Bulletin of Electrical Engineering and Informatics
Vol. 15, No. 6, December 2026, pp. 5136~5148
ISSN: 2302-9285, DOI: 10.11591/eei.v15i6.11675

A convolutional neural network-long short-term memory deep learning classification framework for deepfake detection

Mohd Khaled Y. Shambour, Fadwa Feras Abu Hara, Basil Mohammed Abu-taleb, Mohammad Hassan Ali Al-Onaizan, Mohammad Klaib
Department of Intelligent Systems Engineering, Faculty of Engineering and Design, Middle East University, Amman, Jordan

Article Info

Article history:

Received Oct 20, 2025
Revised Aug 23, 2026
Accepted Sep 12, 2026

Keywords:

Deep learning
Deepfake
Long short-term memory
ResNeXt50
Video sequences

ABSTRACT

The credibility of digital media is being greatly undermined by the rise of deepfake (DF) technology. To help solve this issue, this paper proposes a hybrid spatiotemporal deep learning model that combines the ResNeXt50 convolutional backbone and the long short-term memory (LSTM) model to capture both frame-level spatial artifacts and temporal inconsistencies across video sequences. The novelty of this paper lies in the development of a lightweight convolutional neural network (CNN)-LSTM model that integrates selective fine-tuning of ResNeXt50 with mean-based temporal pooling, enabling improved generalization and computational efficiency for robust DF detection in real-world scenarios. The model was trained and evaluated on the WildDF dataset of 42,456 video sequences. Experimental results revealed that the proposed CNN-LSTM outperformed five baseline models, achieving 79.54% accuracy, 75.80% precision, 73.78% recall, and an F1-score of 74.77%. The proposed approach provides a practical framework for combining spatial and temporal information in DF detection while maintaining a lightweight architecture. This framework can serve as an intelligent classification component within an advanced intelligent classification system, supporting automated and reliable identification of manipulated visual content.

This is an open access article under the CC BY-SA license.

Corresponding Author:

Mohd Khaled Y. Shambour
Department of Intelligent Systems Engineering, Faculty of Engineering and Design
Middle East University
Amman 11831, Jordan
Email: m.shambour@meu.edu.jo

1. INTRODUCTION

The increasing and accelerating use of modern digital technologies has extensive implications on innovation, healthcare, economic development, and various other sectors [1]–[5]. But with these advantages, among the most serious threats that we have is the development of deepfake (DF) technology [6]. DFs represent malicious ways of spreading misinformation when used to damage reputations, impersonate some people, or shape the opinion of the masses [7]. With the advancing technology of DF generators, real and fake video contents become harder to identify even with the help of machines or humans [8], [9]. Therefore, DF has severe societal consequences that required the practical and applicable solutions. Synthetic media pose a threat to the democratic processes as they destroy the trust of the population, put their personal security at risk due to impersonation and spread of illegal material, and put financial systems at risk due to fraud. It also compromises the integrity of the media because it makes the verification of the content more challenging.

To address these risks, this work proposes a simple and scalable model, combining convolutional neural network (CNN) and long short-term memory (LSTM) models, designed for practical application, focusing on efficiency, generalizability, and minimizing critical errors, including the false negatives.

Although CNN-LSTM hybrid models have been introduced to detect DFs, the current work presents several contributions that make it more unique compared to previous methods:

- First, a freeze-based fine-tuning strategy is applied, where all the ResNeXt50 layers are frozen except layer4. This is unlike typical methods which either overfit the whole backbone (risks overfitting on small data) or freeze all layers (inadequate capacity to adapt to DF-specific artifacts).
- Second, a challenging in-the-wild benchmark dataset, called WildDF, was selected in this work. This dataset has a variety of real-world conditions, instead of other existing hybrid models which are only tested on controlled datasets. Such a selection allows evaluating the generalization to the conditions of the real deployment.
- Third, a lightweight efficiency (single-layer LSTM, 512 units, and selective fine-tuning) was utilized, which is appropriate in real-time applications, taking into account the scalability issues that are frequently overlooked in detection research.

These issues are enforced by the reality that DFs can be very realistic and have temporal continuity between frames [10]. Beyond achieving high technical accuracy, there is a critical need for scalable detection systems that can be integrated into real-world social media platforms. Unlike computationally intensive foundation models, our proposed lightweight hybrid architecture is designed for efficient deployment, offering a more scalable solution for monitoring vast streams of digital content.

The novelty of this work lies in the development of a lightweight hybrid architecture that integrates a selectively fine-tuned ResNeXt50 backbone [11] with an LSTM network using mean-type time pooling to effectively bridge the performance gap in detecting diverse manipulations within the WildDF dataset.

The ResNeXt50 applies to extract detailed spatial information in the individual frames of a video, whereas the temporal dynamics of the video at the sequence of frames is modelled by the LSTM network. The main objective is to detect visual artifacts in single frames, as well as to detect unnatural temporal trends thus, essentially, and making DFs detectable.

The main objective of the research is to improve the detection of DF videos by employing a lightweight CNN–LSTM architecture to generalize spatiotemporally, and to test its performance on different wild data quantitatively by analyzing the classification accuracy, loss stability, and confusion-matrix-based error characteristics on the WildDF dataset.

The training and assessment of the model is performed on the WildDF dataset [12] with a high diversity of the real and fake videos examples that can reproduce a realistic scenario. This assists the model to learn to generalize and be reliable with unseen data. The remainder of this paper is structured as; section 2 presents the related works, while section 3 details the proposed methodology. Section 4 presents the results and discussion, and section 5 concludes the study with key findings and future works.

2. RELATED WORKS

Recent work on DF research highlights that not only the sophistication of the methods of generation is increasing, but the methods of detecting them also have some limitations. To provide a clearer analytical synthesis, prior works are reviewed according to four major methodological categories: spatial CNN-based detectors, temporal and spatiotemporal architectures, multimodal and auxiliary-based detection frameworks, and generalization-focused and foundation-model-driven approaches.

2.1. Spatial convolutional neural network-based deepfake detectors

CNN backbone approaches are early and still current used detectors regarded DF detection as a frame-level binary classification task, like Xception, EfficientNet, and ResNet, which were often trained on datasets resembling CelebA/FaceForensics++. For instance, Nelson *et al*. [13] utilized XceptionNet on CelebA images and achieved 95.56% accuracy by revealing subtle spatial artifacts through depth-wise separable convolutions.

Recent research has fused convolution and attention to simultaneously grasp local texture and global context. Convolution- and attention-based weighted ensemble network (CAE)-Net couples EfficientNet, ConvNeXt, and DeiT using wavelet transforms, thus crowning with the 94.63% and 97.37% AUC scores on the DFWild-Cup DF image benchmark, and also displaying some robustness against adversarial perturbations [14]. Lightweight yet specialized 3D CNNs have also been recommended in which the fusing of spatial-rich model features is performed over short frame sequences to spot video DFs with a minimal number of parameters compared to conventional 3D CNNs, and the models still beat some state-of-the-art baselines [15]. Spatial and short-term spatiotemporal methods secure their position at the top of curated

benchmarks but they do not rarely depend on particular artifact patterns and may miss capturing longer-range temporal inconsistencies or domain shifts.

2.2. Temporal and spatiotemporal detection architectures

In order to analyze the action beyond single-frame, lots of studies apply explicit temporal modeling incorporating analysis of three dimensions. It is the case that the CNN-LSTM or CNN-gated recurrent unit (GRU) hybrid models learn the spatial features in the frame and then temporal dependencies are modeled. For instance, where hybrids are based on optical flow, the deal between frames is captured and 91.21%, 79.49%, and 66.26% accuracy are achieved for FaceForensics++, Celeb-DF, and DF detection challenge (DFDC), respectively, with a minuscule number of frames per video, thus enabling the detection to be early [16]. The other hybrid CNN-LSTM specifications also disclose high accuracy in DFDC-like video benchmarks by blending frame-wise feature extraction with sequential modelling [17]–[19].

In addition to one single architecture, Zhang *et al*. [20] suggest particle swarm optimization (PSO)-optimized weighted and evolving ensembles comprising 3D CNNs, CNN-recurrent neural networks (RNNs). Their particle-swarm method not just searches network topology but hyper-parameters too, thus, yielding ensembles that statistically outperform other search strategies and previous ensembles on various synthetic video datasets by better utilizing fine-grained spatiotemporal cues. These temporal architectures have greatly boosted the performance on standard video datasets, yet most of the evaluations have been done within the same dataset; only a small part of their strength under heavy distribution shift or new manipulation classes has been explored so far.

2.3. Multimodal and auxiliary-based detection frameworks

Systems that are tightly integrated put audio-visual dependencies under a common learning process. AVoiD-DF comes up with a temporal-spatial encoder, a multi-modal joint decoder, and a cross-modal classifier to make use of the differences between audio and visual streams artificially. AVoiD-DF has been trained on the DefakeAVMiT benchmark and evaluated on FakeAVCeleb and DFDC as well, surpassing many past baselines and demonstrating better generalization across forgery types [21]. Inspired by this, Javed et al. suggest improved multimodal detection that combines the local CNN features (e.g., lips and eyes) with the global vision transformer (ViT) representations and uses diffusion models for denoising and augmentation of both audio and visual inputs. Their framework achieves almost perfect accuracy (0.9987 on FakeAVCeleb and ≥0.9812 on various multimodal datasets), indicating that the combination of local-global modeling and the use of diffusion-based pre-processing could be a way to get strong robustness on the current multimodal benchmarks [22].

The auxiliary watermark-based methods are capable of even more than that, to say they are able to detect and verify. AVSecure unveils an audiovisual watermarking system that is built around a single encoder-decoder by applying it to the watermarking process in both modalities. It is able to resist the common types of post-processing (e.g., JPEG compression) but still recognizes the manipulations such as SimSwap and marks the DFs on VidTIMIT through the joint verification of audio and video [23]. The latter kinds of methods, however, require that the content be controlled at the time of creation and cannot be employed on the already-circulated forgeries.

2.4. Generalization-focused and foundation-model-based approaches

The scene is increasingly surrounded by such works that explicitly claim to deal with the problem of cross-dataset and cross-manipulation generalization. Khan and Nguyen [24] take a comprehensive look into the performance of eight supervised CNN architectures and two self-supervised transformer models distillation with no labels (DINO), contrastive language–image pre-training (CLIP) subjected to the datasets of FakeAVCeleb, CelebDF-V2, DFDC, and FaceForensics++. Their experiments involving both the datasets and the models show that transformer-based detectors are continually superior to CNNs in terms of performance, that models trained using FaceForensics++ and DFDC datasets have better generalization capabilities than models trained on CelebDF-V2 or FakeAVCeleb datasets, and that image augmentations are especially beneficial for transformer models [24].

Researchers in various fields have created different approaches to work, from architectures to training objectives, specifically to tackle the problem of domain shift. The first paper, “Domain Shift Modeling”, described a framework for ViT-based domain shift modeling that is focused on hard sample generation and diverse feature distribution through attention-guided patch masking and feature statistic shift estimation. Domain specific masking (DSM) takes the gap between the observed and unobserved forgery domains and practically closes it by masking informative patches and resampling feature statistics, which leads to its cross-dataset performance being better than that of several state-of-the-art baselines [25]. On a parallel track, Gu *et al*. [26] developed a multiscale frequency–spatial model that merges specialized filters and a feature

pyramid network with attentional feature fusion and thus achieves considerably higher cross-model detection rates, one of which is a 29.1% margin over the next best method on diffusion-generated images.

Self-supervised and disentanglement-based detectors try to separate the content that carries the meaning from the content that is considered artifacts of the manipulation. A new self-supervised feature disentanglement framework was presented by Yan *et al*. [27] where images are mixed autonomously to create training samples and a network breaks down each input into features representing the content and residuals related to the forgery, while the latter are restricted to be orthogonal. Reconstruction conditioned by the forgery features acting as control vectors leads to increased robustness against the unknown forgery types and reduced reliance on specific training distributions. Ensemble-based strategies similarly solve the generalization problem: Wahab *et al*. [28] combine several asymmetric state-of-the-art detectors and show that, across two out-of-domain datasets, no individual model is consistently best, whereas asymmetric ensembling yields more stable cross-dataset performance.

Foundation model-based DF detectors constitute a different promising avenue. MEViT combines EfficientNet vision transformer with meta-learning, a pair-discrimination loss that diverges fake features from real ones and a domain-adjustment loss to lessen manipulations across methods. With training on a section of FaceForensics++, MEViT displays better results not only on the same dataset but also on the unseen manipulations while being competitive on CelebDF-V2 [29]. On the other hand, more general analyses show that vision transformers and CLIP-like pre-training, while offering strong semantics and better out-of-domain robustness compared to CNNs, are also more computationally demanding and still often lack intrinsic temporal reasoning [24], [29].

In summation, the references indicate a shift in the DF detection methodology from artifact-based CNN detectors operating on selected datasets to the ones that are aware of the temporal aspects, multimodal and generalization-centered approaches employing transformers, self-supervision, and ensembles. Nevertheless, it is still an unsolved research challenge to develop robust, efficient, and temporally rich foundation-model-based detectors for detecting DFs that are truly in-the-wild. The novelty gap addressed in this work is the lack of a robust, hybrid spatiotemporal architecture capable of maintaining high predictive strength when confronting the varied manipulations and artifacts present in the WildDF dataset. Table 1 (see in Appendix) shows a summary of detection approaches by modality, architecture, dataset, and reported accuracy.

3. METHOD

This section summarizes the entire pipeline aimed at constructing a strong DF detection system on a hybrid CNN-LSTM model. The methodology is allocated into several processes, which are the environment preparation, obtaining the input data, preprocessing, data augmentation, and building of the dataset by a sequence as shown in Figure 1.

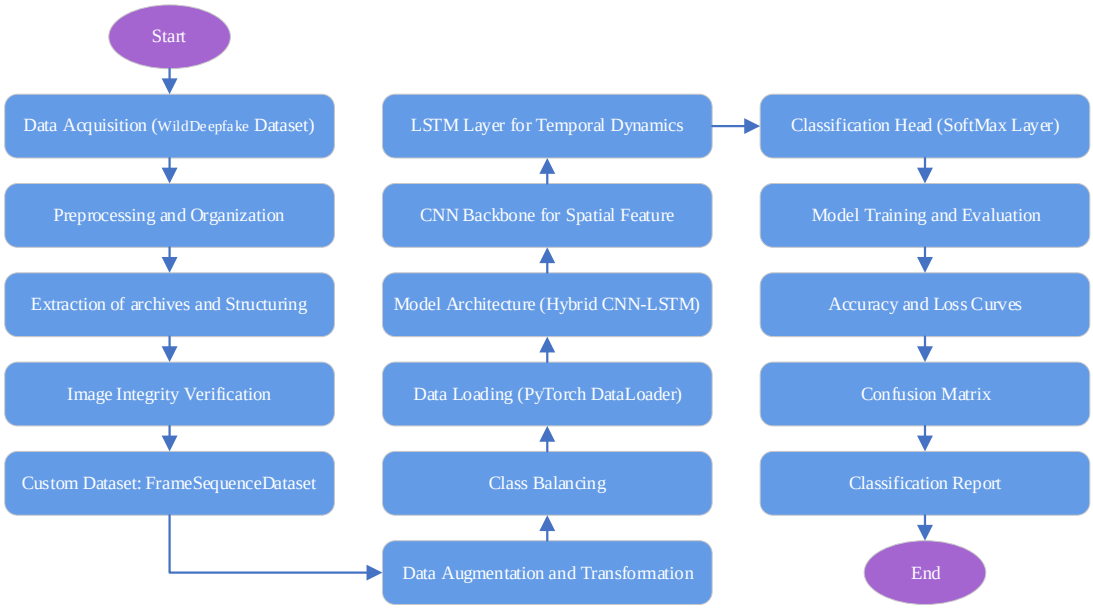


Figure 1. CNN–LSTM method flowchart

3.1. Environment, data acquisition, and preprocessing

The entire DF detection pipeline was developed using Python 3.x within the Google Colab environment. A comprehensive set of libraries was installed to handle all aspects of the pipeline, including dataset processing, preprocessing, deep learning model development, training, and evaluation. This collection of essential tools included tqdm, huggingface_hub, PyTorch Modules, PIL, torchvision transforms, and data utility tools like Dataset, DataLoader, and WeightedRandomSampler.

The dataset utilized for model training and evaluation was the WildDF dataset, which was acquired via the Hugging Face platform. This dataset was specifically chosen because it contains a high diversity of real and fake videos that accurately reproduce a realistic scenario, assisting the model in learning to generalize and be reliable with unseen data. The total dataset utilized for the evaluation phase consisted of 42,456 video sequences. By partitioning the data into an 80% training set and a 20% validation set, the model was exposed to a high diversity of real-world scenarios, assisting it in learning to generalize beyond controlled benchmarks.

Following the extraction process, an automatic integrity checker was used to scan all frames and ensure the quality of the dataset. The PIL.Image.verify() function was specifically employed for this task. This measure ensured dataset quality by automatically identifying and deleting any corrupt images, thereby preventing errors during model ingestion.

3.2. Custom dataset construction

A custom PyTorch dataset class, named "FrameSequenceDataset", was developed to organize and provide ordered sequences of frames as input to the model. This custom class manages the creation of fixed-length inputs, providing sequences such as 4 consecutive frames using a sliding window approach. It also incorporates support for robust data transformations, allowing for heavy augmentation during training while applying only minimal preprocessing for validation and testing. Crucially, the custom dataset class includes error handling functionality that facilitates the automatic skipping of any incomplete or corrupted sequences encountered during the loading process.

3.3. Data augmentation and transformation

To improve generalization and robustness of the model, separate transformation pipelines were applied using torchvision.transforms.Compose() to:

- Training transformations: where all frames are resized to 224×224, in addition to applying random horizontal flip, random rotation (±10º), colour jitter (±30%) random grayscale (10%), and normalization (ImageNet stats).
- Validation and test transformations: where the frames are also resized to 224×224, in addition to applying tensor conversion and normalization with ImageNet statistics.
- Dataset creation: the custom FrameSequenceDataset was initialized with a 4-frame sequence length, employing the full augmentation pipeline for training.

3.4. Class balancing

To remove bias in learning and provide a fair learning procedure between real and fake classes, the WeightedRandomSampler was used with inverse weights of the frequencies of classes. This model was suitable to guarantee an even composition of batches during training which played a major role in balanced classification behavior of the model. In order to specifically alleviate the class imbalance in the training stage, the following steps were used:

- Weights on each sample were calculated in terms of inverse class frequency.
- Samples were used to ensure that the classes were evenly distributed in every batch of training.
- The balance was checked by pre- and post-sampling visualizations of data distribution.

3.5. Data loaders to load dataset

Training, validation, and testing splits were initialized as DataLoader objects, and several methods were applied in this phase, including:

- Shuffling: to provide varied batches and reduce learning bias.
- Batching: allowed to train efficiently on the graphics processing unit (GPU).

Parallel loading for validation and testing datasets, where num_workers=2 and pin_memory=True.

3.6. Model architecture: convolutional neural network–long short-term memory hybrid

To capture both spatial and temporal information from video sequences, a hybrid deep learning architecture was developed by combining a CNN with a LSTM network. In this architecture, the CNN acts as

a spatial feature extractor, processing individual video frames to identify discriminative visual patterns and artifacts associated with DF manipulation, while the LSTM models the temporal relationships among the extracted frame-level features.

3.6.1. Backbone: ResNeXt50 (spatial feature extractor)

The convolutional backbone was the pretrained ResNeXt50-32x4d architecture trained on ImageNet to extract rich spatial representations of each frame. The convolutional layers were retained and the final fully connected and classification layers were removed in order to make it suitable in the extraction of features.

We also froze all layers except layer 4, and performed some fine-tuning to allow effective training and to minimize the likelihood of overfitting to the data. For each input frame, the CNN outputs a 3D tensor of shape [2048, 7, 7], where:

- 2048 represents the number of feature channels, each capturing a distinct type of visual pattern (such as edges, textures, or object parts).
- 7×7 represents the spatial resolution of each feature map i.e., how those features are spatially distributed across the height and width of the frame.

To reduce this rich spatial information into a compact vector, we apply an adaptive average pooling layer (AdaptiveAvgPool2d(1, 1)), which computes the average of each 7×7 feature map, resulting in a single scalar per channel. This transforms the output into a flat vector of size 2048 per frame. Thus, for each video sequence of T frames, the CNN backbone produces a sequence of T feature vectors, each of size 2048.

3.6.2. Temporal dynamics modeling

To capture the temporal dynamics between consecutive frames in a video sequence, we incorporated a LSTM network on top of the CNN feature extractor. This allows the model to learn how visual features evolve over time an essential aspect for identifying subtle and sequential inconsistencies present in DF videos. The LSTM processes a 3D input tensor of shape [batch_size, sequence_length, 2048], representing batches of video sequences where each frame is encoded as a 2048-dimensional feature vector. Configured with 512 hidden units in a single unidirectional layer and post-LSTM dropout of 0.5 for regularization, the model computes the temporal average of all hidden states rather than using just the final state, producing a compact [batch_size, 512] output vector that effectively summarizes the entire video sequence's temporal information. This approach provides a more comprehensive representation than using only the final hidden state while maintaining computational efficiency through the single-layer architecture.

3.6.3. Classification head

The final state of the LSTM is passed to the classification head to provide the prediction for each video sequence. A SoftMax activation function is applied in the final output to convert the model scores into class probabilities for the two classes. The model performs binary classification, where the output label 1 represents a real video and 0 represents a fake video.

4. RESULTS AND DISCUSSION

This section presents the outcomes of training and testing the proposed CNN-LSTM combination model to detect DF. The analysis consists of the criteria of the classification accuracy and the loss trends during the training process, perspective of the confusion matrix performance, and general discussion of the models' performance and potential limitations. Table 2 presents an overview of the classification accuracies of the standalone CNN and LSTM models and the proposed CNN-LSTM hybrid model to give a comparison on the baseline, demonstrating the superior result of the proposed CNN-LSTM model with an accuracy of 79.54%.

Table 2. Summary comparison of baseline vs. hybrid models

Model	Accuracy (%)
CNN (ResNeXt50)	65.61
LSTM	73.25
Proposed CNN–LSTM	79.54

4.1. Training and validation accuracy

Figure 2 shows the training and validation set accuracy of the model in achieving classification within the seven epochs. The model was able to converge quickly and in 4 epochs the training accuracy was 99.9%. This almost flawless fit testifies to the high ability of CNN-LSTM model to learn using the training samples. On the other hand, validation accuracy converged to the range of 95-95.7%, which, though is a high performance with the unseen data, demonstrates a significant difference compared to the training accuracy.

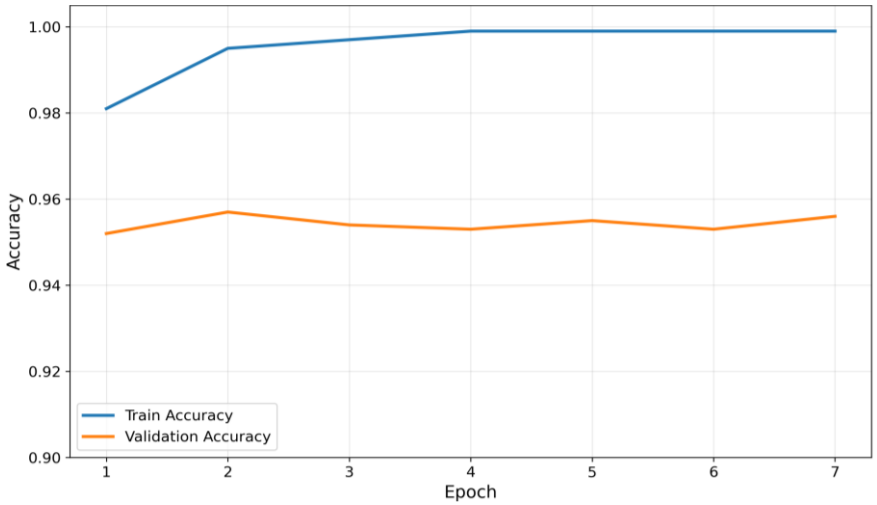


Figure 2. Accuracy results of training and validation sets

The large difference between training and validation accuracy shows that there is slight overfitting. Although the model has good generalization to validation data, its almost perfect training accuracy in comparison with the lower accuracy on the validation renders the possibility of some memorization. However, the cross-validated accuracy clearly shows that the overfitting problem is contained and the model has a strong predictive power.

4.2. Training and validation loss

Training and validation losses change similarly during the same epochs, as shown in Figure 3. The loss diminished rapidly and was close to zero (~0.001) at epoch 4 proving the good performance of the model to fit to the training data.

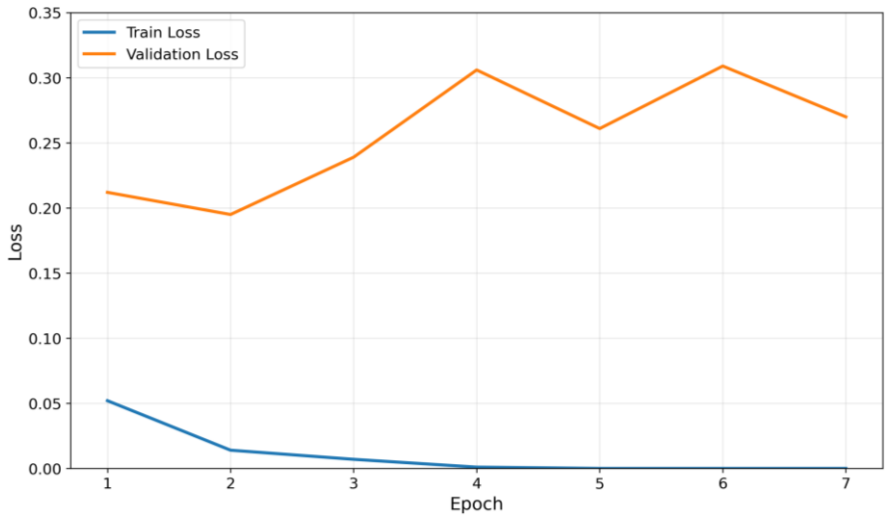


Figure 3. Training and validation losses

Moreover, the validation loss only varied between 0.19 and 0.31 indicating very slight instability. Although a slight disparity between training and validation loss emerged at the very beginning of epoch 3, the validation loss did not wander and become out of control.

The outcomes show small amounts of over-fitting; however, the validation loss was in a reasonable range, indicating that the generalization ability of the model was not significantly affected. The model's near-perfect training accuracy relative to the lower validation accuracy suggests a degree of memorization of training-specific patterns. Dropout (0.5), data augmentation, and early stopping probably contributed to this stable behavior by avoiding extreme divergence in training and validation trends.

This pattern is characteristic of controlled overfitting, in which the model fits the training data well without having an unreasonable generalization. The fact that validation loss stabilizes in a bounded range indicates that regularization measures used helped in preventing extreme deviation between training trend and validation trend as well. These findings confirm the fact that the model possesses adequate generalization to unobservable data despite the presence of training-validation performance gap.

4.3. Confusion matrix analysis

The confusion matrix depicted in Table 3 presents a confusion matrix for the proposed model on the WildDF dataset. Among the 42,456 processed video sequences, the model was able to correctly classify 12,878 fake sequences (true positives) and 20,892 real sequences (true negatives) as such. On the other hand, 4,576 fake sequences were misclassified as real (false negatives), and 4,110 real sequences were classified as fake (false positives).

Table 3. A confusion matrix of the proposed DF detection model

Classx	Predicted	
Actual	TP: 12,878	FN: 4,576
	FP: 4,110	TN: 20,892

Further metrics were calculated as depicted in Table 4 to continue the investigation of the detection behavior of the proposed model. The precision which is the percentage of correctly identified fake videos among all videos predicted as fake, was 75.80%, which means that most of the flagged videos were actually manipulated. The recall which indicates how well the model could recognize the true DF videos stood at 73.78%, thus it was estimated that nearly 26.22% of the fake sequences were left out. The presence of such false negatives is very troublesome in the case of DF detection because fake content which gets undetected could result in real-world hazards of huge magnitude such as the spreading of misinformation and loss of credibility.

Table 4. The performance metrics of the proposed model

Metric	Formula	Result (%)
Accuracy	(TP+TN)/(TP+TN+FP+FN)	79.54
Precision (P)	TP/(TP+FP)	75.80
Recall (R)	TP/(TP+FN)	73.78
F1-score	2×(P×R)/(P+R)	74.77

The F1-score was calculated to give the equilibrium between precision and recall, and it turned out to be at 74.77%, which indicates a reasonable compromise between the detection's reliability and sensitivity. The model's overall classification accuracy was 79.54%, which translates into 33,770 right predictions out of 42,456 samples. Though this accuracy signifies a pretty good overall performance, the comparatively higher false-negative rate points toward the necessity of additional refinement. Therefore, the future enhancements should be directed towards false negatives reduction without jeopardizing the model's overall balance particularly in safety-critical applications.

4.4. Comparisons with other deepfake detection models

To further illustrate the performance of the proposed CNN–LSTM hybrid model, several simulation experiments were performed on five alternative architectures under the same experimental conditions and same WildDF dataset. The compared models are ViT-B/16, EfficientNet-B4, Xception, ResNeXt50 standalone CNN, and standalone LSTM. Table 5 summarizes the performances of the compared models 4 in terms of accuracy, precision, recall, and F1 measures.

Table 5. Performance comparison of the proposed CNN-LSTM model against DF detection models

Model	Accuracy	Precision	Recall	F1
ViT-B/16	0.7771	0.5556	0.8333	0.6667
EfficientNet-B4	0.6497	0.4156	0.7619	0.5378
Xception	0.7197	0.4853	0.7857	0.6000
CNN	0.6561	0.3571	0.3571	0.3571
LSTM	0.7325	0.5000	0.0476	0.0870
Proposed CNN–LSTM	0.7954	0.7580	0.7378	0.7477

The results show that the proposed CNN–LSTM obtained the highest overall accuracy of 79.54% compared to the other models, followed by ViT-B/16 with 77.71%. Although ViT-B/16 showed a high recall of 83.33%, which is a good sign of a high capability of detecting fake sequences, its precision was significantly lower at 55.56%, implying that a high fraction of its positive prediction was a false alarm. This result is clearly observed by F1-scores, which show the differences in performance between these models, where F1-scores were 66.67% and 74.77% for ViT-B/16 and CNN-LSTM, respectively. This indicates that even though ViT-B/16 has a strong global attention mechanism, it has over-predicted in the fake class of such a dataset, which may be due to not explicitly including the time between frame sequences. Moreover, being a large-capacity transformer model, ViT-B/16 has a recall-precision imbalance with high recall rate with 83.33% but comparatively lower precision rate with 55.56%. This trend can be evidence of overfitting to the training distribution, and the model uses the false positive rate to predict the fake class, causing a high false positive rate on unseen data.

Furthermore, the Xception model, which is one of the widely adopted baselines in DF detection literature, achieved an accuracy of 71.97% and a recall of 78.57% but has a low precision score with 48.53%, resulting in a low F1-score of 60.00%. Similarly, EfficientNet-B4 recorded the lowest accuracy among the compared models at 64.97%, with a precision and F1-score of 41.56% and 53.78%, respectively. This indicates a difficulty in reliably distinguishing real from manipulated content on this dataset without temporal context.

The standalone CNN has an accuracy of 65.61% and low rates of precision, recall, and F1-score, each at 35.71%. This indicates that spatial features alone cannot be used to achieve strong detection. On the other hand, the standalone LSTM achieved an intermediate accuracy of 73.25% and low to very low rates for the rest of precision, recall, and F1-score with 50.00%, 4.76%, and 8.70%, respectively. This almost complete recall failure suggests that when the LSTM is used without the rich spatial representations that the CNN backbone offers, it does not effectively detect fake sequences and instead adopts a more largely real-class prediction behavior.

The proposed CNN–LSTM hybrid model addresses the shortcomings of both standalone architectures while incorporating several design choices specifically aimed at mitigating overfitting. ResNeXt50 backbone was partially frozen, leaving the last group of layers fine-tuned, which reduced the number of trainable parameters and minimized the chances of overfitting to training-specific features. An LSTM dropout rate of 0.5 was also used to avoid the co-adaptation of hidden units, and aggressive data augmentation was used in training to raise the input diversity. Moreover, class balancing through WeightedRandomSampler made sure the model was not biased on the dominating class, which further led to the learning that is stable and generalizable. These measures combined allowed achieving the most balanced result of all four metrics with a precision of 75.80% and a recall of 73.78%, which confirms that the combination of spatial and temporal modelling and principled regularization is a key element to making dependable and generalizable DF detector systems.

4.5. Discussion and general performance

The performance results of the proposed CNN-LSTM model give a clear indication of the benefits of combining spatial and temporal modeling for DF detection. The standalone CNN model achieved 65.61% accuracy on the WildDF dataset but struggled to identify the inconsistencies between the consecutive frames. On the other hand, the standalone LSTM model achieved 73.25% accuracy but lacked access the detailed spatial data. The proposed hybrid CNN–LSTM model, however, surpassed both baselines with an accuracy of 79.54%, thus making a strong case for the joint use of spatial artifacts and temporal dynamics.

The results obtained confirm the effectiveness of the combination of ResNeXt50-based spatial feature extraction with LSTM-based temporal modeling. This hybrid architecture can detect both the spatial manipulation artifacts within the individual frames and the temporal inconsistencies across the video segments, which is the key to robust DF detection in less controlled conditions.

With respect to overfitting, the training curves showed that there was a difference between training accuracy (approximately 99.9%) and validation accuracy (approximately 95-95.7%), which implied mild overfitting. This was also supported by the fact that the training loss (~0.001) and the validation loss (0.1931) diverged after the third epoch. However, the overfitting was kept within check and did not lead to poor validation in the long run. The regularization technique used, which includes partial freezing of backbones, LSTM dropouts, data augmentation and weighted sampling of classes, proved helpful in limiting such overfitting and maintaining the overall generalization ability of the model. These results indicate that while the proposed model has strong memorization ability, the applied constraints successfully prevented this from negatively affect its performance on the unseen data.

This superiority of the proposed CNN-LSTM model is confirmed even more by the comparative analysis of five other architecture models. The proposed model achieves better results compared to ViT-B/16, EfficientNet-B4, Xception, standalone CNN, and standalone LSTM on most evaluation metrics (accuracy, precision, recall, and F1-score). It is important to note that ViT-B/16 had the second-highest

accuracy, 77.71% and the highest recall, 83.33%, but its low precision, 55.56%, and F1-score, 66.67%, indicate a tendency toward over-prediction of the fake class. Conversely, the proposed CNN-LSTM had the most balanced metric profile compared to all other compared models, and its precision of 75.80, recall of 73.78 and F1-score of 74.77 confirmed the hypothesis that the combination of spatial feature extraction, temporal dynamics modeling, and principled regularization produces a more reliable and generalizable model for detecting DFs compared to other models.

However, there are still some limitations. The false negative rate of 26.22% is an indication that the temporal feature modeling needs to be more expressive. Attention-based mechanisms, transformer-based sequential encoders, and other advanced techniques can be applied in the future research for deep temporal dependencies that are often hard to model with standard recurrent architectures. Also, the model will have to go through more diverse and competitive datasets to be considered really robust against the breakthroughs in DF generation methods and to be regarded as usable in practical situations.

For practical deployment, several considerations are important:

- API design: the model can be implemented as API which process video frames and produces a detection score and the level of confidence.
- Batch processing: in the case of offline forensic analysis, the model is capable of setting up configurable batch processing on the basis of the length of videos.
- Human-in-the-loop: the model can be effectively applied in the context of triage systems with flagged content being examined by human specialists due to the error profile being currently adequate.

5. CONCLUSION

Undetected DFs have serious legal and political consequences for security and public trust. To address this challenge, this paper proposed a hybrid CNN–LSTM model that captures both spatial artifacts and temporal inconsistencies in video sequences. By integrating ResNeXt50 for spatial feature extraction with LSTM-based temporal modeling, the proposed approach achieved 79.54% accuracy on the WildDF dataset, outperforming CNN (65.61%), LSTM (73.25%), Xception (71.97%), ViT-B/16 (77.71%), and EfficientNet-B4 (64.97%). Although ViT-B/16 achieved competitive accuracy, it suffered from a significant imbalance between recall and accuracy (83.33% recall and 55.56% precision), indicating a tendency toward over-prediction of the fake class. In contrast, the proposed model achieved a more balanced performance, demonstrating the effectiveness of combined spatial–temporal modeling.

Future research will focus on three directions; i) reducing false negatives by employing more transformer-based time models that incorporate attention mechanisms to improve the detection of long-term time dependencies; ii) evaluating cross-dataset generalization on FaceForensics++ and Celeb-DF; and iii) exploring multimodal detection by integrating audio-visual cues to enhance resistance against advanced forgeries. Future work will also investigate sensitivity analysis of key hyperparameters and comparisons with lightweight deep learning models to assess the robustness and computational efficiency of the proposed framework for real-time deployment. Furthermore, explainable, continual, and federated learning techniques will be explored to improve model transparency, adaptability, and privacy-preserving DF detection.

FUNDING INFORMATION

The authors acknowledge the Middle East University, Amman, Jordan, for funding this research project under Decision No. (2025/2024-1604).

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Mohd Khaled Y. Shambour	✓		✓	✓		✓			✓	✓	✓	✓		
Fadwa Feras Abu Hara	✓	✓	✓		✓	✓	✓	✓	✓		✓			
Basil Mohammed Abu-taleb	✓	✓	✓			✓			✓					
Mohammad Hassan Ali Al-Onaizan	✓				✓		✓		✓					
Mohammad Klaib	✓				✓		✓		✓			✓		

C : **C**onceptualization	I : **I**nvestigation	Vi : **Vi**sualization
M : **M**ethodology	R : **R**esources	Su : **Su**pervision
So : **So**ftware	D : **D**ata Curation	P : **P**roject administration
Va : **Va**lidation	O : Writing - **O**riginal Draft	Fu : **Fu**nding acquisition
Fo : **Fo**rmal analysis	E : Writing - Review & **E**diting	

CONFLICT OF INTEREST STATEMENT

The authors state no conflict of interest.

DATA AVAILABILITY

The data that support the findings of this study are available upon reasonable request.

REFERENCES

[1] S. Almuaythir, A. K. Singh, M. Alhusban, and A. O. Daoud, "Robotics technology: catalyst for sustainable development—impact on innovation, healthcare, inequality, and economic growth," *Discover Sustainability*, vol. 5, no. 1, 2024, doi: 10.1007/s43621-024-00744-y.

[2] R. I. Sumon *et al.*, "Automatic Fracture Detection Convolutional Neural Network with Multiple Attention Blocks Using Multi-Region X-Ray Data," *Life*, vol. 15, no. 7, p. 1135, Jul. 2025, doi: 10.3390/life15071135.

[3] E. A. Al-Jallad, "The role of artificial intelligence in developing science journalism strategies and enhancing political awareness," *Middle East Journal of Communication Studies*, vol. 5, no. 2, p. 6, 2026, doi: 10.71220/2585-005-002-006.

[4] M. K. Shambour, "Using Evolutionary Optimization Techniques to Improve the Efficiency of Transportation Scheduling," *Journal of Applied Data Sciences*, vol. 6, no. 2, pp. 1037–1049, May 2025, doi: 10.47738/jads.v6i2.581.

[5] M. Amer and A. M. Atbiqa, "The future of the Arab media industry in utilizing artificial intelligence technologies: A foresight study (2024 to 2034)," *Middle East Journal of Communication Studies*, vol. 5, no. 1, p. 7, 2025, doi: 10.71220/2585-005-001-007.

[6] M. Westerlund, "The emergence of deepfake technology: A review," *Technology Innovation Management Review*, vol. 9, no. 11, pp. 39–52, Jan. 2019, doi: 10.22215/TIMREVIEW/1282.

[7] P. L. Kharvi, "Understanding the Impact of AI-Generated Deepfakes on Public Opinion, Political Discourse, and Personal Security in Social Media," *IEEE Security and Privacy*, vol. 22, no. 4, pp. 115–122, Jul. 2024, doi: 10.1109/MSEC.2024.3405963.

[8] B. Le, S. Tariq, A. Abuadbba, K. Moore, and S. Woo, "Why Do Facial Deepfake Detectors Fail," in *WDC 2023-Proceedings of the 2nd Workshop on the Security Implications of Deepfakes and Cheapfakes, Held in Conjunction with ACM AsiaCCS 2023*, Jul. 2023, pp. 24–28, doi: 10.1145/3595353.3595882.

[9] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, and J. Ortega-Garcia, "Deepfakes and beyond: A Survey of face manipulation and fake detection," *Information Fusion*, vol. 64, pp. 131–148, Dec. 2020, doi: 10.1016/j.inffus.2020.06.014.

[10] M. Masood, M. Nawaz, K. M. Malik, A. Javed, A. Irtaza, and H. Malik, "Deepfakes generation and detection: state-of-the-art, open challenges, countermeasures, and way forward," *Applied Intelligence*, vol. 53, no. 4, pp. 3974–4026, 2023, doi: 10.1007/s10489-022-03766-z.

[11] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proceedings-30th IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2017*, 2017, pp. 5987–5995, doi: 10.1109/CVPR.2017.634.

[12] B. Zi, M. Chang, J. Chen, X. Ma, and Y. G. Jiang, "WildDeepfake: A Challenging Real-World Dataset for Deepfake Detection," in *MM 2020-Proceedings of the 28th ACM International Conference on Multimedia*, Oct. 2020, pp. 2382–2390, doi: 10.1145/3394171.3413769.

[13] L. Nelson, H. Batra, and P. Radha, "Deepfake Detection in Manipulated Images/Audio/Videos: A Three-Stage Multi-Modal Deep Learning Framework," *Inteligencia Artificial*, vol. 28, no. 76, pp. 20–39, Jun. 2025, doi: 10.4114/intartif.vol28iss76pp20-39.

[14] Z. Liu *et al.*, "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," in *Proceedings of the IEEE International Conference on Computer Vision*, Oct. 2021, pp. 9992–10002, doi: 10.1109/ICCV48922.2021.00986.

[15] J. Liu, K. Zhu, W. Lu, X. Luo, and X. Zhao, "A lightweight 3D convolutional neural network for deepfake detection," *International Journal of Intelligent Systems*, vol. 36, no. 9, pp. 4990–5004, Sep. 2021, doi: 10.1002/int.22499.

[16] P. Saikia, D. Dholaria, P. Yadav, V. Patel, and M. Roy, "A Hybrid CNN-LSTM model for Video Deepfake Detection by Leveraging Optical Flow Features," in *Proceedings of the International Joint Conference on Neural Networks*, Jul. 2022, pp. 1–7, doi: 10.1109/IJCNN55064.2022.9892905.

[17] L. V. M. Maguluri, H. N. V. Kothamasu, and S. D. Johnson, "Hybrid Deepfake Detection Using CNN for Spatial Analysis and LSTM for Temporal Consistency," *International Journal of Innovative Science and Research Technology*, pp. 381–387, May 2025, doi: 10.38124/ijisrt/25may346.

[18] V. N. Manju, A. Ritam, S. Supritha, and H. Jyothi, "Deepfake Video Detection Using Hybrid CNN LSTM Architecture Integrating Spatial and Temporal Analysis," in *Proceedings of 2025 International Conference on Emerging Technologies in Computing and Communication, ETCC 2025*, Jun. 2025, pp. 1–5, doi: 10.1109/ETCC65847.2025.11108350.

[19] U. Chaudhary, "Deepfake Detection Using Convolutional and Re-current Neural Network," *International Journal for Research in Applied Science and Engineering Technology*, vol. 13, no. 4, pp. 2572–2575, Apr. 2025, doi: 10.22214/ijraset.2025.68681.

[20] L. Zhang *et al.*, "Video Deepfake classification using particle swarm optimization-based evolving ensemble models," *Knowledge-Based Systems*, vol. 289, p. 111461, Apr. 2024, doi: 10.1016/j.knosys.2024.111461.

[21] W. Yang *et al.*, "AVoiD-DF: Audio-Visual Joint Learning for Detecting Deepfake," *IEEE Transactions on Information Forensics and Security*, vol. 18, pp. 2015–2029, 2023, doi: 10.1109/TIFS.2023.3262148.

[22] M. Javed, Z. Zhang, F. H. Dahri, and T. Kumar, "Enhancing multimodal deepfake detection with local–global feature integration and diffusion models," *Signal, Image and Video Processing*, vol. 19, no. 5, p. 400, May 2025, doi: 10.1007/s11760-025-03970-7.

[23] B. Guo, H. Tai, G. Luo, and Y. Zhu, "AVSecure: An Audio-Visual Watermarking Framework for Proactive Deepfake Detection," in *ICEIEC 2024-Proceedings of 2024 IEEE 14th International Conference on Electronics Information and Emergency Communication*, May 2024, pp. 75–78, doi: 10.1109/ICEIEC61773.2024.10561738.
[24] S. A. Khan and D. T. Dang-Nguyen, "Deepfake Detection: Analyzing Model Generalization Across Architectures, Datasets, and Pre-Training Paradigms," *IEEE Access*, vol. 12, pp. 1880–1908, 2024, doi: 10.1109/ACCESS.2023.3348450.
[25] J. Zhang, J. Ni, and F. Nie, "DSM: Domain Shift Modeling for general deepfake detection," *Signal Processing*, vol. 230, p. 109822, May 2025, doi: 10.1016/j.sigpro.2024.109822.
[26] S. Gu, Z. Qin, L. Xie, Z. Wang, and Y. Hu, "Multiscale Features Integrated Model for Generalizable Deepfake Detection," *International Journal of Intelligent Systems*, vol. 2025, no. 1, Jan. 2025, doi: 10.1155/int/7084582.
[27] B. Yan, P. Liu, Y. Yang, and Y. Guo, "Self-Supervised Feature Disentanglement for Deepfake Detection," *Mathematics*, vol. 13, no. 12, p. 2024, Jun. 2025, doi: 10.3390/math13122024.
[28] H. Wahab, H. Ugail, and I. L. Jaleel, "Ensemble-Based Deepfake Detection using State-of-the-Art Models with Robust Cross-Dataset Generalisation," *arXiv preprint*, 2025, doi: 10.48550/arXiv.2507.05996.
[29] V. N. Tran, H. S. Le, P. Choi, S. H. Lee, and K. R. Kwon, "MEViT: Generalization of Deepfake Detection with Meta-Learning EfficientNet Vision Transformer," *IEEE Open Journal of the Computer Society*, vol. 6, pp. 794–805, 2025, doi: 10.1109/OJCS.2025.3568044.

APPENDIX

Table 1. A summary of detection approaches by modality, architecture, dataset, and reported accuracy

Ref.	Type	Modality	Core architecture	Main dataset(s)	Reported accuracy/AUC
[13]	Three-stage	Image	XceptionNet	CelebA	95.56% accuracy
[13]	multimodal	Audio	CNN+LSTM	DEEP-VOICE	98.5% accuracy
[13]	framework	Video	XceptionNet+LSTM	FaceForensics++, DFDC, Celeb-DF	97.57% accuracy
[14]	CAE-Net ensemble	Image	EfficientNet, ConvNeXt, DeiT+wavelet (CNN+ViT ensemble)	DFWild-Cup (SP Cup 2025)	94.63% accuracy, 97.37% AUC
[15]	Lightweight 3D CNN	Video	Lightweight 3D CNN with SRM features	FF++, Deepfake-TIMIT (HQ/LQ), DFDC-pre, Celeb-DF	>99% on FF++ and TIMIT; 98.07% Celeb-DF; 93.98% DFDC-pre
[16]	Optical-flow CNN–LSTM	Video	CNN+LSTM on optical-flow maps	DFDC, FF++, Celeb-DF	66.26% (DFDC), 91.21% (FF++), 79.49% (Celeb-DF)
[20]	PSO evolving ensemble	Video	PSO-based evolving ensemble model combining I3D and MC3	FaceForensics++	96.20% accuracy, 94.25% AUC
[22]	Multimodal local–global+diffusion	Audio–visual	CNN (local)+ViT (global)+diffusion preprocessor	FakeAVCeleb, AV-Deepfake1M, TVIL, LAV-DF	99.87%, 98.25%, 99.15%, 98.12% accuracy
[21]	AVoiD-DF	Audio–visual	Temporal–spatial encoder+ multimodal joint decoder + cross-modal classifier	DefakeAVMiT, FakeAVCeleb, DFDC	95.3% accuracy, 97.6% AUC on DefakeAVMiT dataset. 83.7% accuracy, 89.2% AUC on FakeAVCeleb . 91.4% accuracy, 95.8% AUC on DFDC.
[28]	MEViT	Image/frame	EfficientNet vision transformer+meta-learning (MEViT-B7(EfficientNetB7)	intra-dataset FaceForensic++ validations (IDFF-2-23), celebDF-v2	80.1 % AUC, 72.5% accuracy
[24]	Generalization study (CNN vs transformers)	Image/frame	CNNs: Xception, Res2Net-101, EfficientNet-B7	FakeAVCeleb, CelebDF-V2, DFDC, FaceForensics++	92%-94% ACC (Intra-dataset avg)
[25]	DSM	Image/frame	ViT with AGPM+FSSE	Multiple public image datasets (e.g., FF++, DFDC)	Better cross-dataset AUC than SOTA baselines
[26]	Image (spatial+frequency features)	Image	FPN+AFF+ResNet50	Total_mix (includes GANs, DMs, and advanced models)	81.5% (Trained on ProGAN)- 88.1% (Trained on ADM)
[27]	Self-supervised feature disentanglement	Image	Encoder–decoder with disentangled content/forgery features	Training dataset: FF++; Testing AUC: CDFv2, DFD, DFDCP, DFDC	The average values for the AUC range between 75.3% and 79.9%.
[23]	AVSecure watermarking	Audio–visual	Unified AV watermark encoder–decoder+AV feature encoder	VidTIMIT	89.99%–100% AUC (according to deepfake method used: SimSwap, StarGAN, DiffAE, Wav2Lip, and SV2TTS)

BIOGRAPHIES OF AUTHORS



Mohd Khaled Y. Shambour is currently a professor specializing in intelligent systems at the Department of Intelligent Systems Engineering, Faculty of Engineering and Design, MEU, Amman, Jordan. He received his Ph.D. in Computer Science from Universiti Sains Malaysia (USM), Malaysia. Prior to his current position, he spent nine years at the Hajj and Umrah Research Center in Makkah, KSA, where he contributed to research addressing complex operational challenges associated with hajj and umrah, with a focus on intelligent optimization, scheduling, resource allocation, and decision-making. His research interests include scheduling techniques, pattern recognition, metaheuristic optimization, and machine learning, particularly their application to complex real-world systems. He can be contacted at email: m.shambour@meu.edu.jo.



Fadwa Feras Abu Hara is a distinguished student in the bachelor of Intelligent Systems Engineering program at MEU. Her technical skills include programming KNX devices. She possesses specialized technical skills in programming KNX devices, drawing electrical diagrams, and performing on-site system troubleshooting. By applying her academic knowledge to real-world challenges, her work successfully bridges the gap between engineering theory and practical application. She can be contacted at email: 202120315@stu.meu.edu.jo.



Basil Mohammed Abu-taleb is a dedicated intelligent systems engineering student at MEU with a focus on automation. He completed an intensive 7-month internship and trainer role at Quartz for Control Systems in late 2025. He specialized in commissioning KNX-based systems, including Siemens actuators and shutter modules. His contributed to presales meetings and ensured technical accuracy through site validation and testing. His internship solidified his expertise in communication protocols and collaborative engineering design. He can be contacted at email: 202120239@stu.meu.edu.jo.



Mohammad Hassan Ali Al-Onaizan is an assistant professor in the Department of Intelligent Systems Engineering at the Faculty of Engineering and Design, MEU, Amman, Jordan. He specializes in the electromagnetic properties of 3D Dirac semimetals and the behavior of magnetic nanocomposites. His interdisciplinary research, published in reputable journals such as life and diagnostics, effectively bridges computational intelligence with innovative materials science, contributing to both theoretical understanding and practical technological advancements. He can be contacted at email: m.alonaizan@meu.edu.jo.



Mohammad Klaib is an assistant professor in the Department of Intelligent Systems Engineering at the Faculty of Engineering and Design, MEU, Amman, Jordan. He received the bachelor's degree in Electronic Engineering from Al-Quds University. The master's degree in Computer Engineering from Near East University, and the Ph.D. degree in Software and Hardware testing from University Sains Malaysia. He is currently working as the head of Department Intelligent Systems Engineering, Middle East University in Jordan. His research interests include software testing, parallel computing, algorithms design, and AI. He can be contacted at email: mklaib@meu.edu.jo.