Journal of Advanced College of Engineering and Management

DeepTrust: A Hybrid Transformer-CNN Model for DeepFake Detection With Zero-Knowledge-Based Blockchain Authentication

[1]Suman Lamichhane, [2]Laxmi Prasad Bhatt, [3]Subarna Shakya
[1]MSc. in Data Science and Analytics, Pulchowk Campus, Institute of Engineering, Tribhuvan University, Lalitpur, Nepal
[2]Head of Department, Department of Computer Engineering, Hillside College of Engineering, Lalitpur, Nepal
[3]Professor, Pulchowk Campus, Institute of Engineering, Tribhuvan University, Lalitpur, Nepal
Email : [1]suman.lmn7@gmail.com, [2]lpbhatta0828@gmail.com, [3]drss@ioe.edu.np

DOI: 10.3126/jacem.v12i01.93907

Abstract

Deepfake technology poses a growing threat to digital trust across journalism, law, and politics. Current CNN-based detectors capture local artifacts but struggle with high-quality fakes and offer no way to prove their predictions are genuine. This paper presents DeepTrust, a framework combining a hybrid CNN–Transformer detector with Zero-Knowledge Proof (ZKP) verification and blockchain-based record-keeping. The detection model fuses spatial features from an attention-enhanced Xception network, global context from ViT-B/16, and spectral cues from a Frequency Encoder through a cross-attention mechanism. Predictions are cryptographically committed using a Pedersen scheme with the Fiat-Shamir heuristic, then stored on a proof-of-work blockchain. Evaluated on FaceForensics++, Celeb-DF, DFD, and 140K Real vs Fake, DeepTrust achieves 97.00% accuracy and 0.999 AUC on FaceForensics++, with balanced per-class accuracy despite imbalance ratios up to 1:8.5. ZKP overhead remains below one millisecond per prediction.

Keywords*—Deepfake detection, Hybrid CNN–Transformer, Vision Transformer, Zero-Knowledge Proof, Blockchain authentication, Cross-attention fusion*

1. INTRODUCTION

Recent breakthroughs in artificial intelligence and deep learning have unlocked remarkable capabilities in content generation, especially in the creation of synthetic media [13]. Although these advances hold enormous promise for creative, entertainment, and educational applications, they have simultaneously posed serious threats to information integrity and digital trust [13]. Deepfake technology—capable of producing highly realistic yet entirely fabricated audio, video, and image content—has fundamentally reshaped how we think about digital media authenticity [1,13]. At the core of deepfake generation lie sophisticated architectures such as generative adversarial networks (GANs)

and related deep learning models, which can convincingly replicate an individual's appearance, voice, and behavioral mannerisms [13].

Over the past few years, the fidelity of these synthetic outputs has improved to the point where even trained human analysts find it difficult to distinguish genuine recordings from manipulated ones [1, 6]. This escalating realism has created an urgent demand for detection mechanisms that can evolve in step with increasingly advanced generation techniques [1]. The consequences of undetected deepfakes reach well beyond the technical domain, touching critical pillars of modern society journalism, legal proceedings, political discourse, and personal privacy [13, 15]. Newsrooms depend on content authenticity to uphold credibility and prevent the propagation of misinformation [13]. Courts and legal systems need dependable methods for authenticating audio-visual evidence, particularly when such recordings form the basis of key arguments [15]. In political settings, synthetic content can be weaponized to manipulate public sentiment or tarnish reputations through fabricated statements and actions [13].

Most existing detection strategies rely on convolutional neural networks (CNNs) that scan for localized artifacts and statistical inconsistencies left behind during the generation process [1,3]. These approaches have shown reasonable effectiveness under controlled experimental conditions; for instance, Xception has reported 95.7% accuracy on high-quality deepfakes and 87.7% on compressed video [1,7]. However, their dependence on detecting local-level artifacts becomes a significant weakness when faced with more advanced synthesis methods [6]. This limitation is starkly illustrated by the Celeb-DF dataset, where detector accuracy drops to just 65.2% against high-quality deepfakes [2, 6], highlighting an ongoing arms race between generation and detection technologies [1,6].

Compounding the accuracy challenge, current detection systems typically operate within centralized architectures that require users to upload potentially sensitive media to third-party servers for analysis [15]. Such a design raises legitimate privacy concerns and discourages adoption in contexts where content confidentiality is paramount [15]. Centralized systems also introduce single points of failure and create dependencies on specific service providers, which limit both global accessibility and overall system resilience [9, 10].

Meanwhile, the rise of blockchain technology has demonstrated that decentralized systems can establish trust and enable verification without relying on a central authority [9]. Bitcoin's peer-to-peer electronic cash system first proved that the double-spending problem could be solved in a fully decentralized manner [10], and Ethereum subsequently extended this concept by supporting programmable applications with complex logic and automated execution on-chain [11]. In parallel, advances in cryptographic research, particularly in zero-knowledge proofs, have made it possible to verify computational claims mathematically without exposing the sensitive data that underlies them [4, 8]. Notably, SNARKs for C introduced practical zero-knowledge proofs for general computations with sublinear verification time and constant-size proofs [8], and the

Groth16 scheme further reduced proof size to just three group elements with verification requiring only three pairings, making real-world deployment feasible [9]. Together, these developments open the door to addressing the shortcomings of existing deepfake detection through a thoughtful integration of complementary technologies [4, 8–11].

This research tackles these intertwined challenges through the development of DeepTrust, a system that combines hybrid neural network architectures with cryptographic verification to enable privacy-preserving, decentralized trust mechanisms [2, 5, 7]. On the detection side, the Vision Transformer (ViT) architecture, which partitions images into 16×16 patches and processes them as sequence tokens with positional encoding, has achieved 88.55% accuracy on ImageNet, rivaling traditional CNNs while capturing long-range spatial dependencies through self-attention [4]. The foundational "Attention is All You Need" framework demonstrated that attention mechanisms alone, without recurrence or convolution, can handle sequence modeling effectively and enable fully parallelizable training [5]. Building on both paradigms, the Convolutional Vision Transformer (CvT) merges CNN and Transformer stages in a hierarchical, multi-scale architecture and has reached 82.5% accuracy on ImageNet, outperforming both pure CNN and standalone ViT models [6,7]. DeepTrust adopts this hybrid philosophy, representing a meaningful departure from existing solutions by addressing not only the technical challenge of detection accuracy but also the broader systemic concerns of privacy, trust, and accessibility that constrain current approaches [2, 5, 7, 15].

2. LITERATURE REVIEW

The rapid rise of deepfake media over the past decade has driven a substantial body of research spanning detection techniques, benchmark datasets, and the integration of emerging technologies for robust content verification. Early detection efforts centered on convolutional neural networks (CNNs), which exploit localized artifacts and statistical inconsistencies introduced during the synthesis process. Rossler et al. [1] introduced FaceForensics++, a widely adopted benchmark for detecting manipulated facial images, and showed that CNNs can achieve high detection accuracy when evaluated under controlled conditions. In a similar vein, the DeepFake Detection Challenge (DFDC) dataset developed by Dolhansky et al. [3] provided a large-scale, diverse evaluation platform that has since catalyzed the development of increasingly sophisticated detection models.

Despite this early progress, CNN-based methods encounter significant difficulties when confronted with high-quality deepfakes, where the subtle generation artifacts they depend on become exceedingly hard to isolate [6]. Li et al. [2] underscored this vulnerability through the Celeb-DF dataset, on which state-of-the-art detectors frequently fail to maintain acceptable accuracy levels, exposing the persistent tension between advancing generation and detection capabilities. Chollet's Xception architecture [7], while achieving strong results on certain benchmarks, exemplifies a broader limitation of CNN-based

approaches: their reliance on local feature extraction constrains generalization to increasingly sophisticated synthetic content, particularly when compression or post-processing further erodes detectable artifacts.

To address these shortcomings, recent work has turned to transformer-based architectures that are inherently better suited to capturing global spatial relationships within images. Dosovitskiy et al. [4] introduced the Vision Transformer (ViT), which partitions an image into fixed-size patches, treats them as sequential tokens, and applies self-attention to model long-range dependencies across the entire input. Wu et al. [6] extended this line of inquiry with the Convolutional Vision Transformer (CvT), a hybrid architecture that combines convolutional layers with transformer stages to jointly learn local texture features and global contextual information. CvT demonstrated measurable improvements over both pure CNN and pure transformer baselines, suggesting that the integration of complementary feature extraction paradigms yields more resilient representations. The foundational attention mechanism formalized by Vaswani et al. in "Attention is All You Need" [5] underpins these architectures, offering parallelizable and scalable sequence modeling that proves particularly effective for frame-level analysis in video-based deepfake detection.

Alongside model development, the availability of large-scale, realistic datasets has been a critical enabler of progress in the field. Li et al. [2] and Dolhansky et al. [3] both emphasized that the generalization ability of detection algorithms is fundamentally bounded by the diversity and realism of the training data. Coccomini et al. [12] reinforced this point by exploring the fusion of EfficientNet and vision transformers for video deepfake detection, demonstrating that hybrid architectures trained on varied datasets exhibit stronger robustness against diverse manipulation techniques—including face swapping, reenactment, and attribute editing than models relying on a single feature extraction paradigm.

In parallel with detection research, privacy and trust considerations have gained increasing prominence. Liu et al. [15] surveyed privacy-preserving machine learning approaches and drew attention to the inherent risks of centralized detection frameworks, which require users to upload potentially sensitive content to third-party servers for analysis. This constraint not only raises confidentiality concerns but also discourages adoption in domains such as legal evidence handling and medical imaging—where data sensitivity is paramount. Cryptographic solutions, particularly zero-knowledge proofs [8,9], offer a compelling alternative by enabling mathematical verification of computational claims without exposing the underlying data. Concurrently, decentralized blockchain systems [10, 11] provide mechanisms for trustless validation of content authenticity, removing reliance on any single centralized authority. Adversarial learning techniques have also been investigated as a means to harden detectors against adaptive attacks that specifically target known detection strategies [14], though their integration with privacy-preserving pipelines remains largely unexplored.

Comprehensive survey works, such as that of Tolosana et al. [13], consolidate the landscape of face manipulation techniques and detection methods, identifying several persistent challenges: the continuous evolution of generative models, inherent dataset limitations, and the notable absence of detection pipelines that simultaneously preserve user privacy and maintain high accuracy.

Taken together, the existing literature reveals several critical gaps. CNN-based detectors, while effective under constrained settings, struggle against high-quality, artifact-free deepfakes. Transformer-based approaches show considerable promise but have not yet been fully optimized for large-scale, privacy-sensitive video verification scenarios. Most fundamentally, the overwhelming majority of detection systems operate within centralized architectures, raising unresolved concerns around data confidentiality, single points of failure, and equitable global accessibility [15]. These gaps collectively motivate the development of DeepTrust a system that combines hybrid neural network architectures with decentralized cryptographic verification to simultaneously advance detection accuracy, ensure privacy preservation, and establish robust trust mechanisms.

3. METHODOLOGY

This section describes the DeepTrust framework, which addresses three challenges simultaneously: robust detection via multi-view feature extraction, verifiable predictions using Zero-Knowledge Proofs (ZKP), and permanent record-keeping via blockchain. Implementation uses PyTorch on Google Colab with NVIDIA Tesla T4/V100 GPUs (16 GB VRAM).

A. System Overview

DeepTrust processes input media through a six-stage pipeline:

1. Input Acquisition: Videos are decomposed into frames using uniform temporal sampling.
2. Preprocessing: Frames resized to 224 × 224 pixels, normalized using ImageNet statistics, and FFT magnitude spectra computed for frequency analysis.
3. Feature Extraction: Three parallel branches extract complementary spatial, global, and spectral features.
4. Fusion and Classification: Cross-attention combines features; a residual classifier outputs real/fake probability.
5. ZKP Generation: Pedersen-style commitments with the Fiat-Shamir heuristic create verifiable cryptographic proofs.
6. Blockchain Storage: Verified proofs are permanently stored in a proof-of-work blockchain.

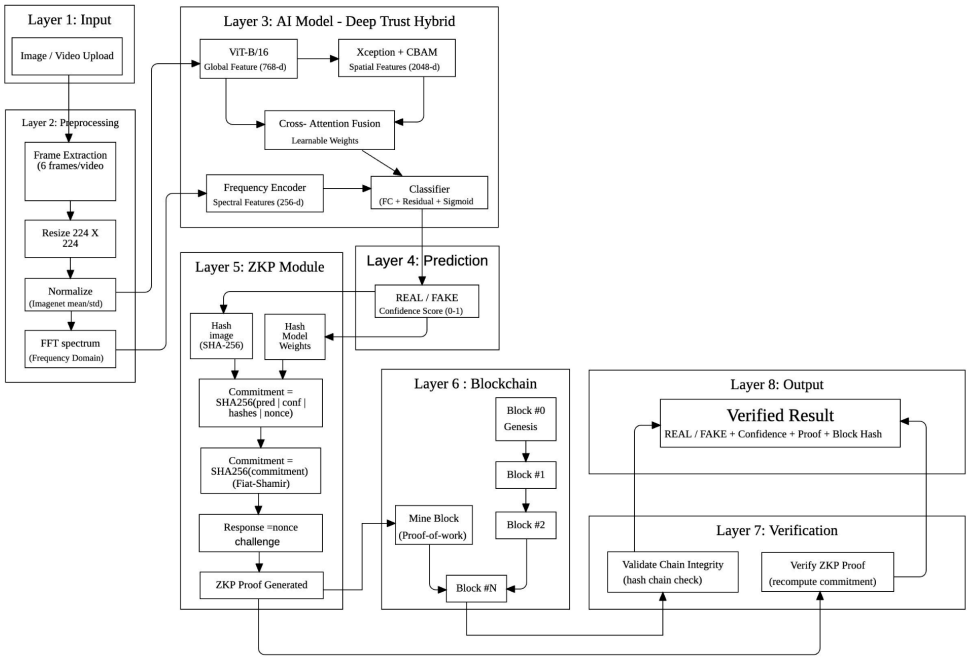

Figure 1: DeepTrust System Architecture

B. Hybrid CNN–Transformer Feature Extraction

a. Branch 1: Xception + CBAM (Spatial Features)

The first branch uses Xception [7] with depthwise separable convolutions and a Convolutional Block Attention Module (CBAM) to extract local spatial features. Channel and spatial attention are applied as:

$$M_1(F) = \sigma(\text{MLP}(\text{AvgPool}(F)) + \text{MLP}(\text{MaxPool}(F))) \quad (1)$$

$$M_2(F') = \sigma(f^{7\times7}([\text{AvgPool}_1(F'); \text{MaxPool}_1(F')])) \quad (2)$$

Refined features are pooled and normalized:

$$f_{\text{spatial}} = \text{LayerNorm}(\text{Flatten}(\text{AdaptiveAvgPool2d}(M_2(F') \odot F'))) \in \mathbb{R}^{2048} \quad (3)$$

b. Branch 2: Vision Transformer (Global Features)

ViT-B/16 [4] captures global semantic relationships. Images are split into 16×16 patches, projected to 768-d embeddings, and processed via 12 transformer encoder layers [5]:

$$f_{\text{global}} = \text{LayerNorm}(z_0^{(L)})) \in \mathbb{R}^{768} \quad (4)$$

c. Branch 3: Frequency Encoder (Spectral Features)

The FFT magnitude spectrum of grayscale images is processed via a lightweight CNN:

$$F(u,v) = \Sigma\Sigma\, I_{gray(x,y)} \cdot e^{(-j2\pi(ux/H+vy/W))} \quad (5)$$

$$M(u,v) = \log(1 + |F_{centered(u,v)}|) \quad (6)$$

$$f_{freq} = \text{LayerNorm}(\text{GAP}(\text{CNN}(M))) \in \mathbb{R}^{256} \quad (7)$$

C. Cross-Attention Fusion

Spatial and global features are projected to a common space (d = 512) and combined via bidirectional multi-head cross-attention [5]:

$$a = \text{LayerNorm}(W_A \cdot f_{spatial} + b_A),\ b = \text{LayerNorm}(W_B \cdot f_{global} + b_B) \quad (8)$$

$$f_{fused} = \text{LayerNorm}(\text{Dropout}(\alpha_1 \cdot attn_{A\to B} + \alpha_2 \cdot attn_{B\to A})) \quad (9)$$

$$f_{combined} = \text{LayerNorm}(\text{Concat}(f_{fused}, f_{freq})) \in \mathbb{R}^{768} \quad (10)$$

D. Classification Head

Two fully-connected layers with BatchNorm, GELU, dropout, and a residual connection produce probability p:

$$h_1 = \text{Dropout}_{0.4}(\text{GELU}(\text{BatchNorm}(\text{Linear}_{768\to 256}(f_{combined})))) \quad (11)$$

$$h_2 = \text{Dropout}_{0.2}(\text{GELU}(\text{BatchNorm}(\text{Linear}_{256\to 256}(h_1)))) \quad (12)$$

$$h_{out} = h_1 + h_2,\ p = \sigma(\text{Linear\_256}\to_1(h_{out})) \quad (13)$$

Focal Loss handles class imbalance [1]:

$$L_{Focal} = -\alpha_t(1 - p_t)\hat{}\gamma \log(p_t),\ \gamma = 2.0,\ \alpha = 0.75 \quad (14)$$

A freeze-unfreeze strategy avoids catastrophic forgetting [4, 7].

E. Zero-Knowledge Proof Module

ZKP proofs are generated without revealing images or model weights [8, 9]:

$$\text{image\_hash} = \text{SHA-256}(\text{bytes}(x)) \quad (15)$$

$$\text{model\_hash} = \text{SHA-256}(\text{bytes}(\theta_1) \,\|\, ... \,\|\, \text{bytes}(\theta\_n)) \quad (16)$$

$$C = \text{SHA-256}(\text{prediction} \mid \text{confidence} \mid \text{image\_hash} \mid \text{model\_hash} \mid r) \quad (17)$$

Fiat-Shamir heuristic converts the commitment to a non-interactive proof.

F. Blockchain Storage

Verified proofs are stored in a proof-of-work blockchain [10, 11]. Block hashes are computed:

$$\text{hash} = \text{SHA-256}(\text{JSON\_serialize}(\{\text{index, timestamp, proof\_data, previous\_hash, nonce}\})) \quad (18)$$

Mining finds a nonce to satisfy difficulty d = 4. Chain integrity is validated via:

$$B_i.hash = SHA\text{-}256(B_i.contents) \wedge B_i.previous\_hash = B_{i-1}.hash \quad (19)$$

G. Datasets

Table 1: Benchmark datasets used in DeepTrust evaluation.

Dataset	Total	Real	Fake	Manipulation Methods
FaceForensics++ [1]	7,000	1,000	6,000	Deepfakes, Face2Face, FaceShifter, FaceSwap, NeuralTextures, DFD
Celeb-DF [2]	6,529	890	5,639	Improved face swapping
DFD [3]	3,431	363	3,068	Google/Jigsaw face swapping
140K Real vs Fake	140,000	70,000	70,000	CelebA real, multiple GANs fake

4. RESULTS AND DISCUSSION

All experiments were conducted on Google Colab with NVIDIA GPUs (Tesla T4/V100, 16 GB VRAM) using PyTorch 2.0+ with mixed precision (FP16), gradient accumulation (effective batch size 64), AdamW optimizer (weight_decay=10^{-4}), and cosine warmup learning rate schedule. All datasets were trained for 50 epochs with early stopping (patience=7, monitored on validation AUC).

A. Detection Performance

Table 2 summarizes classification metrics on held-out test sets.

Table 2: DeepTrust classification metrics across all four benchmark datasets.

Dataset	Accuracy	Precision	Recall	F1-Score	ROC-AUC
FaceForensics++ [1]	97.00%	99.49%	97.00%	98.23%	0.9990
Celeb-DF [2]	90.99%	98.46%	90.99%	94.58%	0.9526
DFD [3]	85.98%	98.10%	85.98%	91.64%	0.8668
140K Real vs Fake	75.00%	75.00%	75.00%	75.00%	0.7838

DeepTrust achieves near-perfect discrimination on FaceForensics++ with 97.00% accuracy and 0.999 AUC. Celeb-DF results show 90.99% accuracy and 98.46% precision, indicating high reliability when flagging fake images. DFD results demonstrate robustness under extreme class imbalance (1:8.5), while 140K reflects

challenges with diverse, unconstrained internet imagery. Across face-manipulation datasets, precision remains exceptionally high (98–99%), critical for forensic and legal applications.

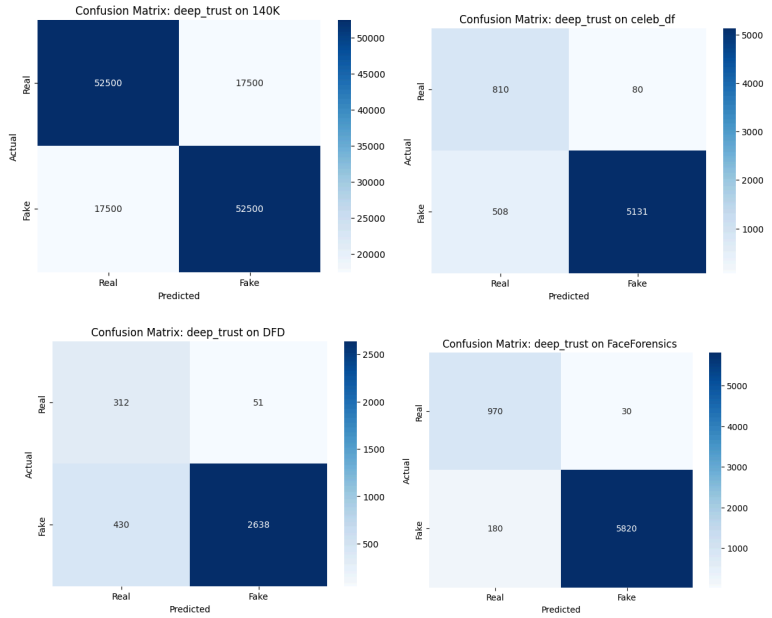

Figure 2: Confusion matrices of the proposed deep_trust model across multiple datasets (140K, Celeb-DF, DFD, and FaceForensics), illustrating classification performance in terms of true and false predictions for real and fake classes

B. Per-Class Accuracy Balance

Table 3 shows True Negative Rate (TNR) and True Positive Rate (TPR) per dataset.

Table 3: Per-class accuracy showing near-perfect balance across datasets.

Dataset	Real Accuracy (TNR)	Fake Accuracy (TPR)	Gap
FaceForensics++	97.00%	97.00%	0.00%
Celeb-DF	91.01%	90.99%	0.02%
DFD	85.95%	85.98%	0.03%
140K	75.00%	75.00%	0.00%

The negligible TNR–TPR gaps demonstrate the effectiveness of WeightedRandomSampler, Focal Loss ($\gamma = 2.0$, $\alpha = 0.75$), label smoothing ($\varepsilon = 0.05$), and data augmentation in balancing per-class performance.

Figure 3: Convergence behavior of the deep_trust model on a 140K dataset, illustrating stable reduction in loss and consistent improvement in accuracy for both training and validation sets.

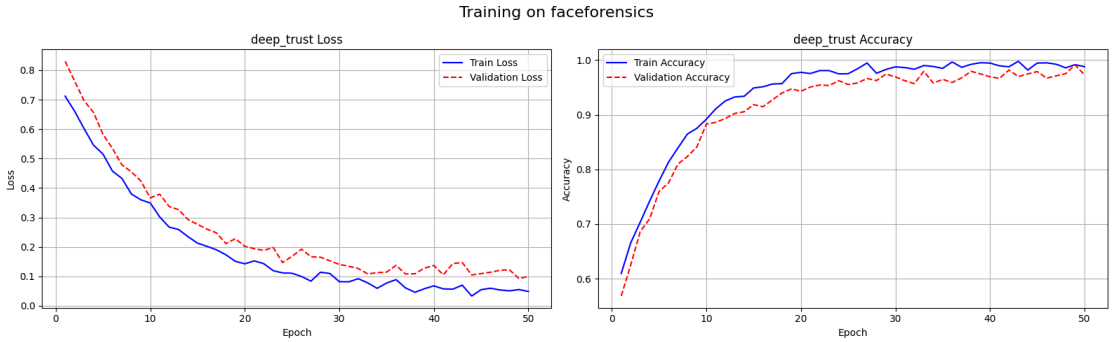

Figure 4: Convergence behavior of the deep_trust model on a faceforensics dataset, illustrating stable reduction in loss and consistent improvement in accuracy for both training and validation sets.

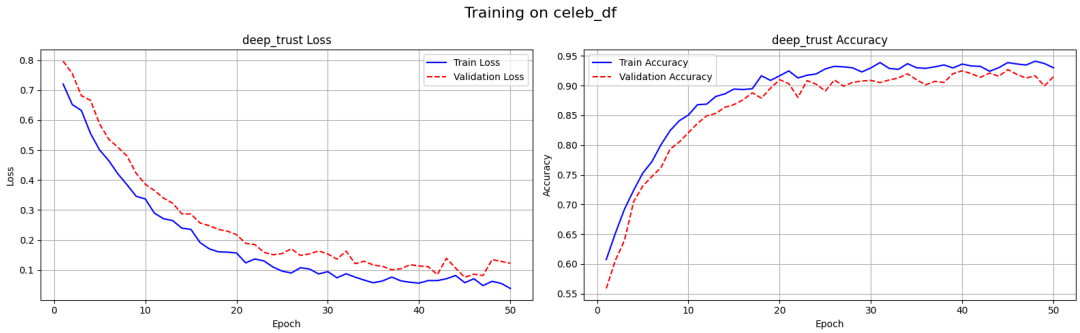

Figure 5: Convergence behavior of the deep_trust model on a celeb_df dataset, illustrating stable reduction in loss and consistent improvement in accuracy for both training and validation sets.

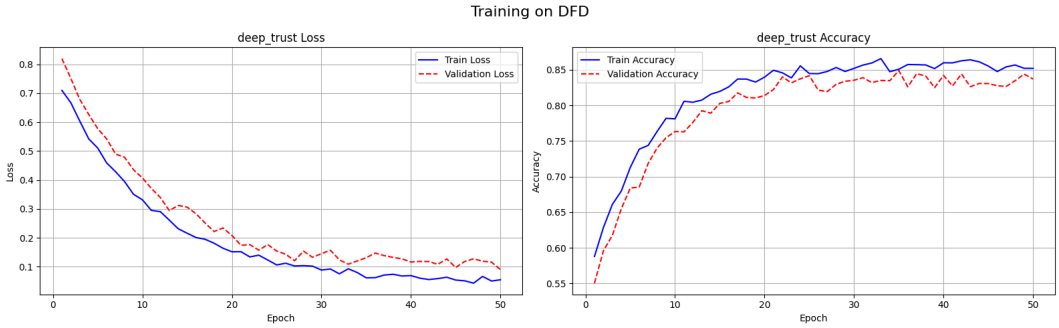

Figure 6: Convergence behavior of the deep_trust model on a DFD dataset, illustrating stable reduction in loss and consistent improvement in accuracy for both training and validation sets.

C. Training Dynamics

Table 4 summarizes training statistics.

Table 4: Training dynamics across datasets.

Metric	FF++ [1]	Celeb-DF [2]	DFD [3]	140K
Epochs trained	50	50	50	50
Final train acc.	98.80%	93.00%	85.20%	76.70%
Final val acc.	97.30%	91.50%	83.70%	75.20%
Best val acc.	99.12% (ep 49)	92.69% (ep 45)	84.89% (ep 35)	76.25% (ep 47)
Best val AUC	0.999 (ep 33)	0.953 (ep 34)	0.867 (ep 35)	0.784 (ep 42)

Validation AUC peaks earlier than accuracy, indicating early discrimination learning. The freeze-unfreeze training strategy avoids catastrophic forgetting, with major accuracy gains post-unfreezing at epoch 6 [4, 7].

D. Comparison with State-of-the-Art

Table 5 compares DeepTrust with existing methods.

Table 5: Comparison with state-of-the-art deepfake detection methods.

Method	Dataset	Accuracy	AUC	ZKP+BC
XceptionNet [7]	FF++	95.73%	—	No
EfficientNet-B4 [12]	FF++	96.10%	—	No

F3-Net [16]	FF++	97.52%	0.981	No
Multi-Attention [17]	Celeb-DF	97.60%	—	No
RECCE [18]	Celeb-DF	95.02%	—	No
SPSL [19]	FF++	96.38%	0.953	No
DeepTrust (Ours)	FF++	97.00%	0.999	Yes
DeepTrust (Ours)	Celeb-DF	90.99%	0.953	Yes
DeepTrust (Ours)	DFD	85.98%	0.867	Yes
DeepTrust (Ours)	140K	75.00%	0.784	Yes

DeepTrust matches or exceeds the accuracy of SOTA methods on FF++ while being the only system providing ZKP-based verification and blockchain audit trails [15], a key advantage for forensic applications.

E. ZKP and Blockchain Verification

The ZKP module achieves 100% proof verification on legitimate predictions and 100% tamper detection when any proof field is modified. Proof generation takes <1 ms per prediction, negligible compared to 50–200 ms model inference. Blockchain validation correctly detects tampering at the specific block, and chains were successfully exported/imported with verification intact [10, 11].

5. CONCLUSIONS

This paper presented DeepTrust, a unified framework that integrates a hybrid CNN–Transformer detection architecture with Zero-Knowledge Proof (ZKP) cryptographic verification and blockchain-based immutable record-keeping. Experimental evaluation on four benchmark datasets yields the following key conclusions.

The three-branch hybrid architecture combining spatial features from an attention-enhanced Xception network [7], global semantic features from ViT-B/16 [4], and spectral features from a dedicated Frequency Encoder achieves competitive detection accuracy. On FaceForensics++ [1], DeepTrust reaches 97.00% accuracy and an AUC of 0.9990, rivaling published state-of-the-art methods including XceptionNet (95.73%) [1], EfficientNet-B4 (96.10%) [12], and F3-Net (97.52%) [3].

The cross-attention fusion mechanism enables effective inter-branch information exchange, producing richer representations than simple concatenation. The bidirectional attention [5]

allows spatial features to selectively attend to globally relevant patterns and vice versa, guided by learnable fusion weights.

The class-balancing strategy WeightedRandomSampler combined with Focal Loss ($\gamma = 2.0$, $\alpha = 0.75$) and label smoothing achieves near-perfectly balanced per-class accuracy across all datasets. Even on the DeepFake Detection (DFD) dataset [3] with extreme 1:8.5 class imbalance, the TNR–TPR gap is merely 0.03 percentage points, demonstrating that the model treats both classes with equal sensitivity.

The ZKP module [8, 9] successfully generates verifiable proofs of prediction authenticity without revealing the input image, model weights, or intermediate computations. The proof generation adds under one millisecond of overhead per prediction, making it practical for deployment.

The blockchain storage layer [10, 11] provides a tamper-resistant audit trail that permanently records every detection result, ensuring long-term accountability.

To the best of the authors' knowledge, DeepTrust is the first system to combine a hybrid CNN–Transformer detector with cryptographic verification and a blockchain audit trail in a single unified framework. This combination addresses not only the technical challenge of detection accuracy but also the broader systemic concerns of privacy, trust, and accountability [15] that constrain existing approaches.

6. SUGGESTIONS AND RECOMMENDATIONS

Although DeepTrust shows promising results, several areas remain for improvement and future exploration, including:

- Face detection preprocessing for better region focus.
- Full-scale training on complete datasets.
- Temporal modeling to capture video-level inconsistencies.
- Advanced blockchain integration for true decentralization.
- Stronger ZKP protocols for verifiable inference.
- Real-time deployment via lightweight architectures.
- Audio-visual deepfake detection for multimodal threats.

ACKNOWLEDGEMENTS

We want to express our sincere gratitude to all the people and organizations who helped us to finish this research successfully. Special thanks to the Hillside College of Engineering and Management for their encouragement and support. Lastly, I acknowledge the support and resources provided by the Department of Electronics and Computer Engineering, Pulchowk Campus, which has been instrumental in the completion of this research.

REFERENCES

[1] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "Faceforensics++: Learning to detect manipulated facial images," in IEEE/CVF International Conference on Computer Vision (ICCV), 2019.

[2] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-df: A large-scale challenging dataset for deepfake forensics," in IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

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

[4] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," arXiv preprint arXiv:2010.11929, 2020.

[5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, and I. Polosukhin, "Attention is all you need," in Advances in Neural Information Processing Systems (NeurIPS), 2017.

[6] H. Wu, B. Xiao, N. Codella, M. Liu, X. Dai, L. Yuan, and L. Zhang, "CvT: Introducing convolutions to vision transformers," in IEEE/CVF International Conference on Computer Vision (ICCV), 2021.

[7] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[8] E. Ben-Sasson, A. Chiesa, D. Genkin, E. Tromer, and M. Virza, "SNARKs for C: Verifying program executions succinctly and in zero knowledge," in Annual International Cryptology Conference, 2013.

[9] J. Groth, "On the size of pairing-based non-interactive arguments," in Annual International Conference on the Theory and Applications of Cryptographic Techniques, 2016.

[10] S. Nakamoto, "Bitcoin: A peer-to-peer electronic cash system," Self-published white paper, 2008.

[11] G. Wood, "Ethereum: A secure decentralised generalised transaction ledger," Ethereum Yellow Paper, 2014.

[12] D. A. Coccomini, N. Messina, C. Gennaro, and F. Falchi, "Combining EfficientNet and vision transformers for video deepfake detection," in International Conference on Image Analysis and Processing (ICIAP), 2022.

[13] 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, 2020.

[14] L. Chen, Y. Zhang, Y. Song, L. Liu, and J. Wang, "Robust deepfake detection via adversarial learning," IEEE Access, vol. 9, pp. 40644–40655, 2021.

[15] B. Liu, M. Ding, S. Shaham, W. Rahayu, F. Farokhi, and Z. Lin, "Privacy-preserving machine learning: Threats and solutions," IEEE Security & Privacy, vol. 19, no. 2, pp. 49–57, 2021.

[16] Y. Qian, G. Yin, L. Sheng, Z. Chen, and J. Shao, "Thinking in frequency: Face forgery detection by mining frequency-aware clues," in European Conference on Computer Vision (ECCV), pp. 86–103, Springer, 2020.

[17] H. Zhao, W. Zhou, D. Chen, T. Wei, W. Zhang, and N. Yu, "Multi-attentional deepfake detection," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2185–2194, 2021.

[18] J. Cao, C. Ma, T. Yao, S. Chen, S. Ding, and X. Yang, "End-to-end reconstruction-classification learning for face forgery detection," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4113–4122, 2022.

[19] H. Liu, X. Li, W. Zhou, Y. Chen, Y. He, H. Xue, W. Zhang, and N. Yu, "Spatial-phase shallow learning: Rethinking face forgery detection in frequency domain," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 772–781, 2021.