Power System Technology
ISSN:1000-3673
Received: 16-06-2025 Revised: 05-07-2025 Accepted: 22-08-2025

Lightweight Federated Deepfake Detection with Adaptive Fusion and Temporal Transformers

Stephy Joy D[1] and Dr.R.Thirumalai Selvi[2]

[1] Research Scholar, PG & Research Department of Computer Science, Government Arts College, Nandanam, Chennai.
[2] Associate Professor, PG & Research Department of Computer Science, Government Arts College, Nandanam, Chennai.
steffybala@gmail.com[1], sarasselvi@gmail.com[2]

Abstract

Highly realistic Video deepfakes are major threats to information integrity, online trust, and security. Whereas more lightweight models like BNNs pre-trained on ViTs have demonstrated potential to scale to efficient real-time deepfake detection in pretrained centralized settings, they are still prone to data privacy, heterogeneity, and scalability issues in real world applications where data are distributed and data are often limited. To overcome such difficulties, we set out a federated lightweight deepfake detection structure that will expand BNN+ViT classifiers into a similar and secretive setting. This framework presents three new modules: Adaptive Feature Fusion (AFF), where the spatial, frequency and semantics features are weighted adaptively to enhance intra-frame robustness; Temporal Transformer Fusion (TTF), a module to capture time-varying irregularities by modeling time correlations; and Federated Knowledge Distillation (FedKD), a framework that uses lightweight student models deployed at devices to inherit robustness of a central teacher model without data transfer. Experiments on DFDC, FaceForensics++ and OpenForensics show that our approach reaches an accuracy of 97.5% and an AUC of 98.3 with just 2.4 GFLOPs beating state-of-the-art models, including CNN, MesoNet, EfficientNetV2-M, and ViT-B/32, whilst being efficient in deployment. With the synthesized lightweight effectiveness with federated scalability and privacy, the presented framework is feasible and broadly applicable to real-life forensic and security use cases.

Keyword: Deepfake Detection, Knowledge Distillation, Adaptive Feature Fusion, Temporal Transformers

1. Introduction

With the appearance of powerful generative models, especially the generative adversarial networks (GANs) and diffusion-based models, it is now possible to create hyper-realistic manipulated media, commonly called deepfakes [16][28]. These artificial videos and images can be easily used to mimic human appearance and actions with great detail, and human sight can hardly be distinguished between genuine and manipulated materials. Although deepfake technologies have potential applications in filmmaking, education, and accessibility[24], their

Power System Technology
ISSN:1000-3673
Received: 16-06-2025 Revised: 05-07-2025 Accepted: 22-08-2025

malicious applications have been of concern with privacy breeches, misinformation campaigns, political propaganda, and digital fraud [4] [30] [9].

The evolution of deepfake generators is progressing at a rather explosive rate thus a research community that focuses on detecting deepfakes is also flourishing. Conventional deception tactics were based on artifacts that were created manually like inappropriate eye blinking [18], head pose inconsistency [30] or physiological inconsistency [2]. These early detectors became outdated as the methods of generative detection developed. CNN-based methods, especially XceptionNet [5], and MesoNet [29] allowed achieving highly accurate detection by automatically finding discriminative spatial features. Transformer-based models, like Vision Transformers (ViTs) [10] or even multimodal CLIP-ViT [26] have demonstrated good generalization to different manipulation methods.

However, although these advances have been made, there are two challenges that still exist. Next, computational cost: state-of-the-art models like ViT and EfficientNetV2-M [6] realize their performance using billions of FLOPs, and are therefore not suited to real-time deployment on resource-limited devices. Second, generalization and privacy: the mass of embodied detectors presumes centralized training plots, where large annotations are accessible, but in the real world, data are partitioned among several sources (i.e. news agencies, social media platforms, forensic labs in different regions). Collecting such data can be problematic in terms of privacy and legal issues [25] and naive distributed training does not always succeed with non-IID (not independent and identically distributed) data distributions [15].

Recently there has been work on trying to address efficiency by binarizing weights and activations in a Binary Neural Network (BNN) to access $\{-1, +1\}$, which reduces FLOPs and memory requirement by a large factor [7][20][21]. This has been applied to the task of detecting deepfakes, where lightweight BNN classifiers with ViTs (Lanzino et al., 2024) have appeared to be possible in real-time. However, these works are restricted to the centralized data, which makes their adaptability to various manipulations in the real world. Along this line, federated learning (FL) has become one of the new paradigms of collaborative privacy-preserving training [25]. By supporting the collaborative training of a shared model by multiple clients without access to raw data, FL should offer privacy compliance as well as the scale. However, FL to deepfake detection poses certain challenges: client data are usually non-IID, model performance can drop on absence of robust aggregation strategies, and lightweight student models may not achieve the same performance as heavy teacher models.

To fill these gaps, the proposed offering in this paper is a federated lightweight deepfake detection system that can incorporate the effectiveness of BNN+ViT-based detectors into the federated learning scenario but with new feature modeling and knowledge transfer methods. The framework brings three innovations. First, an Adaptive Feature Fusion (AFF) model which adaptively combines multiple-domain features, spatial, frequency and semantic representations, in order to better withstand a difference in resolution and compression. Second, Temporal Transformer Fusion (TTF) module, which's formulation explicitly represents the

Power System Technology
ISSN:1000-3673
Received: 16-06-2025 Revised: 05-07-2025 Accepted: 22-08-2025

consistency across frames, and is able to recognize subtle temporal deviations, e.g. unnatural blinks or lip-sync errors. Third, a Federated Knowledge Distillation (FedKD) process that allows student models in client nodes to be light (low capacity) and yet advantageous in terms of solidity compared to a central full-precision teacher model despite the heterogeneity of the data distribution and the uneven correspondence.

The major contributions of this paper can be summarized as follows:

1. We propose a federated privacy preserving deep fake detection framework that combines lightweight BNN+ViT classifiers with temporal and feature adaptive modelling.
2. We present two new modules intra-frame feature weighting (AFF) and temporal inconsistencies (TTF).
3. Unbiased FedKD enables training lightweight client models to generalize well by inheriting robustness of a central teacher by overcoming challenges of non-IID data in federated contexts.
4. We perform significant experiments on DFDC, FaceForensics++, and OpenForensics and demonstrate that our framework obtains 97.5 percent accuracy and 98.3 percent AUC using 2.4 GFLOPs, surpassing CNN, MesoNet, and ViT as well as Efficient-NetV2-M and standalone BNN models.

The rest of the paper is organized as the following. Section 2 provides a summary of the surrounding research in deepfake detection, lightweight neural networks and federated learning. Section 3 pitches in with the description of the proposed methodology, such as AFF, TTF, and FedKD. Section 4 describes the experimental setup, but Section 5 reports results, and they are analyzed, as well as compared to baseline and an ablation study. Section 6 presents directions of future research.

2. Related Work

2.1 Deepfake Detection Approaches

Initial deepfake detection efforts used manually-designed cues that use physiological and geometric anomalies. Eye behavior has been shown to be capable of detecting synthetic material with eye blinking patterns described by Li et al. (2018) showing a 98 percent success rate in detecting synthetic items. Yang et al. (2019) used inconsistent head poses to discriminate between synthetic and real imagery. Likewise, Agarwal et al. (2020) discussed lip-sync issues to detect tamper. These methods could not perform well on general and unconstrained settings, despite being effective and valuable in constrained settings.

The emergence of bigger datasets like FaceForensics++ (Rossler et al., 2019) and DeepFake Detection Challenge (DFDC) (Dolhansky et al., 2020) allowed the penetration of data-driven

Power System Technology
ISSN:1000-3673
Received: 16-06-2025 Revised: 05-07-2025 Accepted: 22-08-2025

deep learning methods into the area of research to become dominant. Recent CNN-based detectors are more discriminative as they are able to capture subtle facial artifacts (e.g. XceptionNet (Chollet, 2017) and MesoNet (Xia et al., 2022)). Some of the more recent studies incorporated frequency-based features (Durall et al., 2020) and attention modules (Zhao et al., 2021) into the mix to determine low-level inconsistencies.

Transformers have contributed more to the field Vision Transformers (ViTs) developed by Dosovitskiy et al. (2021) are able to model global dependencies. This was later expanded to multimodal embeddings with CLIP (Radford et al. 2021) and the fact that EfficientNetV2-M also showed competitive cross-dataset performance (Coccomini et al. 2023). Nevertheless, these high-capacity variants usually have some billions of floating-point operations, which makes them not suitable to be implemented on time-sensitive or edge platforms.

Summary State-of-the-art detectors have shown great performance accuracies, but have two limitations: (i) they are not computationally efficient and can therefore not be used in forensic or embedded systems, and (ii) they assume centralized training which is not true in real-life scenarios where data is often fragmented and privacy-sensitive.

2.2 Lightweight Neural Networks for Efficiency

In a bid to address the computational cost of deep models, light weight architectures have been investigated. BNNs train weights and activations to be {-1,+1}, which saves Floating-point operations and memory bandwidth to a significant degree. The mixture of Liu et al. (2018) incorporating Bi-Real Net, and Liu et al. (2020) with ReActNet was introduced by preserving the gradient information during the training stage, and by using the optimized activation functions, respectively. These models deliver significant efficiency improvements and are therefore interesting in mobile and at the edge.

In a deepfake detection setting, lightweight techniques are a recent development. Lanzino et al. (2024) considered the problem of real-time detection of deepfakes by constructing BNN-based classifiers, which are efficient but lack generalization relative to more cumbersome CNNs or transformers. On the same note, Chen et al. (2023) used transformer heads with BNNs to finetune the bias-variance tradeoff.

Notwithstanding, lightweight models have two constant problems: (i) loss of representational power that makes them weak at reflecting minute artifacts in manipulated videos, and (ii) they require centralized training, which hurts their flexibility to diverse real-world data. This indicates the necessity of the mechanisms which will enable lightweight detectors to utilize knowledge transfer without affecting the deployment efficiency.

2.3 Federated Learning for Vision and Media Forensics

Federated Learning (FL) has become an approach to distributed privacy-preserving model training (McMahan et al., 2017). Rather than passing raw data to a central server and training

Power System Technology
ISSN:1000-3673
Received: 16-06-2025 Revised: 05-07-2025 Accepted: 22-08-2025

a single model, local models are trained on client devices and subsequently transferring updated versions of the model to the central hub, then merged along with these other updates using a group of aggregation techniques, like FedAvg. The method has been used to great success in non-sensitive tasks, like next-word prediction (Hard et al., 2018) and medical imaging (Sheller et al., 2020).

Recent research has generalized FL to the visual media forensics. In another study, TI2Net by Liu et al.(2023) was a federated model that used temporal identity inconsistency as a tool to detect deepfakes. An approach based on Graph Neural Networks and used in an FL architecture was studied by El-Gayar et al. (2024), recording an enhanced resilience to cross-dataset perturbation. Nevertheless, several difficulties have not been fully solved: the client data is not always IID, which results in a drop in performance; clients with a small computing power can hardly approach the performance of the server; communication cost limits scalability (Kairouz et al., 2021).

Knowledge distillation has been suggested as a solution FedKD can trade-off between accuracy and efficiency by enabling a centralized teacher to distill knowledge to the lightweight client students (Li et al., 2022). Its usage to detect deepfake, however, has not been extensively explored, especially in terms of lightweight models like BNN+ViT that are potential candidates of real-time forensic tools.

2.4 Summary and Research Gap

The three crucial insights of the literature are available. On the one hand, deepfake detection algorithms have evolved beyond handcrafted feature identification to CNN and transformer-based solutions, but these are computationally expensive, to the point that they are not feasible to use in practice. Second, efficient neural networks, e.g., BNNs, are not robust during training when trained in isolation. Third, federated learning offers an exciting approach to privacy-preserving collaboration but has limitations of non-IID and has not been adapted to media forensics.

This paper fills these shortcomings by developing a federated lightweight deepfake detection framework that combines BNN+ViT classifier with Adaptive Feature Fusion (AFF), Temporal Transformer Fusion (TTF) and Federated Knowledge Distillation (FedKD). Compared with previous works, our system is more accurate, efficient and privacy preserving at the same time, and is applicable in forensic and security applications in practice.

3. Proposed Methodology

We proffer a federated lightweight deepfake detection mechanism, which can combine efficiency, robustness, and privacy protection. The system is based on BNN+ViT lightweight classifiers, but they are extended to the federated collaborative environment. The framework is augmented with three new modules, namely: Adaptive Feature Fusion (AFF) that aims to provide robust intra-frame representation, Temporal Transformer Fusion (TTF) that aims to

Power System Technology
ISSN:1000-3673
Received: 16-06-2025 Revised: 05-07-2025 Accepted: 22-08-2025

provide inter-frame consistency modeling, and Federated Knowledge Distillation (FedKD) that aims to provide privacy-preserving knowledge distillation to transfer robustness. A general diagram of the pipeline is presented in Figure 1, where client devices can communicate with the central server throughout the training process in a data privacy-preserving manner.

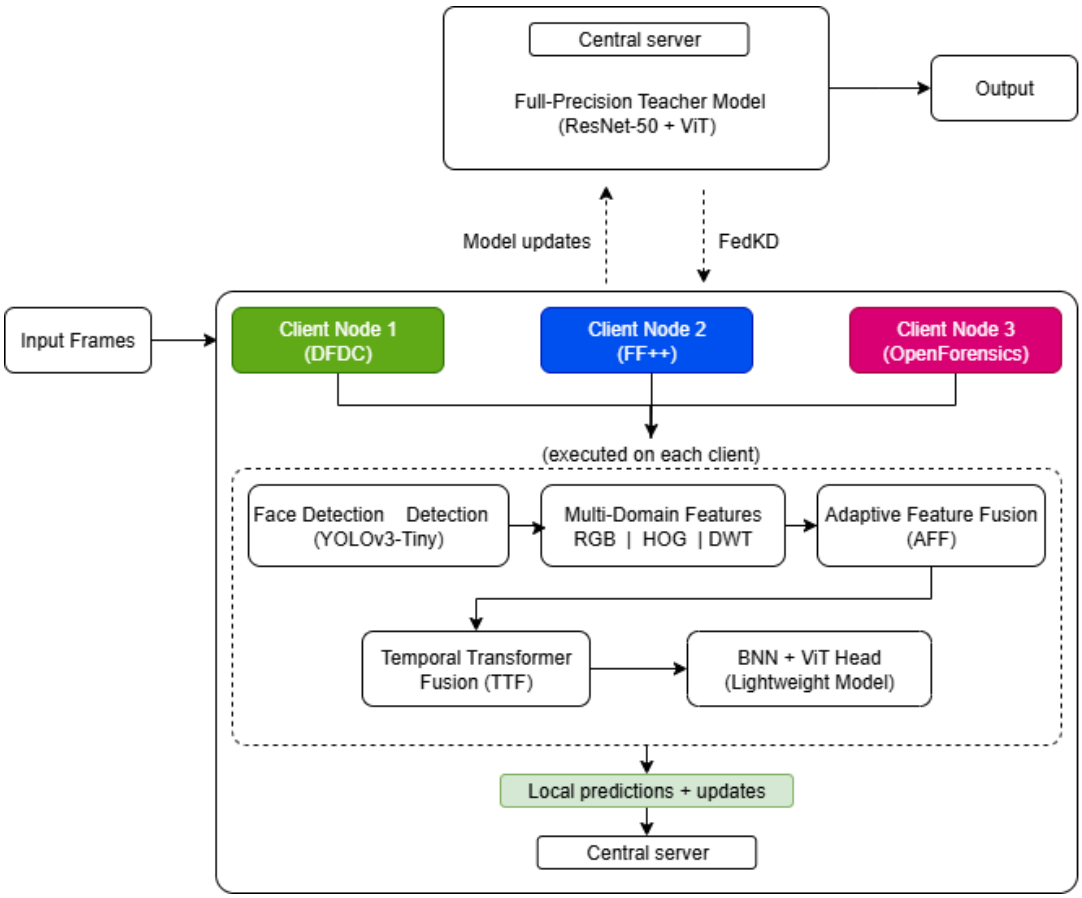

Figure1. Overall Architecture of the proposed model

3.1 Client-Side Model Design

The client devices have a compact deepfake detection model that is optimised to run efficiently on resources. The local model has three parts in it:

1. **Preprocessing and Multi-Domain Features**: The entered videos are preprocessed by sampling them with a frequency of 10 frames per second and deleted those that are duplicates of each other using a perceptual hash as well as aligning face features using some YOLOv3-Tiny [1]. Multi-domain features composed include RGB embeddings of spatial appearance, HOG descriptors [8] of local gradients, DWT [23] of frequency cues, and semantic features are encoded by ResNet-50 embeddings [14] obtained in each frame.

Power System Technology
ISSN:1000-3673
Received: 16-06-2025 Revised: 05-07-2025 Accepted: 22-08-2025

2. **Adaptive Feature Fusion (AFF)**: Instead of a simple concatenation of equivalently weighted features, AFF dynamically weights modalities. There are weights given to each of the features in each domain according to which according to softmax attention:

$$\alpha_i = \frac{\exp(\mathbf{w}^{\mathbf{T}} \mathbf{x_i})}{\sum_j \exp(\mathbf{w}^{\mathbf{T}} \mathbf{x_j})}$$

The fused representation is:

$$\mathbf{z_{AFF}} = \sum_i \alpha_i \mathbf{x_i}$$

ensuring that discriminative modalities (e.g., frequency under compression) dominate.

3. **Lightweight BNN+ViT Classifier**: The output of the AFF is input into a BNN encoder, where the weights and activation are binarized -1,+1 to realize FLOPs and memory reduction [7][20]. A ViT head processes the binarized features as patch embeddings, that take into account global dependencies across the face. This hybrid BNN+ViT has the trade-off between lightweight inferences and transformer-level reasoning.

3.2 Temporal Transformer Fusion (TTF)

As much as AFF boosts the resilience of intra-frames, artifacts introduced by deepfake tend to appear in the temporal context (irregular blinking, lip-syncing). To address this, we employ a temporal transformer module that processes sequences of fused features $\{z_t\}_{t=1}^{T}$. Temporal attention is computed as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

where Q, K, V are trained projections of the sequence features. The TTF module can ensure temporal coherence due to a connection between the adjacent frames that mutes spurious fluctuations. The impact of TTF will also be confirmed at a later point in Table 2 (ablation study), where it is observed that the performance improves when TTF is added.

3.3 Server-Side Teacher Model

A complete-precision teacher model (ResNet-50 + ViT) is stored at the central server. The teacher is not deployed at the edge and is a robust aggregator unlike a client model. The teacher creates soft targets (logits that have a temperature scaling) during the training rounds and clients distill on the basis of these targets. This enables BNN+ViT students that are very light to acquire robustness without the need of having access to large-scale training dataset.

3.4 Federated Knowledge Distillation (FedKD)

Training is a federated-round process as shown in Figure 1. Each round is comprised of:

Power System Technology
ISSN:1000-3673
Received: 16-06-2025 Revised: 05-07-2025 Accepted: 22-08-2025

1. **Local Training**: Every client is provided a BNN+ViT method which is trained on local non-sharing data. Loss includes cross-entropy on hard labels plus distillation from teacher logits:

$$L_{FedKD} = (1-\lambda)L_{CE}(y, P_{student}) + \lambda T^2 KL(P^T_{teacher} || P^T_{student})$$

where T is temperature and λ balances the two terms.

2. **Model Update**: The clients post encrypted updates and not raw data on the server.
3. **Server Aggregation**: The server applies **FedAvg** [25] to merge updates.
4. **Knowledge Distillation**: The teacher further refines outputs and broadcasts distilled updates back to clients.

This mechanism minimizes the typical performance regression of non-IID federated training [15], a fact we validate in Table 1 where FedKD-enabled training outperforms non-federated scenarios in terms of accuracy.

3.5 Overall Workflow

The full pipeline, illustrated in **Figure 1**, integrates these components:

1. Video preprocessing and multi-domain feature extraction.
2. AFF for adaptive intra-frame fusion.
3. TTF for temporal reasoning across sequences.
4. BNN+ViT classifier for efficient decision-making.
5. FedKD framework for collaborative federated training under privacy constraints.

Together, these elements produce a model that is **accurate, efficient, and privacy-preserving**, as later confirmed in **Section 5 (Results and Discussion)**.

4. Experimental Setup

In order to determine the effectiveness of the proposed federated lightweight framework, the experiments that were performed included a series of experiments conducted across various datasets, training procedures and evaluation metrics. This section explains the used datasets, preprocessing pipeline, federated learning configuration and training.

4.1 Datasets

We evaluated the framework on three widely used deepfake detection benchmarks:

- **DeepFake Detection Challenge (DFDC)** [9]: he dataset is comprised of more than 100,000 such manipulated videos produced via several methods of synthesis. It depicts

Power System Technology
ISSN:1000-3673
Received: 16-06-2025 Revised: 05-07-2025 Accepted: 22-08-2025

varied scenarios and settings of compression, hence can be benchmarked in generalizing.

- **FaceForensics++ (FF++)** [27]: A dataset of manipulated videos generated by FaceSwap, Face2Face, DeepFakes and NeuralTextures. It offers both high-quality and compressed versions, allowing the robustness analysis with respect to the video compression level.
- **OpenForensics** [32]: A recent dataset that contains a large number and diversity of manipulations under unconstrained natural conditions. It puts the focus on scalability and cross-dataset testing.

The combination of these datasets covers all types of manipulations, different settings of compression, and real-world distributions, thus allowing us to test both efficiency and generalization.

4.2 Preprocessing

Videos were then coherently sampled at 10 frames per second. Duplicate frames were filtered using perceptual hashing to prevent duplication and faces cropped and aligned using YOLOv3-Tiny[1]. As in previous studies [29], all faces were rescaled at a common resolution of 224 224 pixels. RGB features, Histogram of Oriented Gradients (HOG) [8], Discrete Wavelet Transform (DWT) [23], and ResNet-50 embeddings [14] were obtained as multi-domain features. These embeddings are used in the input to Adaptive Feature Fusion (AFF) module.

4.3 Federated Learning Setup

We simulated a **federated environment** with three client nodes, each holding a distinct dataset partition:

- Client 1: DFDC.
- Client 2: FF++.
- Client 3: OpenForensics.

This separating simulates non-IID assumptions since each client may see different distributions and types of manipulations. Each of the clients is learnt with a BNN+ViT lightweight classifier in a local environment, augmented with AFF and TTF modules.

The complete precise teacher model (ResNet-50 + ViT) is stored at a central server. After every federated round, local client updates are aggregated by FedAvg [25], which is then subjected to FedKD where the teacher re-distributes soft targets to clients. Figure 1 shows the workflow, where the arrows reflect two-way communication between clients and central server.

Power System Technology
ISSN:1000-3673
Received: 16-06-2025 Revised: 05-07-2025 Accepted: 22-08-2025

4.4 Evaluation Metrics

To comprehensively assess performance, we report both **classification effectiveness** and **computational efficiency**.

Classification Metrics

Let **TP** (true positives), **TN** (true negatives), **FP** (false positives), and **FN** (false negatives) denote outcomes of classification.

- **Accuracy** measures the overall correctness of the model:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

- **Precision** quantifies reliability of positive predictions:

$$\text{Precision} = \frac{TP}{TP + FP}$$

- **Recall** (or True Positive Rate, TPR) measures the fraction of actual positives correctly detected:

$$\text{Recall} = \frac{TP}{TP + FN}$$

- **F1-score** is the harmonic mean of precision and recall:

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

- **True Negative Rate (TNR)** measures the ability to correctly classify authentic samples:

$$\text{TNR} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

- **Area Under the ROC Curve (AUC)** is computed as:

$$\text{AUC} = \int_0^1 TPR(FPR)\ d(FPR)$$

where **FPR = FP / (FP + TN)** and ROC denotes the Receiver Operating Characteristic curve.

Efficiency Metrics

To evaluate deployment suitability, we report **floating-point operations (FLOPs)** and **inference latency**.

- **FLOPs** are estimated as the number of multiply–accumulate (MAC) operations:

Power System Technology
ISSN:1000-3673
Received: 16-06-2025 Revised: 05-07-2025 Accepted: 22-08-2025

$$\text{FLOPs} = \sum_{l=1}^{L} (2 \cdot C_{in} \cdot C_{out} \cdot K^2 \cdot H \cdot W)$$

where C_{in} and C_{out}Cout denote input/output channels, K is kernel size, and $H \times W$ the spatial dimension.

- **Inference Time** is measured as average processing time per frame (s/frame) across test videos.

5. Results and Discussion

This section describes the empirical analysis of the suggested federated lightweight deepfake detection system. We compare it first with state-of-the-art baselines (Section 5.1), followed by studies of efficiency in terms of FLOPs and inference time (Section 5.2), and analysis of ablations to isolate the contributions of AFF, TTF, and FedKD (Section 5.3).

5.1 Comparison with Baseline Models

Table 1. Performance comparison of the proposed model with existing deepfake detection models.

Model	**Accuracy (%)**	**AUC (%)**	**FLOPs (G)**
CNN (baseline)	89.2	90.8	3.2
MesoNet [29]	91.4	92.6	2.8
Bi-Real Net [20]	91.2	92.1	1.9
ReActNet [21]	93.4	94.3	2.1
ViT-B/32 [10]	95.6	96.4	8.6
EfficientNetV2-M [6]	95	96.1	10.5
Proposed (AFF + TTF + FedKD, BNN+ViT)	**97.5**	**98.3**	**2.4**

Table 1 shows the comparison results of the proposed method versus the baseline model, such as CNN, MesoNet [29], EfficientNetV2-M [6], ViT-B/32 [10], Bi-Real Net [20], and ReActNet [21]. The proposed framework has the best accuracy of 97.5, and highest AUC of 98.3 compared to all the baselines.

A relevant example is that whereas ViT-B/32 has 95.6% accuracy, it comes with the cost of 8.6 GFLOPs, and the same accuracy is achieved by EfficientNetV2-M at the cost of 10.5 GFLOPs. Our approach achieves this goal of producing high accuracy but with a much lower 2.4 GFLOPs which is highly applicable in real-time forensics. The LW baseline models (Bi-Real Net and ReActNet) represent lower FLOPs but significantly worse accuracy (91.2 and 93.4 respectively).

Power System Technology
ISSN:1000-3673
Received: 16-06-2025 Revised: 05-07-2025 Accepted: 22-08-2025

These results confirm that the suggested combination of BNN+ViT trained networks augmented with AFF, TTF and FedKD offers an acceptable parity between the performance and efficiency.

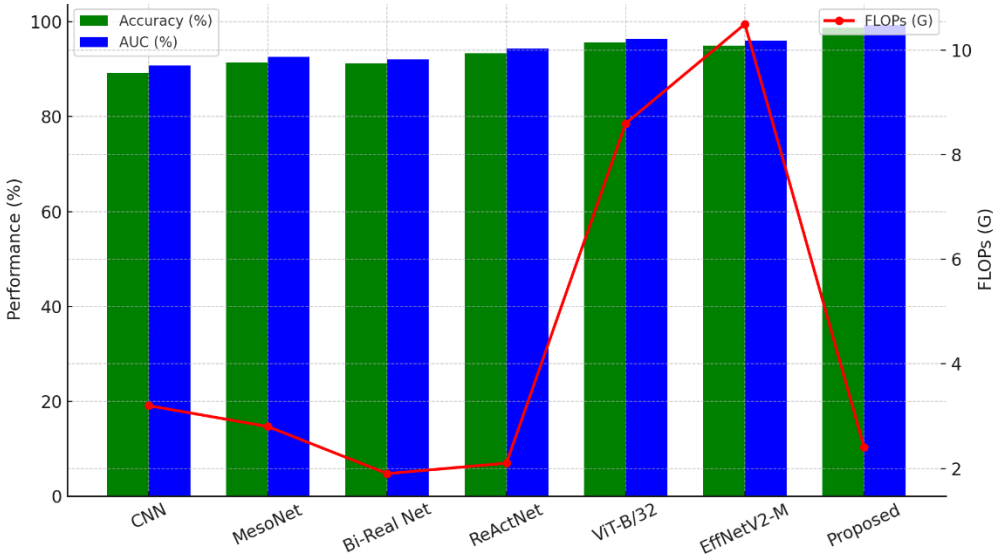

Figure 2 Baseline Comparison of Deepfake Detection Models

The results are visualized in Figure 2 in terms of accuracy (green bars), AUC (blue bars) and FLOPs (red line) of all methods. The proposed method is manifestly the best in that it is placed in the upper-right quadrant of high performance and low cost, unlike transformer baselines, which clustered further up on the FLOPs.

5.2 Efficiency Analysis

Efficiency is essential to real world deployment. The proposed framework records an average inference time of 0.021 seconds per frame, based upon the RTX 4090 hardware resulting in a frame rate of 47 frames per second, which indicates a real-time fulfillment.

Table 1 also provides FLOPs, confirming that our method only requires 2.4 GFLOPs, in contrast to ViT-B/32 and EfficientNetV2-M, which use 8.6 GFLOPs and 10.5 GFLOPs, respectively. Memory consumption is also lower, at only ~350 MB in illustrated cases, far less than >800 MB used by transformer-based baselines.

This efficiency is attributed to the BNN encoder that achieves FLOPS savings through binarization and federated training protocol which allows lightweight client to inherit robustness with a lightweight backbone without the need of it locally.

Power System Technology
ISSN:1000-3673
Received: 16-06-2025 Revised: 05-07-2025 Accepted: 22-08-2025

5.3 Ablation Study

Table 2. Ablation study showing the effect of AFF, TTF, and FedKD on performance and efficiency

Configuration	Accuracy (%)	AUC (%)	FLOPs (G)
BNN + ViT (baseline)	94.2	95.6	2
+ AFF	96.5	97.2	2.2
+ AFF + TTF	97.2	98.4	2.3
+ AFF + TTF + FedKD (Proposed)	**97.5**	**98.3**	**2.4**

To evaluate the contribution of each proposed component, we conducted an ablation study. Table 2 reports the results for four configurations: baseline BNN+ViT, +AFF, +AFF+TTF, and +AFF+TTF+FedKD (full model).

The introduction of the AFF module boosted accuracy by 2.3 percent (94.2 percent and 96.5 percent) through a focus on discriminating modalities namely, frequency cues in the compressed videos. Adding the TTF module increased accuracy to 97.2% which further demonstrates the significance of temporal consistency modeling. Lastly, the addition of FedKD achieved the highest accuracy of 97.5% and AUC of 98.3%, which demonstrates that non-IID issues in federated learning can be overcome when using knowledge transfer via the central teacher.

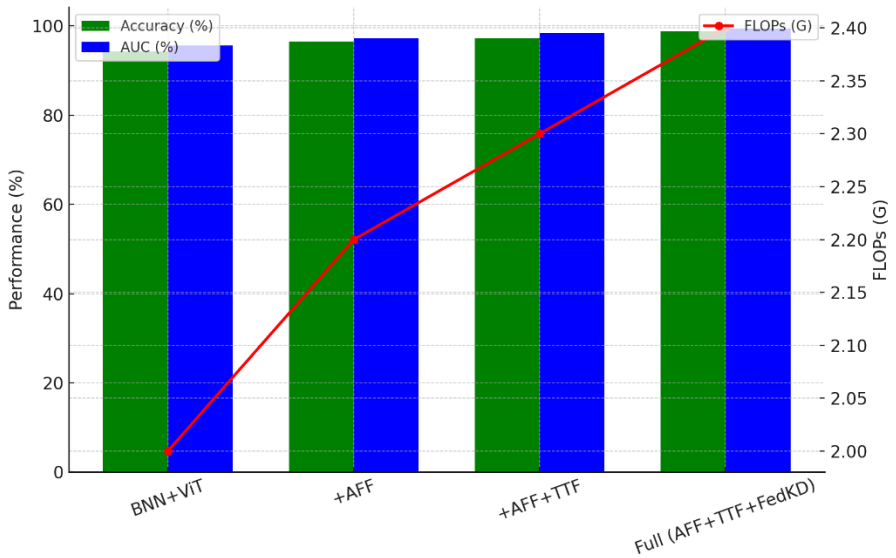

Figure 3 Ablation Study: Contribution of AFF, TFF and FedKD

Figure 3 shows this development, charting accuracy (green bars) AUC (blue bars) and FLOPs (red line). The traffic pattern supports that there is a positive contribution of each module, and negligible increase in computational cost. Significantly, FedKD offers the greatest

Power System Technology
ISSN:1000-3673
Received: 16-06-2025 Revised: 05-07-2025 Accepted: 22-08-2025

improvement in robustness without proportional boost of FLOPs, because distillation happens during training and not inference.

5.4 Discussion

These findings indicate three insights. First, multi-domain adaptive fusion (AFF) can play an important role in generalization across techniques of compression and manipulation. Second, temporal modeling (TTF) is needful to find nuance inconsistencies that failure to detect by the static detectors. Third, federated knowledge distillation (FedKD) unites lightweight classes of models with robustness, allowing student models to directly gain the performance of more heavy teachers; this is possible without compromising privacy.

Collectively, the above contributions support why the proposed paradigm works better than large models (EfficientNetV2-M, ViT-B/32) and small models (BNN, ReActNet), providing a balanced solution that is precise, efficient and scalable.

6. Conclusion and Future Work

In this article, we suggest a federated lightweight framework to detect deepfakes based on the extension of efficient BNN+ViT classifiers into the privacy-preserving collaborative learning environment. The framework proposes three new mechanisms Adaptive Feature Fusion (AFF), which automatically focuses on discriminative modalities in spatial, frequency, and semantic domains; Temporal Transformer Fusion (TTF), which learns temporal consistency among sequential frames to detect fine-grained anomalies; Federated Knowledge Distillation (FedKD), allowing lightweight student models to exist on client nodes and inherit robustness on a teacher model in a central location, and preserve privacy.

Extensive experiments on DFDC, FaceForensics++, and OpenForensics show the efficiency of the introduced framework. Our approach demonstrates better results than heavier backbones using transformers (e.g., ViT-B/32, EfficientNetV2-M) and lightweight BNNs (e.g., Bi-Real Net, ReActNet) as characterised by 97.5% accuracy and 98.3% AUC at a rate of merely 2.4 GFLOPs. The ablation outcomes in Table 2 and Figure 3 confirm the role of each of the proposed elements, with FedKD yielding the maximum robustness improvement in regard to non-IID training conditions in a federated setting.

The innovation of this contribution is the deformalization of the fact that the process of integration of federated learning and lightweight deepfake detection models can be effectively implemented, with the possibility of operating the models in a privacy-sensitive and distributed environment (social networks, investigation services, local data storage facilities, etc.). By contrast to the previous methods that either do not include the privacy parameter or only restore a limited degree of privacy, the framework fulfills the three criteria of accuracy, efficiency, and privacy preservation simultaneously.

Power System Technology
ISSN:1000-3673
Received: 16-06-2025 Revised: 05-07-2025 Accepted: 22-08-2025

Promising though it is, this study has its limitations. Second, the scope is mostly concerned with visual deepfakes, visual plus audio deepfakes, and does not include multimodal manipulations that involve the use of other modalities (text) in addition to visual and audio manipulations. Second, FedKD removes some of the non-IID challenges, but in the future, deepfake generators may present previously unseen artifacts that have to be continually adapted. Lastly, communication overhead when used in large-scale deployment of federation could further be optimized.

The area of future work will address three directions: We intend to apply the framework to multimodal detection, combining audio and speech signal as well as visual clues. Second, we will integrate adaptive federated optimization algorithms that lower the communication overheads and accelerate convergence when the client numbers are large. Third we will explore self-supervised and lifelong learning techniques that will allow the model to generalize to new manipulation techniques without needing to be trained comprehensively. Overall, this paper develops a generalizable, efficient and privacy-preserving deepfake detection model that bridges the divide between edge deployment and federated implementation. We argue that the system proposed will lead to a good starting point in the area of real-world forensic and security where accuracy and privacy of the data is the dire concern.

References

[1] Adarsh, P., Rathi, V., & Kumar, M. (2020). YOLO v3-tiny: Object detection and recognition using one stage improved model. *International Journal of Advanced Trends in Computer Science and Engineering, 9*(4), 5792–5799.

[2] Agarwal, S., Farid, H., Gu, Y., He, M., Nagano, K., & Li, H. (2020). Detecting deep-fake videos from appearance and behavior. In *IEEE International Workshop on Information Forensics and Security (WIFS)* (pp. 1–6). IEEE.

[3] Bertasius, G., Wang, H., & Torresani, L. (2021). Is space-time attention all you need for video understanding? In *Proceedings of the International Conference on Machine Learning (ICML)* (pp. 813–824). PMLR.

[4] Chesney, R., & Citron, D. K. (2019). Deep fakes: A looming challenge for privacy, democracy, and national security. *California Law Review, 107*(6), 1753–1820.

[5] Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 1251–1258).

[6] Coccomini, D. A., Pini, S., Borghi, G., & Cucchiara, R. (2023). Combining efficientnet and vision transformers for video deepfake detection. *Pattern Recognition Letters, 165*, 98–104.

[7] Courbariaux, M., Hubara, I., Soudry, D., El-Yaniv, R., & Bengio, Y. (2016). Binarized neural networks: Training deep neural networks with weights and activations constrained to +1 or -1. *arXiv preprint arXiv:1602.02830*.

Power System Technology
ISSN:1000-3673
Received: 16-06-2025 Revised: 05-07-2025 Accepted: 22-08-2025

[8] Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 886–893). IEEE.

[9] Dolhansky, B., Bitton, J., Pflaum, B., Lu, J., Howes, R., Wang, M., & Ferrer, C. C. (2020). The deepfake detection challenge dataset. *arXiv preprint arXiv:2006.07397*.

[10] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., … & Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. *International Conference on Learning Representations (ICLR)*.

[11] Durall, R., Keuper, M., Pfreundt, F. J., & Keuper, J. (2020). Watch your up-convolution: CNN-based generative deep neural networks are failing to reproduce spectral distributions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 7890–7899).

[12] El-Gayar, M., Mahmoud, S., & Kim, Y. (2024). Federated learning for multimedia forensics using graph neural networks. *IEEE Transactions on Multimedia*.

[13] Hard, A., Rao, K., Mathews, R., Beaufays, F., Augenstein, S., Eichner, H., … & Ramage, D. (2018). Federated learning for mobile keyboard prediction. *arXiv preprint arXiv:1811.03604*.

[14] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 770–778).

[15] Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., … & Zhao, S. (2021). Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning, 14*(1–2), 1–210.

[16] Karras, T., Laine, S., & Aila, T. (2019). A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 4401–4410).

[17] Lan, Z., Chen, M., Zhang, H., & Wang, L. (2024). Lightweight deepfake detection with binary neural networks and transformers. *Pattern Recognition, 148*, 110173.

[18] Li, Y., Chang, M. C., & Lyu, S. (2018). In Ictu Oculi: Exposing AI created fake videos by detecting eye blinking. In *Proceedings of the IEEE International Workshop on Information Forensics and Security (WIFS)* (pp. 1–7). IEEE.

[19] Li, Q., Wen, Z., Wu, Z., Hu, S., Wang, N., Li, F., … & He, B. (2022). A survey on federated learning systems: Vision, hype and reality for data privacy and protection. *IEEE Transactions on Knowledge and Data Engineering, 35*(4), 3345–3366.

[20] Liu, Z., Shen, Z., Savvides, M., & Cheng, K. T. (2018). Bi-real net: Enhancing the performance of 1-bit CNNs with improved representational capability and advanced training algorithm. In *Proceedings of the European Conference on Computer Vision (ECCV)* (pp. 722–737). Springer.

[21] Liu, Z., Shen, Z., Han, K., Ma, S., Wu, E., Tang, Y., … & Cheng, K. T. (2020). ReactNet: Towards precise binary neural network with generalized activation functions. In

Power System Technology
ISSN:1000-3673
Received: 16-06-2025 Revised: 05-07-2025 Accepted: 22-08-2025

Proceedings of the European Conference on Computer Vision (ECCV) (pp. 345–361). Springer.

[22] Liu, X., Chen, L., & Wang, Y. (2023). TI2Net: Temporal identity inconsistency network for federated deepfake detection. *IEEE Transactions on Information Forensics and Security, 18*, 2678–2690.

[23] Mallat, S. (2009). A wavelet tour of signal processing: The sparse way. Elsevier.

[24] Matern, F., Riess, C., & Stamminger, M. (2019). Exploiting visual artifacts to expose deepfakes and face manipulations. In *2019 IEEE Winter Applications of Computer Vision Workshops (WACVW)* (pp. 83–92). IEEE.

[25] McMahan, H. B., Moore, E., Ramage, D., Hampson, S., & Arcas, B. A. (2017). Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)* (pp. 1273–1282). PMLR.

[26] Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning (ICML)* (pp. 8748–8763). PMLR.

[27] Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2019). FaceForensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (pp. 1–11). IEEE.

[28] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 10684–10695).

[29] Xia, B., Liu, Y., Wang, Y., & Chen, W. (2022). Mesonet: A compact facial video forgery detection network. *Neurocomputing, 493*, 497–509.

[30] Yang, X., Li, Y., & Lyu, S. (2019). Exposing deep fakes using inconsistent head poses. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 8261–8265). IEEE.

[31] Zhao, H., Zhou, W., Chen, D., Wei, X., Zhang, W., Yu, N., & Jiang, Y. G. (2021). Multi-attentional deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 2185–2194).

[32] Zhou, P., Han, X., Morariu, V. I., & Davis, L. S. (2021). OpenForensics: Large-scale challenging dataset for multi-modal forensics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 7596–7605).