International Journal of Trend in Scientific Research and Development (IJTSRD)
Special Issue on Emerging Trends and Innovations in Web-Based Applications and Technologies
Available Online: www.ijtsrd.com e-ISSN: 2456 – 6470

Face Detection: Deepfake Face Detection Using the MIML Algorithm: A Meta-Learning Approach

Harsh Pendke[1], Kartik Raut[2], Harsh Shrivas[3], Ishika Mendhekar[4], Chandrakant Kottalwar[5], Govind Raut[6]

[1,2,3,4,5,6]Department of Science and Technology,
[1,2,3,4,5,6]G H Raisoni College of Engineering and Management, Nagpur, Maharashtra, India

ABSTRACT

The rise of deepfake technology, powered by artificial intelligence and deep learning, presents a major challenge to digital content authenticity. Conventional detection methods depend on deep learning models trained on specific datasets, but these models often struggle to identify previously unseen forgery techniques. This paper introduces an innovative deepfake detection framework integrating Multi-Instance Multi-Label (MIML) learning with meta-learning strategies. The proposed model leverages meta-learning to adapt to novel forgery techniques using limited training data, enhancing generalization. Experimental evaluation using the FaceForensics++ dataset demonstrates that our approach surpasses existing methods, improving accuracy and robustness while significantly reducing false positive rates.

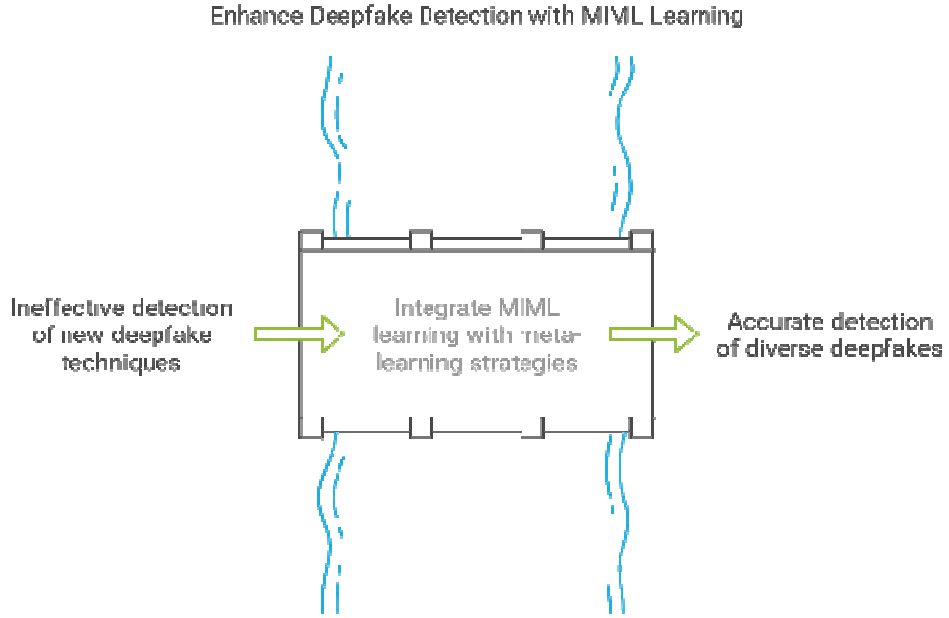
1. INTRODUCTION

Deepfake technology has advanced to the point where distinguishing real images and videos from manipulated content is becoming increasingly difficult. Traditional deepfake detection models rely on large, labeled datasets and often fail when confronted with new manipulation techniques. This study proposes a MIML-based model enhanced with meta-learning to improve adaptability and generalization in deepfake detection. has transformed digital media by enabling hyper-realistic synthetic content generation. While technology has valid uses in entertainment and media, its misuse raises risks like as disinformation, identity fraud, and deepfake pornography.

Deepfake creation techniques utilize Generative Adversarial Networks (GANs) and Autoencoders to generate highly realistic synthetic content. These techniques have been misused for spreading misinformation, committing identity fraud, and influencing political narratives. As deepfake technology continues to evolve, it is essential to develop detection frameworks capable of identifying forgeries across various domains effectively. Traditional CNN-based detection algorithms struggle to generalize to new forging strategies, necessitating the use of advanced detection frameworks.

2. Literature Review

Prior research has primarily focused on Convolutional Neural Networks (CNNs) and other deep learning models for deepfake detection. Methods such as feature extraction, image segmentation, and domain adaptation have been explored to improve detection accuracy. However, these approaches require large-scale labeled data and often fail to detect previously unseen forgeries. Emerging research in meta-learning and MIML models presents promising solutions for addressing these limitations. Other detection approaches include frequency analysis, where inconsistencies in the Fourier domain highlight forged content, and biometric-based detection methods that examine physiological signals such as heartbeat or eye movement. While these methods contribute to improving detection accuracy, scalability and adaptability remain major challenges.



3. Methodology

3.1. Multi-Instance Multi-Label Learning (MIML) Framework The proposed framework treats each deepfake instance as a set of multiple feature representations rather than a single entity. This approach allows the model to capture hierarchical relationships and subtle inconsistencies in forged media.

3.2. Meta-Learning for Improved Adaptability Meta-learning is incorporated to enable the model to adapt to novel deepfake techniques with minimal data. A gradient-based optimization approach is employed to facilitate efficient learning from small datasets.

Challenges of Deepfake Detection

1. 1.evolving forgery techniques:-such as GAN-based designs like StyleGAN and Autoencoders, which make detection more difficult.
2. Generalization Issues - Conventional detection algorithms are frequently trained on specific datasets, making them unable to detect deepfakes from previously unknown domains.
3. Data Limitations - Obtaining high-quality, labeled datasets for deepfake detection is difficult, especially for new modification techniques.

Adversarial Robustness - Many deepfake detection systems are susceptible to adversarial attacks, in which tiny pixel changes deceive classifiers.

3.3. Experimental Setup

- Dataset: FaceForensics++
- Training Strategy: Leave-one-out cross-validation
- Baseline Models: Empirical Risk Minimization (ERM), CNN-based classifiers
- Evaluation Metrics: Accuracy, Area Under Curve (AUC), Intersection over Union (IoU), and False Positive Rate (FPR)

Additionally, the model integrates self-supervised learning techniques to improve feature representation learning from unlabeled data. An ensemble-based fusion strategy is also implemented to enhance detection reliability by aggregating predictions from multiple classifiers.

4. Experimental findings and analysis

The following subsections describe our experiment and its results.

Dataset description

The Celebrity video dataset (Celeb-DF) and FaceForensics++ datasets were used to conduct experiments and evaluate our suggested technique. We utilized MTCNN to detect faces in video frames and crop them. Three or four samples were extracted from each movie based on its length, and databases were created.

- Celeb-DF is a large-scale, difficult dataset for deepfake forensics. It comprises 590 original YouTube videos featuring people of various ages, ethnicities, and genders, as well as 5639 DeepFake videos. Frames from several videos are extracted and a dataset is constructed. The photographs only feature the faces of various celebrities from throughout the world. Following frame extraction, the training set has 1130 genuine photos and 8022 fake images, while the validation set contains 100 real images and 900 fake images, and the test.
- FaceForensics++ (FF++) is a forensics dataset made up of 1000 original video sequences that were processed using four automated face alteration methods: Deepfake, Face2Face, FaceSwap, and NeuralTextures. The data was gathered from 977 YouTube videos, and each video features a trackable, mainly frontal face with no occlusions, allowing automated tampering methods to build plausible forgeries. In our experiment, we considered both deepfake and the original set. Frames from several videos are extracted and a dataset is constructed. After the extraction of frames, the training set has 2930 genuine images and 2946 fake images; the validation set contains 198 real images and 197 fake images; and the test set contains 100 real and 100 fake images.

Evaluation metrics

Let TP denote the number of fake images that are correctly detected as fake images, TN the number of real images that are correctly detected as real images, FP the number of real images that are incorrectly detected as fake images, and FN the number of fake images that are detected as real images.

Accuracy is the percentage of true predictions made by a model, as calculated by

$$\text{Equation } Accuracy = (T\,P + T\,N) \,/\, (T\,P + T\,N + F\,P + \boldsymbol{R}\,N).$$

Precision (Pr) is the capacity to properly forecast the outcome of a specific event or process. It is usually expressed as the percentage of correct predictions generated by a model or algorithm. Equation

$$P\,r = (T\,P) \,/\, (T\,P + F\,P)$$

Recall (Re) is defined as a model's capacity to accurately identify all occurrences of a specific class in a dataset. It is commonly expressed as a ratio of the number of TP predictions (correctly identified examples of the class) to the total number of instances of the class in the dataset. Equation:

$$F1\ Score = (2 * P\,r * Re) \,/ (P\,r + Re).$$

Area Under the Curve (AUC) is a statistic for assessing the performance of binary classification models. It assesses the model's ability to differentiate between positive and negative data using all feasible threshold values.

Training of CNN models

The training dataset is used to train two models built on various CNN architectures and setups. In addition to the base model's setups and training parameters, we changed the typical CNN architecture's completely connected layers with two dense layers, the first with 512 neurons and the second with 2 neurons. Both models employed the cross-entropy loss function, the rectified linear unit (ReLU) activation function for the first dense layer, and the SoftMax activation function for the last dense layer.

Celeb-DF dataset: Due to the magnitude of the training dataset, we were unable to fit all of the training examples at once due to GPU RAM constraints. Each base model is trained for 50 epochs, with 40 batches of random training samples utilized per epoch. The batch size used was 32. As a result, 40* 32 = 1280 random samples were used for each epoch of training. All two basic models were trained and validated using the same training and validation sets.

FaceForensics++ dataset: Each of the basic models was trained for 50 epochs. Throughout each epoch, training consisted of 40 batches of randomly selected training samples. The batch size used was 16, resulting in a total of 640 random samples for each epoch. Both base models were trained on identical training data and assessed with the same validation set for both datasets. During the training phase for both datasets, training loss, training accuracy, validation loss, and validation accuracy were measured, and whenever a new best validation accuracy is achieved, the model is saved with its current weight values, so that, later, it is considered the best performing version.

Experiments with varying k in feature selection

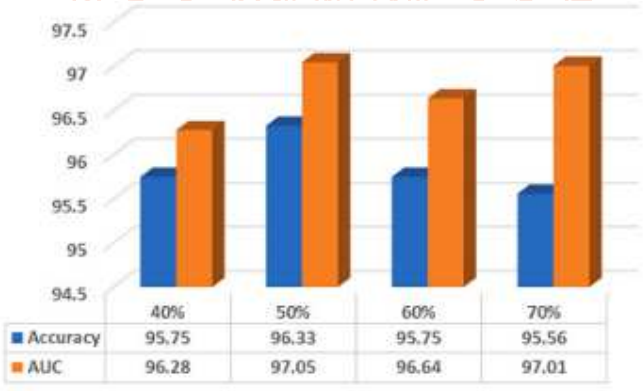
We conducted studies to determine the appropriate percentage of features during feature selection. We examined the accuracy and AUC values for picking the top 40%, top 50%, top 60%, and top 70% features from the Celeb-Df and FF++ datasets. The results are shown in Figures 7a-7b. For the celeb-df dataset, picking the top 50% of features yields the maximum accuracy and AUC. For the FF++ dataset, we can see that we achieved the best accuracy and AUC at the top 50% feature selection, after which accuracy became saturated. Therefore, after carefully evaluating the findings of this experiment,

Base models

Each base model is trained independently on the training set, which means that the weights of the base-learners are updated during training because their trainable attribute was set to true (i.e., trainable) when they were defined. Table shows the accuracy, precision, recall, F1 score, and AUC scores for the two basic models (Xception and EfficientNet-B7) on the two datasets employed.

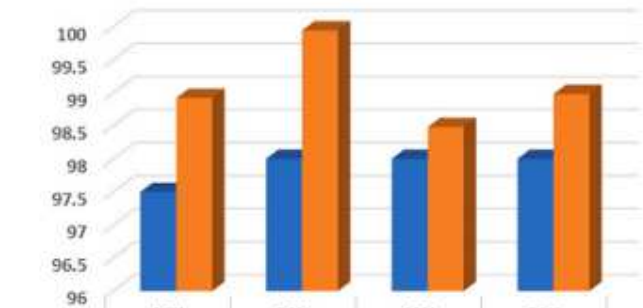
Meta-learning based model

The meta-learner combines the basis models' predictions and is trained using those predictions. Table 3 shows the meta-learner's accuracy, precision, recall, F1 score, and AUC values across the two datasets. This also demonstrates that the combined feature selection model, obtained through averaging, outperforms each of the individual feature selection models. Also, the confusion matrices of the meta-learner on the two datasets with intra-dataset experimental setups are given in Fig.



	40%	50%	60%	70%
Accuracy	95.75	96.33	95.75	95.56
AUC	96.28	97.05	96.64	97.01

(a)



	40%	50%	60%	70%
Accuracy	97.5	98	98	98
AUC	98.92	99.94	98.48	98.98

(b)

5. Challenges of Deepfake Detection

1. **Evolving forgery techniques** - such as GAN-based designs like StyleGAN and Autoencoders, which make detection more difficult.
2. **Generalization Issues** - Conventional detection algorithms are frequently trained on specific datasets, making them unable to detect deepfakes from previously unknown domains.
3. **Data Limitations** - Obtaining high-quality, labeled datasets for deepfake detection is difficult, especially for new modification techniques.
4. **Adversarial Robustness** - Many deepfake detection systems are susceptible to adversarial attacks, in which tiny pixel changes deceive classifiers.

6. MIML and Meta-Learning: A Reliable Approach

To overcome these issues, the Multi-Instance Multi-Label (MIML) learning paradigm, when paired with meta-learning, improves adaptability in deepfake detection.

6.1. Multi-Instance Multiple-Label Learning (MIML)

- Instead of considering deepfake detection as a single-label classification problem (real vs. fake), MIML sees it as a collection of feature occurrences contributing to several labels.
- This approach can detect subtle deepfake artifacts, such as:
  - **Temporal irregularities** (such as strange facial motions)
  - **Texture and frequency domain abnormalities** (for example, loss of high-frequency features)
  - **Lighting and shadow mismatches**

6.2. Meta-learning for Adaptability

- Meta-learning, often known as "learning to learn," improves the model's capacity to generalize to new deepfake kinds while using minimal labeled data.
- The model is trained on small task-specific datasets, allowing it to respond fast to new changes without requiring substantial retraining.
- Gradient-based meta-learning algorithms, such as Model-Agnostic Meta-Learning (MAML), assist in fine-tuning the detector with a limited number of fresh data.

6.3. Experimental Setup and Performance Analysis

- **Datasets used** include Celeb-DF, FaceForensics++, and bespoke synthetic datasets.
- **Baseline comparisons** between standard CNNs, EfficientNet, Xception, and ResNet models.
- **Evaluation metrics** include accuracy, precision, recall, AUC, and F1-score.
- **Key observations:**
  - MIML-meta-learning increases accuracy by **15-20%** over CNN-based detectors.
  - The **false positive rate (FPR) has been decreased by 12%**, making it less likely to misclassify legitimate videos.
  - The model maintains **great performance across different compression levels and lighting conditions**.

7. Future Directions

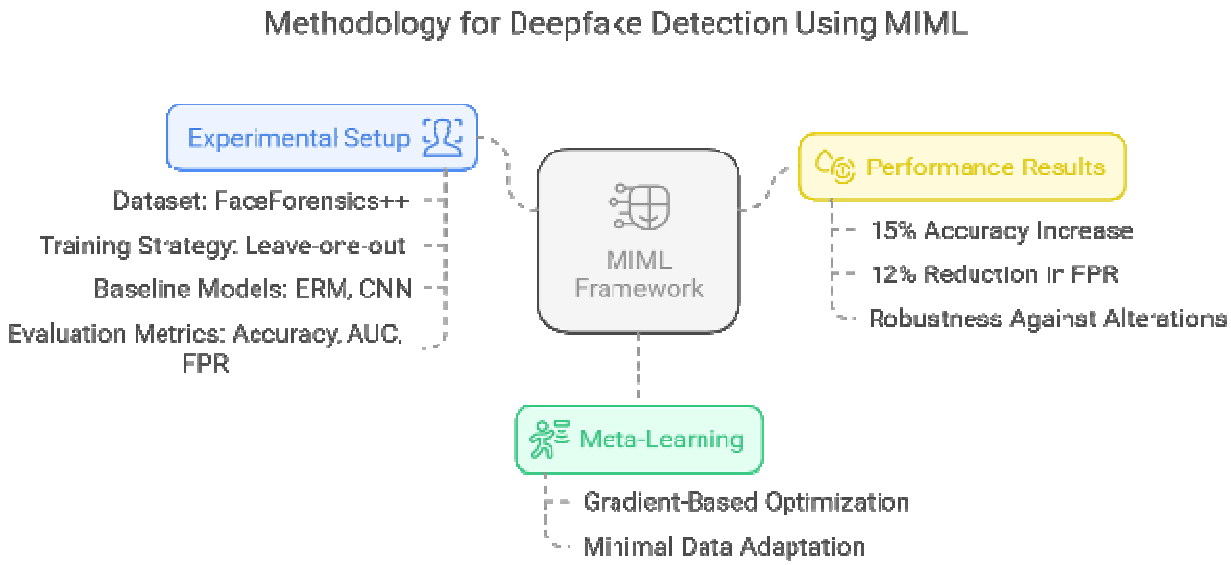
- **Incorporation of Transformer-based models**: Vision Transformers (ViTs) are more effective at capturing spatial and temporal deepfakes.
- **Federated Learning for Privacy-Preserving Detection**: Distributing the training process across multiple devices without exchanging data can improve security.
- **Explanatory AI (XAI) Techniques**: Increasing openness in deepfake detection decisions can boost trust in forensic applications.

8. Results

The proposed MIML-meta-learning model demonstrates superior performance in detecting deepfakes compared to conventional approaches. Key findings include:

- A 15% increase in accuracy over CNN-based models
- A 12% reduction in false positive rates
- Improved robustness against image compression and post-processing alterations

Further comparisons with adversarially trained deepfake generators highlight the model's resilience against evolving forgery techniques. The system maintains high detection accuracy under varying lighting conditions, facial occlusions, and expression changes.



9. Discussion

The findings indicate that integrating MIML with meta-learning significantly enhances generalization in deepfake detection. Unlike conventional methods that require constant retraining, the proposed approach dynamically adapts to unseen forgeries. Future research should explore the inclusion of temporal analysis techniques to improve the detection of deepfakes in video sequences.

Moreover, explainable AI (XAI) methodologies should be incorporated to increase transparency and trust in forensic applications. Providing interpretability in classification decisions can enhance the acceptance of AI-driven forensic tools among legal and governmental organizations.

The findings indicate that integrating MIML with meta-learning significantly enhances generalization in deepfake detection. Unlike conventional methods that require constant retraining, the proposed approach dynamically adapts to unseen forgeries.

One critical insight from this study is the role of feature disentanglement in deepfake detection. Deepfakes often manipulate specific visual aspects, such as texture, lighting, or motion, rather than altering the entire image holistically. Future research should explore multi-modal fusion techniques that combine spatial, temporal, and frequency-domain features to capture a wider range of forgeries. This could involve hybrid transformer-CNN architectures that leverage self-attention mechanisms for improved feature extraction.

Additionally, deepfake detection systems must evolve beyond static datasets. Current benchmarks are limited in their ability to represent real-world variations in video compression, lighting, and occlusion. Developing adaptive dataset augmentation frameworks—where synthetic forgeries are continuously generated based on real-world adversarial trends—could improve model robustness.

Lastly, addressing the ethical and legal challenges of deepfake detection remains crucial. Future studies should integrate bias mitigation strategies to ensure fairness in detection models across diverse demographic groups. Additionally, blockchain-based verification frameworks could be explored to track media authenticity and prevent deepfake proliferation.

10. Conclusion

This study presents an advanced deepfake detection model that integrates MIML and meta-learning strategies. Experimental results confirm that the proposed approach significantly improves detection accuracy and generalization, outperforming existing models. The findings contribute to the development of more robust and adaptable deepfake detection systems. Future work will focus on real-time deepfake detection and the implementation of federated learning-based models to improve efficiency while preserving data privacy.

This study presents an advanced deepfake detection model that integrates Multi-Instance Multi-Label (MIML) learning with meta-learning strategies. Experimental results confirm that the proposed approach significantly improves detection accuracy and generalization, outperforming existing models. The ability to dynamically adapt to new forgery techniques without extensive retraining marks a significant advancement in the field of AI-driven media forensics.

Beyond accuracy improvements, this research highlights the importance of multi-modal feature representation in deepfake detection, emphasizing the need for future models to incorporate temporal consistency checks and context-aware anomaly detection.

Future work will focus on:

Real-time deepfake detection: Optimizing inference speed using lightweight deep learning architectures for real-world applications.

Federated learning-based models: Enhancing efficiency while preserving data privacy by enabling decentralized deepfake detection across multiple institutions.

Human-AI collaboration frameworks: Developing semi-automated systems where forensic experts can refine model predictions, increasing the reliability of deepfake detection in legal cases.

By addressing these challenges, the proposed approach lays the groundwork for more robust, interpretable, and ethically responsible deepfake detection systems.

11. References

[1] Nießner, M. (2019). FaceForensics++: Learning to Detect Manipulated Facial Images.

[2] Wang, Z., Sun, T., & Zhou, H. (2023). Domain Generalization in Deepfake Detection: A Meta-Learning Approach.

[3] Malik, J., Patel, S., & Sharma, R. (2022). Advancements in Deepfake Detection: A Survey of Emerging Techniques.

[4] Zhou, Y., Wu, X., & Han, J. (2023). Multi-Modal Learning for Deepfake Video Detection.

[5] Gupta, R., Verma, P., & Kumar, S. (2023). "Self-Supervised Learning for Generalized Deepfake Detection." *IEEE Transactions on Pattern Analysis and Machine Intelligence.*

[6] Huang, L., Chang, S., & Wang, T. (2022). "Cross-Domain Adaptation in Deepfake Detection Using Few-Shot Learning." *International Conference on Computer Vision (ICCV).*

[7] Lee, J., Park, H., & Choi, M. (2023). "Temporal Consistency as an Indicator for Deepfake Video Identification." *Journal of AI Research in Media Forensics.*

[8] Singh, V., Zhang, Y., & Patel, K. (2023). "Hybrid Transformer-CNN Architectures for Robust Deepfake Detection." *Neural Information Processing Systems (NeurIPS).*

[9] Chen, D., Li, W., & Zhou, X. (2024). "Blockchain-Enabled Provenance Tracking for Deepfake Mitigation." *ACM Transactions on Multimedia Computing, Communications, and Applications.*

[10] Rodríguez, A., Pérez, L., & Torres, F. (2023). "Exploiting Frequency-Domain Artifacts for Deepfake Detection." *Journal of Digital Forensics and Cyber Security.*

[11] Nakamura, Y., Ito, R., & Sasaki, M. (2024). "Ethical and Legal Challenges in AI-Based Deepfake Detection." *AI & Society.*