WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

AI-Powered Lightweight Deepfake Filter

Bhavesh Jaware[a], Bhavesh Patil[b],Gaurav Shinde[c], Dr Varsha Patil[d]

[a]Department of Computer Engineering Matoshri College of Engineering & Research Centre Eklahare , Nashik-422105 , India
[b]Department of Computer Engineering Matoshri College of Engineering & Research Centre Eklahare , Nashik-422105 , India
[c]Department of Computer Engineering Matoshri College of Engineering & Research Centre Eklahare , Nashik-422105 , India
[d]Project Guide, Computer Engineering Matoshri College of Engineering & Research Centre Eklahare , Nashik-422105 , India

A B S T R A C T :

The increasing prevalence of deepfake media has sparked global concern, particularly regarding its misuse in misinformation, cybercrime, and social manipulation. While advanced AI models have demonstrated high detection accuracy, their computational complexity makes them impractical for deployment on resource-limited devices such as smartphones. This paper presents a novel approach—an AI-powered lightweight deepfake filter optimized through knowledge distillation. A compact student model learns from a larger, more complex teacher model to deliver near-equivalent detection performance with significantly lower computational demand. Experiments conducted on benchmark datasets demonstrate that the proposed lightweight model achieves high accuracy while being deployable in real-time on mobile platforms. This research aims to bring deepfake detection capability directly into users' hands, enabling instant content verification and promoting safer digital interaction.

Keywords: Deepfake Detection, Lightweight AI, Knowledge Distillation, Mobile Deployment, Real-Time Classification, Edge AI

1. Introduction

The digital age has enabled the rapid generation and distribution of multimedia content, but it has also opened doors to sophisticated manipulation tools—none more disruptive than deepfakes. Deepfakes refer to AI-generated synthetic media, where individuals' faces, voices, or actions are replaced with astonishing realism. These creations are powered primarily by Generative Adversarial Networks (GANs) and encoder-decoder architectures.

Although the technology behind deepfakes has legitimate applications in entertainment, accessibility, and education, its potential misuse for misinformation, political sabotage, and identity fraud presents a grave challenge. This has led to the development of numerous detection algorithms; however, most high-accuracy models are compute-intensive, limiting their deployment to cloud infrastructure or high-end systems.

The gap between powerful detection models and real-world usability forms the basis of this research. We propose a compact, energy-efficient, and mobile-ready deepfake filter using knowledge distillation. This technique allows a small model (student) to imitate the predictions of a large, accurate model (teacher), thereby enabling deepfake detection on devices with limited resources—such as smartphones or edge devices.

Our solution is focused on balancing detection accuracy and computational cost, opening a path for real-time, on-device media verification for the general public.

2. Literature Survey

The domain of deepfake detection has seen several breakthroughs over the past five years, especially with the advancement of deep learning techniques.

2.1 Conventional Approaches

Early deepfake detection methods focused on identifying inconsistencies in blinking, head posture, and facial warping. For instance, Matern et al. (2019) used handcrafted visual artifacts to detect facial inconsistencies, while Korshunov et al. (2018) evaluated frame-by-frame facial movement anomalies.

2.2 Deep Learning-Based Models

Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks have been widely adopted for detection. FaceForensics++ (Rossler et al., 2018) trained CNNs on manipulated videos, achieving significant detection performance. More recent approaches use transformer-based architectures for spatio-temporal pattern recognition, achieving even better results but at high computational cost.

2.3 Limitations in Mobile Deployment

Despite high accuracy, most existing models cannot run efficiently on mobile or edge hardware due to:

- High memory and storage requirements
- Long inference times
- Battery drain due to heavy processing

2.4 Knowledge Distillation in Lightweight AI

Knowledge distillation (KD) has recently emerged as an effective way to compress deep learning models. In KD, a student model is trained to replicate the softened output of a teacher model, capturing its learning behavior. Works such as Hinton et al. (2015) demonstrated that student models could achieve 90%+ performance of teacher models at a fraction of the size. This approach has been applied in image classification, NLP, and recently, media forensics.

2.5 Research Gap

While KD is well studied in other domains, its application in **mobile deepfake detection** remains largely unexplored. Our work fills this gap by developing a KD-based deepfake detection model suitable for mobile environments, with promising real-world applications.

3. Methodology

Our approach is built on a two-model training pipeline:

3.1 Teacher Model

A full-sized CNN (e.g., ResNet-50 or EfficientNet-B4) is trained on a labeled deepfake dataset to perform binary classification (real vs. fake). This model achieves high accuracy but is computationally heavy.

3.2 Student Model

A much smaller CNN (e.g., MobileNet or TinyResNet) is trained using:

- *Soft Targets*: The probabilistic outputs of the teacher model
- *Hard Targets*: The ground truth labels

The student minimizes a combination of:

- *Cross-entropy loss* (hard labels)
- *Kullback-Leibler divergence* (teacher predictions)

This setup allows the student model to inherit the performance capabilities of the teacher, while remaining lightweight.

4. Implementation

4.1 Dataset

We used a subset of the *FaceForensics++* dataset, containing manipulated and authentic videos, and extracted frames for training.

4.2 Tools & Frameworks

- *Framework*: TensorFlow Lite for mobile deployment
- *Models*: ResNet-50 (teacher), MobileNetV2 (student)
- *Training*: Google Colab with GPU acceleration
- *Evaluation*: Accuracy, model size (MB), inference time (ms)

4.3 Training Procedure

1. Train the teacher model to convergence (accuracy > 95%)
2. Freeze the teacher and train the student using knowledge distillation loss
3. Quantize the student model for mobile compatibility
4. Deploy on an Android simulator for testing

5. Results & Discussions

Metric	Teacher (ResNet-50)	Student (MobileNet)
Accuracy (%)	96.2	91.4
Model Size (MB)	98	13
Inference Time (ms/frame)	320	42
Energy Consumption	High	Low
Mobile Compatibility	No	Yes

The student model retained 91.4% accuracy of the teacher with 86% smaller size, making it suitable for real-time applications on mobile devices. It can detect and flag manipulated images in under half a second, enabling seamless integration into apps or browsers.

6. Conclusion & Future Work

TThis study demonstrates that *knowledge distillation* can be effectively used to create lightweight deepfake detection models suitable for *mobile deployment*. By training a student model on the outputs of a high-performing teacher, we achieve an optimal trade-off between performance and efficiency. Our student model enables real-time deepfake detection on smartphones, making media verification tools more accessible to end users.

Future Directions:

- Explore *multi-modal distillation* including audio and text signals
- Extend detection to *live video streams* and social media filters
- Implement *on-device training updates* for personalization
- Integrate with *browser plugins* for media verification at source

Our approach promotes ethical AI use by democratizing access to verification tools and empowering users to challenge digital misinformation effectively.

REFERENCES

1. Rossler, A. et al. (2018). FaceForensics++: Learning to Detect Fake Faces. *ICCV*.Zhao, J., Wu, X., & Li, Y. (2020). Multi-task Learning for Fake Face Detection with Consistency Constraints. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
2. Hinton, G. et al. (2015). Distilling the Knowledge in a Neural Network. arXiv preprint arXiv:1503.02531.
3. Matern, F., Thies, J. (2019). Detecting Deepfakes with Artifacts. SpringerLink.
4. Korshunov, P., Marcel, S. (2018). Deepfakes: A Threat to Face Recognition. IEEE BTAS.
5. Tan, M., Le, Q. (2019). EfficientNet: Rethinking Model Scaling. ICML.
6. Sandler, M., et al. (2018). MobileNetV2: Inverted Residuals and Linear Bottlenecks. CVPR.
7. Zhang, X. et al. (2021). TinyML: Machine Learning for Embedded Systems. O'Reilly Media.