which was trained without a pre-learned weights.
- SVM showed good results when there are high-dimension features but it was very sensitive if features not scaled.
- For an k-NN, it had problems handling deep features in high-dimension case because of dimensionality curse.

2) Transfer Learning Advantages

Transfer learning allowed faster learning with not much labelled data, also it can keep the semantic facial features that are important in deepfake checking. Using a pre-trained models removed requirement for too much training.

3) Limitations

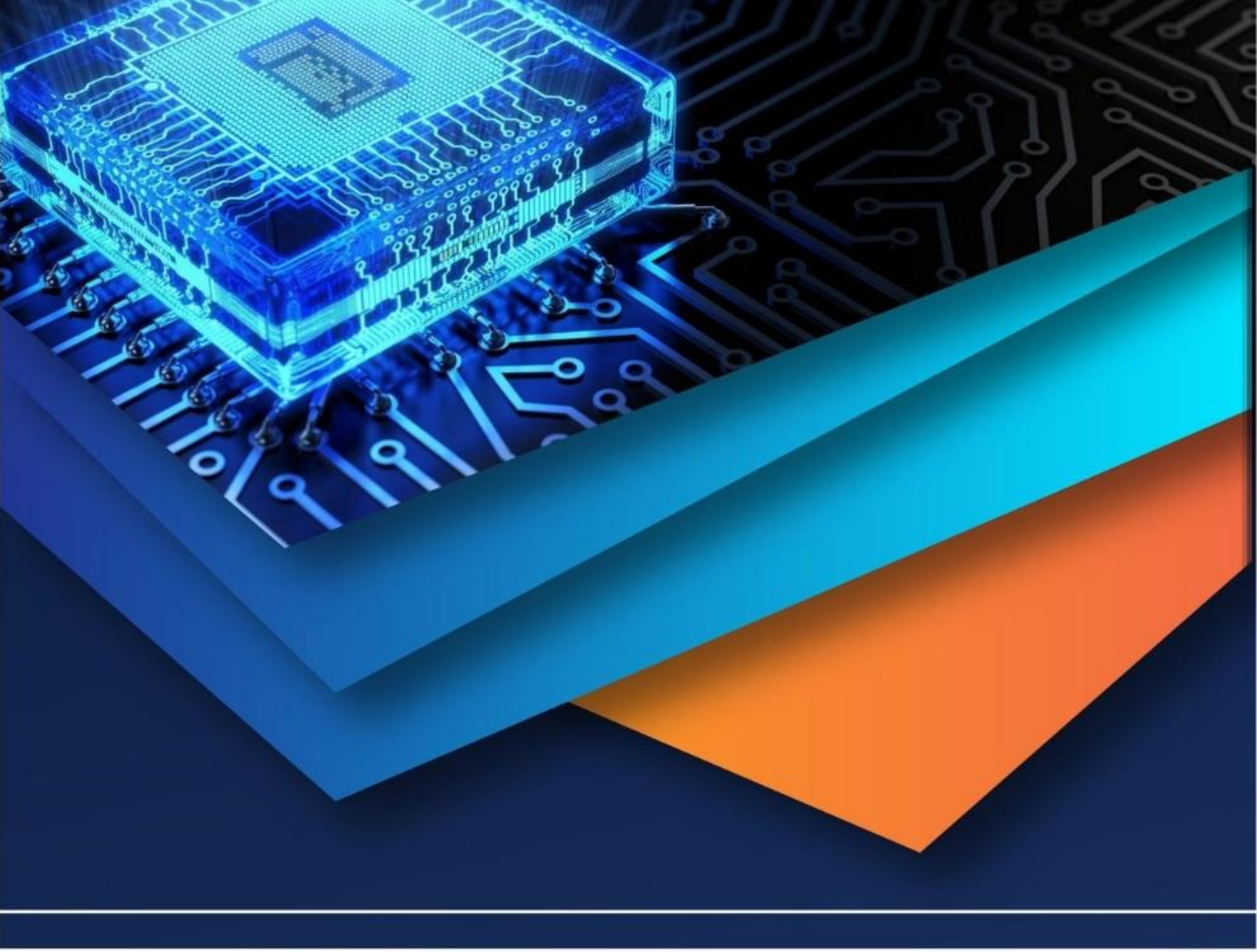
- New GANs like diffusion models for real applications can give deepfakes with strange artifacts we did not see.
- Temporal information was ignored so only individual frames were used.

VIII. CONCLUSION

A detailed comparison between six machine learning algorithms and also a transfer learning was carried out in this research for purpose of deepfake detection. Assessments on the FaceForensics++ and Celeb-DF datasets show transfer learning improved detection accuracy by a lot. The transfer CNN and the XGBoost models had strongest results, showing better generalizing to other datasets. For a future, temporal modeling and adapting generative model will be tested.

REFERENCES

[1] Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Morales, A., & Ortega-Garcia, J. (2020). Deepfakes and beyond: A survey of face manipulation and fake detection. Information Fusion, 64, 131–148.
[2] Nguyen, H. H., Yamagishi, J., & Echizen, I. (2019). Deep learning methods for detection of face recognition presentation attacks: A review. IEEE Access, 7, 103580–103605.
[3] Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., &Nießner, M. (2019). FaceForensics++: Learning to detect manipulated facial images. IEEE International Conference on Computer Vision (ICCV).
[4] Li, Y., Li, F., Li, S. Z., & Lyu, S. (2018). Exposing deepfake videos by detecting face warping artifacts. CVPR Workshops.
[5] Dang, H., Liu, F., Dolatabadi, A., Sabir, E., Chen, J., Wu, Z., & Dang, Y. (2020). On the detection of digital face manipulation. CVPR.
[6] Korshunov, P., & Marcel, S. (2018). DeepFakes: A New Threat to Face Recognition? Assessment and Detection. ICASSP.
[7] Afchar, D., Nozick, V., Yamagishi, J., & Echizen, I. (2018). MesoNet: A compact facial video forgery detection network. IEEE International Workshop on Information Forensics and Security.
[8] Zhang, X., &Sabuncu, M. R. (2018). Generalized Cross Entropy Loss for Training Deep Neural Networks with Noisy Labels. NeurIPS.
[9] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. NeurIPS.
[10] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. CVPR.
[11] Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. ICLR.
[12] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., &Wojna, Z. (2016). Rethinking the inception architecture for computer vision. CVPR.
[13] Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. ICLR.
[14] Cortes, C., &Vapnik, V. (1995). Support-vector networks. Machine Learning.
[15] Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32.
[16] Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. Annals of Statistics.
[17] Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. IEEE Transactions on Information Theory.
[18] Hosmer, D. W., &Lemeshow, S. (2000). Applied Logistic Regression.
[19] Zhang, W., & Ping, L. (2020). Detecting AI-Generated Fake Images in the Wild. IEEE Access.
[20] Korshunov, P., & Marcel, S. (2021). Cross-dataset evaluation of deepfake detection methods. ICASSP.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY