GOCE DELCEV UNIVERSITY - STIP
FACULTY OF COMPUTER SCIENCE

ISSN 2545-4803 on line
DOI: 10.46763/BJAMI

BALKAN JOURNAL OF APPLIED MATHEMATICS AND INFORMATICS (BJAMI)

YEAR 2025 VOLUME 8, Number 2

AIMS AND SCOPE:
BJAMI publishes original research articles in the areas of applied mathematics and informatics.

Topics:
1. Computer science;
2. Computer and software engineering;
3. Information technology;
4. Computer security;
5. Electrical engineering;
6. Telecommunication;
7. Mathematics and its applications;
8. Articles of interdisciplinary of computer and information sciences with education, economics, environmental, health, and engineering.

BALKAN JOURNAL
OF APPLIED MATHEMATICS AND INFORMATICS
(BJAMI), Vol 8

ISSN 2545-4803 on line
Vol. 8, No. 2, Year 2025

EDITORIAL BOARD

TABLE OF C O N T E N TS

Balkan Journal of Applied Mathematics and Informatics Online ISSN 2545-4803
Volume 8 Number 2 Year 2025 UDC: 077.3-021.111:004.78.056-047.44

COMPARATIVE EVALUATION AND ANALYSIS OF DIFFERENT DEEPFAKE DETECTORS

JORDAN POP-KARTOV, ALEKSANDRA MILEVA, AND CVETA MARTINOVSKA BANDE

Abstract. Deepfakes, synthetic media generated using deep learning, pose significant risks to the integrity, security, and trust of information. Reliable detection is therefore critical, yet existing models often fail when exposed to real-world distortions such as compression, occlusion, and lighting variations. This paper presents a comparative evaluation of deepfake detection models, including XceptionNet, EfficientNet, MesoNet, and Vision Transformers, across multiple benchmark datasets such as FaceForensics++, DFDC, Celeb-DF, and Wild-Deepfake. Models are assessed not only under pristine conditions but also under controlled distortions that reflect realistic deployment environments. The results show that XceptionNet and the fine-tuned Vision Transformers achieve the strongest accuracy and robustness, maintaining competitive performance across domains, while MesoNet demonstrates computational efficiency but suffers from reduced reliability under challenging conditions. EfficientNet provides a balance between parameter efficiency and detection quality but lags behind in cross-dataset generalization. The findings highlight clear trade-offs between robustness, efficiency, and deployment feasibility, emphasizing that lightweight models are best suited for edge scenarios, whereas more complex architectures remain preferable in cloud or high-resource environments. The study concludes with open challenges and future research directions, including the integration of multimodal cues, domain adaptation, and explainable detection frameworks, to improve resilience against increasingly sophisticated deepfake generation techniques.

1. INTRODUCTION

Deepfakes are synthetic media, typically videos or audio recordings, created using advanced deep learning models, particularly Generative Adversarial Networks (GANs) and autoencoders. These techniques allow for highly realistic facial swaps, lip-syncing, and voice imitation. Although such technologies have useful applications in film, education, accessibility, and virtual reality, they also pose serious threats to the integrity of information and media.

Date: November 4, 2025.
Keywords. Deepfake detection; Generative Adversarial Networks; Computer Vision; Robust- ness; XceptionNet; Vision Transformers; Benchmarking.

The increasing use of deepfakes to spread misinformation, forge identities, and manipulate public perception presents an urgent need for reliable detection systems. Media organizations, social networks, and government institutions are especially vulnerable, as malicious actors exploit deepfakes to erode trust and disseminate false content.

Publicly available datasets such as FaceForensics++ [14], DFDC [15], WildDeep- fake [18], and SHIELD [13] are essential not only for training but also for cross-domain testing. These datasets vary in manipulation type, resolution, and subject diversity.

The models studied in this paper are XceptionNet [19], EfficientNet [20], MesoNet [21], Vision Transformers (ViTs) [22], and ConvLSTM [23]. They represent different families of detection strategies. XceptionNet, derived from image classification tasks, was one of the best performing models in DFDC [1]. EfficientNet offers a parameter-efficient variant. The compact design of MesoNet enables deployment in low-power devices, while ViTs and ConvLSTM introduce attention and temporal awareness. Their implementation details and pretrained weights are available in public repositories, as documented in references [3, 4, 28].

In summary, this paper builds upon a growing body of literature that emphasizes the need for practical, robust, and explainable deepfake detection. It incorporates insights from leading benchmarks and surveys to assess how resilient detection models truly are when exposed to realistic and imperfect inputs. Our main contributions are as follows:

- We provide a comprehensive comparison of deepfake detection models, including XceptionNet, EfficientNet, MesoNet, Vision Transformers, and Con- vLSTM, highlighting their strengths and weaknesses across multiple datasets.
- We evaluated these models on widely used datasets (DFDC[15], Celeb-DF[16], FaceForensics++ [14], WildDeepfake [18]), ensuring a fair and re-producible benchmarking environment.
- We analyze the robustness of detection methods under realistic conditions, such as compression, noise, and other distortions, which often degrade performance in real-world scenarios.
- We discuss the trade-offs between model complexity, detection accuracy, and deployment feasibility (e.g., lightweight models vs. large-scale architectures).
- We outline open challenges and future research directions toward more resilient and explainable deepfake detection systems.

The remainder of this paper is structured as follows. Section 2 reviews related work on deepfake generation and detection techniques. Section 3 introduces the

detection models studied in this work and the datasets used for evaluation, followed by Section 4, which reports and discusses the experimental results. Finally, Section 5 concludes the paper and outlines the directions for future research.

2. RELATED WORK

The idea of conducting a comparative evaluation of deepfake detection models and their robustness against real-world distortions is derived heavily from existing benchmark efforts, surveys, and challenge initiatives.

The Deepfake Detection Challenge (DFDC) [1] was a groundbreaking initiative launched by Facebook and its partners, providing a massive dataset of over 100,000 videos to evaluate detection models in diverse scenarios. It emphasized generalization and encouraged the development of models capable of operating on videos that vary in compression, background, ethnicity, and gender.

Following DFDC, DeepfakeBench [2] and SoK benchmark [3] proposed structured evaluation frameworks. DeepfakeBench introduced a unified testing suite with common metrics and conditions, while the SoK benchmark laid out a taxonomy for model classification and emphasized the need for standardized benchmarking in the field. Together, they brought much-needed rigor and comparability to the evaluation process [4].

Earlier deepfake detection efforts were based on hand-crafted features and statistical inconsistencies. For example, models looked for irregular blinking or texture mismatches in facial regions [5]. These early attempts, while useful in constrained scenarios, were often brittle. As deepfake generation improved, such cues became less reliable. This limitation motivated the transition toward deep learning–based approaches, such as MesoNet [7] and CNN-based pipelines [8], which learn discriminative visual patterns directly from data rather than relying on predefined features.

Modern detection models use deep neural networks capable of learning subtle and complex patterns. Surveys such as Zhang, 2023 [4] and Zhang, 2024 [9] provide an overview of these changes. They highlight the transition from binary classification to more complex spatio-temporal and multimodal analysis, incorporating attention mechanisms and ensemble strategies. Deepfake detection models often assume ideal conditions, yet real-world scenarios introduce significant distortions that degrade performance. These include video compression, lighting variation, occlusions (e.g., sunglasses, masks), motion blur, and audio desynchronization. The need to evaluate detection robustness under such conditions has become a critical concern, as emphasized by recent works on cross-dataset generalization and domain adaptation [11] and augmentation-driven robustness strategies [12].

The *WildDeepfake* dataset addresses this issue by introducing videos “in the wild,” collected from uncontrolled online sources such as social media and public video platforms. These videos naturally contain real-world artifacts such as motion

blur, inconsistent lighting, occlusions, and compression artifacts. WildDeepfake challenges the generalization ability of detectors, revealing that models trained on synthetic and pristine datasets often experience accuracy drops of more than 20% when tested in such uncontrolled environments [6].

Similarly, *SHIELD* [13] proposes a benchmarking framework specifically designed to evaluate the robustness of detection. It systematically tests models against a range of attacks and distortions in both visual and audio modalities. SHIELD highlights how existing detectors fail to generalize when exposed to noise, compression, or multimodal inconsistencies, and advocates for robustness as a core design goal rather than an afterthought.

In this work, we build on these efforts by conducting a systematic survey of robustness across commonly encountered real-world distortions. Unlike prior datasets that focus on isolated cases or lack modality diversity, our analysis compares multiple detection models with consistent distortion settings. We simulate controlled variations in compression rates, illumination conditions, occlusion types, and temporal/audio misalignment to evaluate performance stability. Our goal is to quantify robustness gaps and identify architectural patterns that promote generalization in diverse scenarios.

3. Methodology

3.1. **Datasets.** To comprehensively evaluate the robustness of deepfake detection models, we utilize multiple publicly available datasets, each of which presents diverse types of manipulation and real-world conditions. These include FaceForensics++ [14], DFDC [15], WildDeepfake [17]. By combining these datasets, we ensure coverage of a wide range of forgery artifacts, compression levels, and video qualities, which is critical for assessing generalization and robustness.

3.2. **Data Preprocessing and Augmentation.** To prepare the input data for training and evaluation, we extract individual frames from videos and apply a series of preprocessing steps designed to improve model generalization across diverse conditions. Each frame is resized to an input resolution of 299×299. For temporal representation, we sample 20 frames per video at uniform intervals, capturing both spatial and temporal information relevant for detection. The training process is configured with 30 epochs and a batch size of 8, balancing computational efficiency with stable gradient updates.

During training, various data augmentation techniques are applied to video frames. These include random horizontal flipping to simulate orientation changes, slight random rotations to introduce viewpoint variability, and color jittering to mimic variations in lighting conditions such as brightness, contrast, and saturation. These augmentations are standard in computer vision tasks and help improve robustness by exposing the model to diverse visual variations [27, 24].

Following augmentation, the frames are converted to tensor format and normalized with standard ImageNet mean and standard deviation values, which is a common practice when using pretrained convolutional neural networks [**?**]. Normalization stabilizes training by ensuring consistent input feature distributions.

During evaluation, a simpler preprocessing pipeline is used: frames are converted to tensors and normalized identically but without augmentation, ensuring a consistent and fair assessment of model performance [25].

These transformations are implemented within dataset classes that manage frame loading, transformation application, and label handling for supervised learning [26].

3.3. **Model Architectures and Training.** Our experiments focus primarily on convolutional neural network (CNN) architectures known for their effectiveness in image and video forgery detection. Models such as XceptionNet, EfficientNet, and MesoNet are trained from scratch or fine-tuned using pretrained weights where available. Training is performed using mini-batch stochastic gradient descent with cross-entropy loss as the objective function.

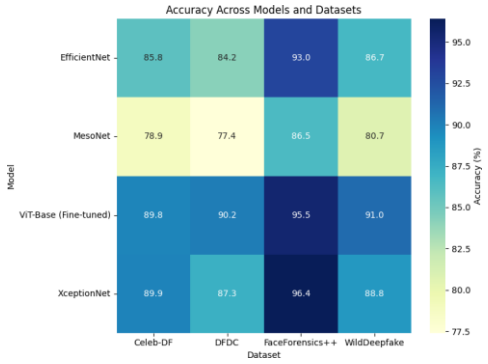
Hyperparameters such as learning rate, batch size, and number of epochs are optimized through validation experiments. Early stopping is used to prevent overfitting. Additionally, we explore models with temporal awareness (e.g., ConvLSTM) and attention mechanisms (e.g., Vision Transformers) to evaluate the benefits of spatiotemporal and multimodal features.

4. RESULTS

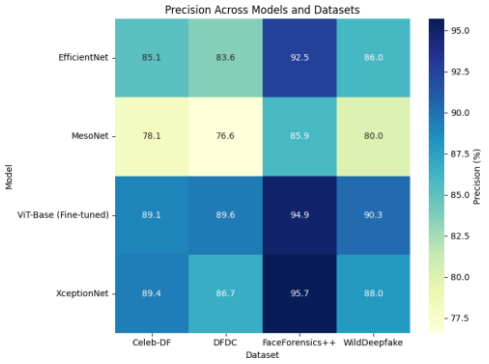
Building on the described dataset preparation, augmentation strategies, and model architectures, the following section presents a comprehensive benchmark of selected deepfake detection models. We evaluated their performance across multiple datasets and under various controlled distortions to assess their robustness and generalization capabilities. This systematic comparison provides insight into the strengths and limitations of current approaches, guiding future improvements in real-world deepfake detection.

The evaluation covers several key performance metrics, including accuracy, precision, recall, and F1-score, measured consistently across clean and distorted data conditions. We specifically analyze how factors such as video compression, varying illumination, occlusions, and temporal misalignments affect detection efficacy. By benchmarking models such as XceptionNet, EfficientNet, MesoNet and Vision Transformers under these challenging scenarios, we aim to identify architectural features and training strategies that contribute to improved resilience.

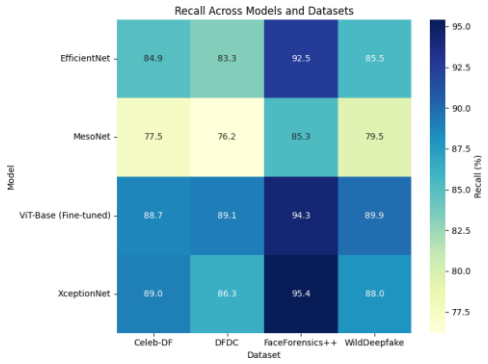
Furthermore, the results highlight the trade-offs between model complexity, computational efficiency, and robustness, which are critical considerations for deploying deepfake detectors in real-world applications. We also discuss cases where certain models exhibit notable failure modes or robustness gaps, emphasizing the need for continued research on adaptive and multimodal detection techniques.



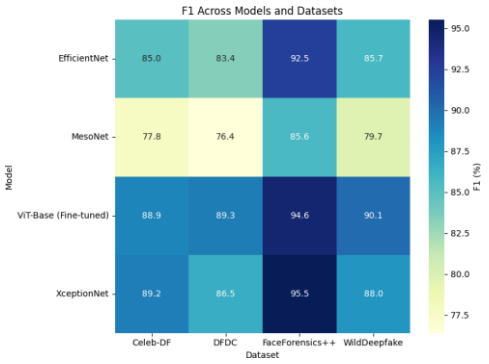
(a) Heatmap of Accuracy Across Deepfake Detection Models and Datasets



(b) Heatmap of Precision Across Deepfake Detection Models and Datasets



(c) Heatmap of Recall Across Deepfake Detection Models and Datasets



(d) Heatmap of F1-Score Across Deepfake Detection Models and Datasets

FIGURE 1. Heatmap comparison of Accuracy, Precision, Recall, and F1-Score across models and datasets.

4.1. **Model Performance Across Different Datasets.**

4.2. Figure 1 presents the performance of selected deepfake detection models across multiple benchmark datasets. Overall, the results indicate that most models achieve high accuracy on controlled datasets such as FaceForensics++, but their performance consistently drops when evaluated on more challenging real-world datasets such as WildDeepfake, Celeb-DF, and DFDC. In particular, XceptionNet achieves the highest performance in Face-Forensics++ with an accuracy of 96.4%, while also maintaining competitive results on cross-dataset evaluations. Similarly, the fine-tuned Vision Transformer (ViT- Base) demonstrates strong generalization, surpassing EfficientNet and MesoNet on WildDeepfake and DFDC, which reflects its advantage in capturing global spatial dependencies.

MesoNet, being a lightweight architecture with only 0.28M parameters, shows significantly lower accuracy across all datasets (77–86%), suggesting that while computationally efficient, it sacrifices robustness. EfficientNet performs better than MesoNet, but still lags behind XceptionNet and ViT-Base, particularly under domain shifts. Importantly, the observed decline in performance on WildDeepfake (e.g., 80.7% for MesoNet and 86.7% for EfficientNet) highlights the difficulty of detecting manipulated content in unconstrained real-world scenarios compared to curated datasets.

Taken together, these findings underline a clear trade-off: complex architectures such as XceptionNet and ViT achieve stronger robustness and cross-dataset generalization, albeit at higher computational cost, whereas lightweight models like MesoNet are more efficient but less reliable in challenging conditions. These results emphasize the importance of balancing accuracy, robustness, and efficiency when designing practical deepfake detection systems for deployment in real-world applications.

TABLE 1. Model Accuracy Under Different Compression Levels

Model	**Compression**	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1 Score**
XceptionNet	Raw	96.4	95.7	95.4	95.5
XceptionNet	HQ (c23)	92.8	92.0	91.7	91.9
XceptionNet	LQ (c40)	86.2	85.4	85.0	85.2
EfficientNet	Raw	93.0	92.5	92.5	92.5
EfficientNet	HQ (c23)	89.4	88.7	88.3	88.5
EfficientNet	LQ (c40)	83.8	83.0	82.5	82.7
MesoNet	Raw	86.5	85.9	85.3	85.6
MesoNet	HQ (c23)	82.3	81.6	81.1	81.3
MesoNet	LQ (c40)	76.9	76.1	75.7	75.9
ViT-Base (Fine-tuned)	Raw	95.5	94.9	94.3	94.6
ViT-Base (Fine-tuned)	HQ (c23)	90.2	89.5	89.1	89.3
ViT-Base (Fine-tuned)	LQ (c40)	84.7	84.0	83.6	83.8

4.3. **Robustness to Compression.** To simulate realistic deployment conditions where videos may undergo varying degrees of compression, we evaluated the robustness of each model in three scenarios: no compression, medium compression (HQ, c24) and high compression (LQ, c40).

Table 1 reports the performance of the deepfake detection models in the FaceForensics++ dataset at different levels of video compression. As expected, all models achieve their highest accuracy in raw uncompressed videos, with XceptionNet reaching 96.4% and ViT-Base closely following at 95.5%. However, performance consistently degrades as compression intensifies. Under high compression (c40), XceptionNet drops to 86.2%, EfficientNet to 83.8%, MesoNet to 76.9%, and ViT-Base to 84.7%, highlighting the sensitivity of detection networks to lossy encoding artifacts.

Among the architectures evaluated, XceptionNet shows the strongest robustness across compression levels, maintaining accuracy greater than 92% even under moderate compression (c23). The Vision Transformer also performs competitively, but its accuracy declines more sharply under severe compression compared to XceptionNet. EfficientNet follows a similar trend, while MesoNet exhibits the steepest decline, losing nearly 10 percentage points from raw to high compression, highlighting its limited resilience despite its computational efficiency.

Overall, these findings suggest that, while models like XceptionNet and ViT can generalize reasonably well under moderate compression, the presence of heavy compression substantially reduces detection reliability. This has critical implications for deployment in real-world scenarios, where social networks and messaging platforms often apply aggressive compression that can conceal deepfake traces.

4.4. **Effect of Other Distortions.** We further assess model performance under realistic distortions frequently encountered in real-world settings, including occlusions (e.g., sunglasses), low lighting, and motion blur.

Table 2 summarizes the robustness of various deepfake detection models under these challenging conditions. As observed, all models experience a drop in performance compared to clean inputs, yet the extent of degradation varies significantly by architecture. XceptionNet remains the strongest performer, achieving 91.7% accuracy under occlusion and 88.4% under low light, demonstrating a relatively stable robustness. The Vision Transformer (ViT-Base) also shows competitive results, with 90.5% under occlusion and 87.6% in low-light scenarios, indicating its capacity for generalization even under challenging conditions.
EfficientNet achieves moderate robustness, performing reasonably well under occlusion (88.2%) but suffering more under low light (85.1%). MesoNet, in contrast, shows the most pronounced vulnerability to distortions, with performance dropping to 79.4% under occlusion and further to 74.9% in low-light settings. This suggests that lightweight models, while computationally efficient, may lack the representational capacity to capture subtle deepfake cues when visual quality is degraded.

TABLE 2. Model Performance with Common Video Distortions

Model	Distortion	Accuracy (%)	Precision (%)	Recall (%)	F1 Score
XceptionNet	Occlusion	91.7	90.9	90.5	90.7
XceptionNet	Low Light	88.4	87.6	87.1	87.3
EfficientNet	Occlusion	88.2	87.5	87.1	87.3
EfficientNet	Low Light	85.1	84.3	83.9	84.1
MesoNet	Occlusion	79.4	78.6	78.0	78.3
MesoNet	Low Light	74.9	74.1	73.6	73.8
ViT-Base (Fine-tuned)	Occlusion	90.5	89.8	89.3	89.6
ViT-Base (Fine-tuned)	Low Light	87.6	86.9	86.5	86.7

TABLE 3. *Computational Efficiency Metrics of Deepfake Detection Models*

Model	Parameters (M)	Inference Speed (FPS)	Model Size (MB)	Training Time (hrs)
XceptionNet	22.9	30	88	12
EfficientNet	19	35	45	10
MesoNet	0.28	50	3	3
ViT-Base (Fine-tuned)	86	25	400	20

Overall, these results highlight that occlusions and lighting variations remain critical challenges for deepfake detectors. Although advanced architectures such as XceptionNet and ViT can retain a relatively high detection accuracy under such distortions, simpler models face substantial performance degradation. These findings reinforce the need for training strategies and data augmentation techniques that explicitly account for real-world distortions to enhance detector robustness in deployment environments.

4.5. **Computational Efficiency and Deployment Considerations.**
4.6. Table 3 summarizes model sizes, inference speed (frames per second), number of parameters, and estimated training times. This information helps to evaluate practical deployment scenarios that range from edge devices to cloud environments.

5. Conclusions

This study presented a comprehensive benchmark and evaluation of deepfake detection models across multiple datasets and real-world distortions. Our experiments highlight that while models such as XceptionNet and EfficientNet achieve high accuracy on pristine datasets like FaceForensics++, their performance degrades under common distortions, including compression artifacts, occlusions, and varying lighting conditions.

Lightweight architectures like MesoNet demonstrate advantages in computational efficiency and deployment suitability on edge devices but often sacrifice some detection robustness. Transformer-based models such as ViT-Base show promise in balancing accuracy and generalization, particularly when fine-tuned on diverse datasets.

Our findings emphasize the inherent trade-offs between model complexity, speed, and robustness. These trade-offs must be carefully considered when designing practical deepfake detectors for real-world applications, especially where computational resources or latency are constrained. Additionally, robustness to adversarial and unseen perturbations remains an open challenge that requires further research, potentially through multimodal approaches and adaptive learning strategies.

Future work should explore augmenting detection frameworks with temporal and audio cues, domain adaptation techniques, and explainability methods to improve trust and reliability. By continuing to rigorously benchmark detection models under realistic conditions, the community can accelerate the development of more resilient and deployable deepfake detection systems critical to preserving media integrity.

References

[1] Dolhansky, B.; Bitton, J.; Pflaum, B.; Lu, J.; Howes, R.; Wang, M.; Canton Ferrer, C. (2020). The DeepFake Detection Challenge (DFDC) Dataset. In Proc. IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW).

[2] Yan, Z.; Zhang, Y.; Yuan, X.; Lyu, S.; Wu, B. (2023). DeepfakeBench: A Comprehensive Benchmark of Deepfake Detection. arXiv preprint arXiv:2303.00747.

[3] Le, B. M.; Kim, J.; Woo, S. S.; Moore, K.; Abuadbba, A.; Tariq, S. (2024). SoK: Systematization and Benchmarking of Deepfake Detectors in a Unified Framework. arXiv preprint arXiv:2401.05578.

[4] Pei, G.; Zhang, J.; Hu, M.; Zhang, Z.; Wang, C.; Wu, Y.; Zhai, G.; Yang, J.; Shen, C.; Tao, D. (2023). Deepfake Generation and Detection: A Benchmark and Survey. arXiv preprint arXiv:2310.10923.

[5] Li, Y.; Chang, M.-C.; Lyu, S. (2018). In Ictu Oculi: Exposing AI Created Fake Videos by Detecting Eye Blinking. In Proc. IEEE International Workshop on Information Forensics and Security (WIFS).
[6] Li, Y.; Yang, X.; Sun, P.; Qi, H.; Lyu, S. (2021). WildDeepfake: A Challenging Real-World Dataset for Deepfake Detection. arXiv preprint arXiv:2009.12037.
[7] Afchar, D.; Nozick, V.; Yamagishi, J.; Echizen, I. (2018). MesoNet: A Compact Facial Video Forgery Detection Network. In Proc. IEEE International Workshop on Information Forensics and Security (WIFS).
[8] Korshunov, P.; Ebrahimi, T. (2018). Deepfakes: A New Threat to Face Recognition? Assessment and Detection. arXiv preprint arXiv:1812.08685.
[9] Yi, J.; Zhang, C. Y.; Tao, J.; Wang, C.; Yan, X.; Ren, Y.; Gu, H.; Zhou, J. (2024). ADD 2023: Towards Audio Deepfake Detection and Analysis in the Wild. arXiv preprint arXiv:2401.02842.
[10] Kundu, R.; Balachandran, A.; Roy-Chowdhury, A. K. (2024). TruthLens: Explainable DeepFake Detection for Face Manipulated and Fully Synthetic Data. arXiv preprint arXiv:2401.06392.
[11] Cheng, J.; Yan, Z.; Zhang, Y.; Luo, Y.; Wang, Z.; Li, C. (2024). Can We Leave Deepfake Data Behind in Training Deepfake Detector? arXiv preprint arXiv:2401.15466.
[12] Sun, K.; Chen, S.; Yao, T.; Liu, H.; Sun, X.; Ding, S.; Ji, R. (2024). DiffusionFake: Enhancing Generalization in Deepfake Detection via Guided Stable Diffusion. In Proc. Advances in Neural Information Processing Systems (NeurIPS).
[13] Shi, Y.; Gao, Y.; Lai, Y.; Wang, H.; Feng, J.; He, L.; Wan, J.; Chen, C.; Yu, Z.; Cao, X. (2024). SHIELD: An Evaluation Benchmark for Face Spoofing and Forgery Detection with Multimodal Large Language Models. arXiv preprint arXiv:2401.06750.
[14] Rössler, A.; Cozzolino, D.; Verdoliva, L.; Riess, C.; Thies, J.; Nießner, M. (2019). Face-Forensics++: Learning to Detect Manipulated Facial Images. In Proc. IEEE International Conference on Computer Vision (ICCV), pp. 1–11. Dataset available at: https://github.com/ondyari/FaceForensics.
[15] Dolhansky, B.; Howes, R.; Pflaum, B.; Baram, N.; Ferrer, C. (2020). The Deepfake Detection Challenge (DFDC) Dataset. arXiv preprint arXiv:2006.07397. Dataset available at: https://ai.facebook.com/datasets/dfdc.
[16] Li, Y.; Yang, X.; Sun, P.; Qi, H.; Lyu, S. (2020). Celeb-DF: A New Dataset for DeepFake Forensics. In Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3207–3216. Dataset available at: https://github.com/yuezunli/celeb-deepfakeforensics.
[17] Zhou, Y.; Han, X.; Mao, H.; Yang, Y.; Liu, J.; Chen, W. (2021). DeepFake Detection in the Wild. In Proc. ACM Multimedia (MM), pp. 2387–2395.
[18] Zi, B.; Chang, M.; Chen, J.; Ma, X.; Jiang, Y.-G. (2020). WildDeepfake: A Challenging Real-World Dataset for Deepfake Detection. In Proc. 28th ACM International Conference on Multimedia (ACM MM), pp. 2730–2738. Dataset available at: https://github.com/OpenTAI/wild-deepfake.
[19] Chollet, F. (2017). Xception: Deep Learning with Depthwise Separable Convolutions. In Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1251–1258. Code available at: https://github.com/titu1994/Keras-Applications.
[20] Tan, M.; Le, Q. (2019). EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In Proc. International Conference on Machine Learning (ICML), pp. 6105–6114. Code and models available at: https://github.com/tensorflow/tpu/tree/master/models/official/efficientnet.

[21] Afchar, D.; Nozick, V.; Yamagishi, J.; Echizen, I. (2018). MesoNet: A Compact Facial Video Forgery Detection Network. In Proc. IEEE International Workshop on Information Forensics and Security (WIFS). Code available at: https://github.com/DariusAf/MesoNet.
[22] Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; et al. (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv preprint arXiv:2010.11929. Code and models available at: https://github.com/google-research/vision_transformer.
[23] Shi, X.; Chen, Z.; Wang, H.; Yeung, D.-Y.; Wong, W.-K.; Woo, W.-c. (2015). Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting. In Proc. Advances in Neural Information Processing Systems (NeurIPS). Code available at: https://github.com/ndrplz/ConvLSTM_pytorch.
[24] Krizhevsky, A.; Sutskever, I.; Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. In Advances in Neural Information Processing Systems (NeurIPS).
[25] He, K.; Zhang, X.; Ren, S.; Sun, J. (2016). Deep Residual Learning for Image Recognition. In Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
[26] Paszke, A.; et al. (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library. In Advances in Neural Information Processing Systems (NeurIPS).
[27]Shorten, C.; Khoshgoftaar, T. M. (2019). A Survey on Image Data Augmentation for Deep Learning. Journal of Big Data, vol. 6, no. 60.
[28] Nirkin, Y.; Wolf, L.; Keller, Y.; Hassner, T. (2022). DeepFake Detection Based on the Discrepancy Between the Face and its Context. IEEE Transactions on Pattern Analysis and Machine Intelligence.

Jordan-Pop-Kartov,
Goce Delcev University,
Faculty of Computer Science
Stip, North Macedonia
E-mail address: jordan.popkartov@ugd.edu.mk

Aleksandra Mileva
Goce Delcev University,
Faculty of Computer Science
Stip, North Macedonia
E-mail address: aleksandra.mileva@ugd.edu.mk

Cveta Martinovska Bande,
Goce Delcev University,
Faculty of Computer Science
Stip, North Macedonia
E-mail address: cveta.martinovska@ugd.edu.mk