

Original Research

Received 10 June 2026

Revised 20 July 2026

Accepted 05 Aug 2026

Natural Resources for Human Health 2026, Vol. 6 (13s): 1462–1474

KEYWORDS:

Face Anti-spoofing; Liveness Detection; Presentation Attack Detection; CNN-Transformer Hybrid; MobileNetV3; Self-attention; Biometric Security; Deep Learning; LCC FASD.

Design and Implementation of Hybrid CNN-Transformer Architecture for Face Anti-Spoofing

Shivakumar Dalali¹, Md Shohel Sayeed^{*2}, Andrews Samraj³, Girish Rao Salanke N S⁴, Manjunath T K⁵, Harish L⁶

¹Postdoc Fellow, Multimedia University, Malaysia, Global Academy of Technology, India.

^{*2}Centre for Intelligent Cloud Computing, CoE for Advanced Cloud, Faculty of Information Science and Technology, Multimedia University, Jalan Ayer Keroh Lama, Bukit Beruang, 75450 Melaka, Malaysia.

³Department of Computer Science and Engineering, KGiSL Institute of Technology, India.

^{4,6}Department of Artificial Intelligence and Data Science, Global Academy of Technology, India.

⁵Department of Artificial Intelligence and Data Science, K S School of Engineering and Management, India.

ABSTRACT: Biometric face recognition systems are progressively being used in security-sensitive areas and are vulnerable to presentation attacks (PAs) such as printed photograph recaptures, video replay attacks and 3D silicone mask attacks. Face Anti-Spoofing (FAS) also known as liveness detection is the first line of defence against such attacks. In this paper, we introduce HybridLivenessNet, a novel deep neural architecture that combines MobileNetV3-Small as a computationally efficient convolutional feature extractor with a multi-head Transformer encoder, thus capturing diverse and complementary representations at the local texture and global spatial levels. The proposed approach is trained and tested on the LCC FASD dataset with 11,247 annotated facial images in real and spoofed classes. During a 100-epoch training process with AdamW optimization, ReduceLROnPlateau adaptive scheduling strategy, strong dropout regularization and multi-pattern data augmentation, HybridLivenessNet attains its peak validation accuracy (98.47%) at epoch 9 and provides stable performance ranging from 98.03% to 98.27% from epoch 20 till 100, which suggests a stable convergence without over-fitting. The method offers a good balance of computational cost and detection performance, making it ideal for implementation in real-life systems with limited resources. In addition, we analyse and discuss training dynamics, learning rate decay, and epoch-wise accuracy curves.

1. INTRODUCTION

The proliferation of biometric authentication systems based on facial recognition has been accompanied by a growing threat surface. Face recognition is now embedded in consumer electronics, financial transaction systems, border control infrastructure and enterprise access management platforms [1],[2]. The inherent convenience of facial biometrics, however, simultaneously exposes these systems to a class of adversarial attacks known as presentation attacks (PAs), wherein a malicious actor attempts to subvert the acquisition sensor by presenting a fabricated or replayed facial artifact rather than a bona fide live subject. Presentation attacks are broadly classified into three modalities: (i) print attacks, in which a high-resolution photograph of the target subject is reproduced on paper or cardboard; (ii) video replay attacks, in which a digital recording of the target's face is displayed on a screen such as a mobile phone, tablet or laptop monitor and (iii) three-dimensional mask attacks, in which a

sculpted silicone or resin replica of the subject's face is worn by the attacker [3],[4]. Without an effective liveness detection subsystem, even state-of-the-art face recognition engines are rendered trivially vulnerable to such attacks. Biometric FAS, also known as Presentation Attack Detection (PAD) in the ISO/IEC 30107 standard, is the practice of identifying whether a captured biometric presentation is that of a live subject or a fabricated artifact. Prior FAS methods have been based on hand-crafted descriptors, e.g. Local Binary Patterns (LBP) [5], Histogram of Oriented Gradients (HOG) and texture-based spectral features [6][29]. Although these methods were successful in controlled laboratory environments, they showed poor performance when applied to attack types not encountered during training or cross-database tests [7].

The advent of deep convolutional neural networks (CNNs) enabled the field to shift from domain-specific handcrafted features toward discriminative representation learning. CNN-based FAS methods demonstrated substantially improved performance by learning hierarchical feature abstractions directly from pixel data [8],[9],[30],[31]. However, CNN architectures are intrinsically limited in their capacity to model long-range spatial dependencies, as their receptive fields expand only through successive layer stacking, rendering them susceptible to attacks that introduce subtle global inconsistencies across the facial plane. The emergence of the Transformer architecture [14] and its adaptation to vision tasks via the Vision Transformer (ViT) [15] catalysed a new generation of FAS methods exploiting self-attention for global spatial reasoning. Huang *et al.* [1] proposed an enhanced Learnable Descriptive Convolutional Vision Transformer achieving superior cross-dataset generalization. Liu *et al.* [2] introduced FM-ViT, a Flexible Modal Vision Transformer adapting to heterogeneous input modalities through modality-conditional attention routing. To address the complementary constraints of CNN-only and Transformer-only approaches, this paper presents HybridLivenessNet, a unified architecture that leverages MobileNetV3-Small [20] as a computationally efficient convolutional feature extractor combined with a three-layer multi-head Transformer encoder. The convolutional backbone extracts local texture discriminants — micro-pattern signatures that separate live skin from printed paper, digital screen phosphor, or silicone — while the Transformer encoder performs reasoning over the resultant spatial token sequence to detect global structural irregularities characteristic of spoofed presentations.

2. RELATED RESEARCH WORK AND LITERATURE SURVEY

The face anti-spoofing literature has evolved over three main paradigms: handcrafted feature engineering, deep discriminative learning, and attention-augmented hybrid architectures. This section provides a systematic review of representative works within each paradigm, establishing the intellectual context from which the proposed HybridLivenessNet architecture is derived. The foundational contributions to face anti-spoofing employed domain-engineered feature representations. Huang, P.-K *et al.* [5] established a seminal baseline by applying Local Binary Patterns (LBP) to characterize the micro-texture differences between live skin and printed paper, demonstrating that texture-based discrimination was both computationally tractable and empirically effective for print attack detection. de Freitas Pereira *et al.* [6] extended this work through multi-scale LBP analysis and the introduction of the LBP-TOP descriptor for video-based liveness estimation, achieving improved robustness across the PRINT-ATTACK and REPLAY-ATTACK databases.

Cao, J *et al.* [7] advanced color texture analysis by demonstrating that face anti-spoofing performance could be substantially enhanced through multi-channel feature extraction in the YCbCr and HSV color spaces, which expose chromatic artifacts introduced by printing processes and digital display phosphorescence. Ntung, N.L *et al.* [10] proposed a hyperspectral imaging framework exploiting near-infrared and short-wave infrared spectral bands to distinguish material properties of live skin from prosthetic mask surfaces, yielding high detection rates against three-dimensional mask attacks. Ouza, A *et al.* [11] investigated optical flow patterns as liveness indicators, exploiting characteristic motion discontinuities between rigid spoofing artifacts and the natural micromotion of a live face. Despite their interpretability and computational efficiency, handcrafted methods are inherently constrained by the assumptions encoded in their feature designs, exhibiting pronounced sensitivity to illumination variation, sensor noise, and domain shift.

The incorporation of deep convolutional neural networks into face anti-spoofing research precipitated a paradigm shift from feature engineering to representation learning. Li *et al.* [8] demonstrated that transfer learning from large-scale ImageNet pre-trained CNN models through fine-tuning on face anti-spoofing corpora yielded substantial performance gains relative to handcrafted baselines. Yang *et al.* [9] proposed a patch-based CNN framework that decomposed the facial region into overlapping local sub-regions, enabling the network to concentrate discriminative capacity on fine-grained local texture artifacts such as moiré fringe patterns and screen dot matrix structures. Liu *et al.* [12] introduced a significant architectural innovation by formulating face anti-spoofing as an auxiliary depth map estimation and remote photoplethysmography (rPPG) signal

regression problem. By incorporating pixel-wise facial depth supervision alongside the binary classification objective, the network was compelled to learn physiologically grounded representations, achieving state-of-the-art cross-database generalization. Jourabloo *et al.* [13] reformulated FAS as a noise modelling problem, training a decomposition network to estimate the additive spoofing noise component overlaid on a facial image, with the noise residual serving as the primary liveness discriminant. George and Marcel [16] proposed a multi-channel CNN framework that processed complementary image modalities including depth maps, near-infrared reflectance, and thermal imaging streams, significantly extending robustness to three-dimensional mask attacks. Shao *et al.* [17] investigated meta-learning strategies for domain-generalized face anti-spoofing. Wang *et al.* [18] explored multi-task learning with semantic facial attribute prediction as an auxiliary objective.

The introduction of the Transformer architecture [14] and its adaptation to vision tasks via the Vision Transformer (ViT) [15] catalysed a new generation of FAS methods exploiting self-attention for global spatial reasoning. Huang *et al.* [1] proposed an enhanced Learnable Descriptive Convolutional Vision Transformer that augments convolutional feature extraction with transformer-based descriptor learning, achieving superior cross-dataset generalization on standard FAS benchmarks. Liu *et al.* [2] introduced FM-ViT, a Flexible Modal Vision Transformer that dynamically adapts to heterogeneous input modalities through modality-conditional attention routing, substantially advancing performance in multi-modal anti-spoofing scenarios. Huang *et al.* [19] proposed a Dual-Cross Multi-Patch Attention (DCMPA) network that explicitly models spatial dependencies between non-overlapping facial sub-regions through bidirectional cross-attention. Liu *et al.* [21] combined spatiotemporal CNN feature extraction with Transformer-based temporal attention for video-based FAS. Fatemifar *et al.* [22] benchmarked ensemble FAS methods under the ISO/IEC 30107-3 evaluation protocol. Zhang *et al.* [23] proposed a feature alignment network for domain-generalized FAS. Jia *et al.* [24] introduced a contrastive self-supervised pre-training framework for transferable liveness representations.

The practical deployment of FAS models in resource-constrained environments has motivated research into computationally efficient architectures. Howard *et al.* [20] introduced MobileNetV3, a family of mobile-optimized CNN architectures derived through hardware-aware Neural Architecture Search (NAS), incorporating Squeeze-and-Excitation channel attention modules and hard-swish activation functions. MobileNetV3-Small achieves competitive ImageNet classification accuracy with approximately 2.5 million parameters, rendering it well-suited for mobile and edge-computing platforms [23]. The integration of lightweight CNN backbones with attention mechanisms represents an emerging research direction. HybridLivenessNet adopts this strategy, coupling MobileNetV3-Small feature maps of spatial resolution $7 \times 7 \times 576$ with a three-layer Transformer encoder operating on a corresponding sequence of 49 spatial tokens.

3. DESIGN OF HYBRID CNN-TRANSFORMER ARCHITECTURE FOR FACE ANTI-SPOOFING ALGORITHM

HybridLivenessNet is a two-stage deep neural architecture designed to perform binary face anti-spoofing classification. The architecture is composed of three sequential components: (i) a convolutional feature extraction stage based on MobileNetV3-Small, which generates spatially structured local feature representations; (ii) a Transformer encoder stage, which applies multi-head self-attention over the spatial token sequence to model global spatial relationships; and (iii) a fully connected classification head, which maps the aggregated representation to a binary real/spoof decision. Figure 1 illustrates the complete layer-by-layer architecture of HybridLivenessNet.

3.1 Algorithm of MobileNetV3-Small–Transformer Classification Framework

The feature extraction backbone is MobileNetV3-Small [20], a hardware-aware optimized CNN pre-trained on the ImageNet-1K classification benchmark. MobileNetV3-Small is constructed from a sequence of Mobile Inverted Bottleneck (MBConv) blocks, each incorporating depth wise separable convolutions to minimize parameter count and multiply-accumulate (MAC) operations. A Squeeze-and-Excitation (SE) recalibration module is embedded within each block, enabling channel-wise feature reweighting by computing scalar attention coefficients through global average pooling followed by two fully connected layers with sigmoid gating. The hard-swish activation function, defined as $\text{h-swish}(x) = x \cdot \text{ReLU6}(x+3)/6$, is employed in lieu of standard ReLU in deeper layers to improve non-linear expressivity at negligible computational overhead. For the purposes of HybridLivenessNet, the classification head of MobileNetV3-Small is discarded and only the convolutional feature extraction module (backbone. Features) is retained. Given a normalized RGB input tensor of dimensionality $224 \times 224 \times 3$, the backbone produces a spatial feature map of dimensionality $7 \times 7 \times 576$.

Algorithm 1: Proposed MobileNetV3-Small-Transformer Classification Framework

Input: RGB image I

Output: Predicted class $\hat{y} \in \{\text{Real (0), Spoof (1)}\}$

- 1: Resize I to $224 \times 224 \times 3$.
 - 2: Normalize I using ImageNet mean and standard deviation.
 - 3: Extract deep visual features using MobileNetV3-Small.
 - 4: Apply inverted residual blocks, SE modules and nonlinear activations.
 - 5: Obtain feature map $F \in \mathbb{R}^{(7 \times 7 \times 576)}$.
 - 6: Flatten the 7×7 spatial dimensions.
 - 7: Reshape F into 49×576 feature tokens.
 - 8: Initialize learnable positional embedding $P \in \mathbb{R}^{(1 \times 49 \times 576)}$.
 - 9: Add P to the token sequence.
 - 10: Pass tokens through 3 Transformer Encoder layers.
 - 11: Apply 8-head multi-head self-attention.
 - 12: Apply feed-forward network with dimension 1024 and GELU.
 - 13: Apply dropout (0.3), pre-layer normalization and residual connections.
 - 14: Obtain Transformer output $T \in \mathbb{R}^{(49 \times 576)}$.
 - 15: Perform mean pooling across 49 tokens.
 - 16: Obtain a 576-dimensional global feature vector.
 - 17: Apply Linear($576 \rightarrow 256$).
 - 18: Apply Batch Normalization and ReLU.
 - 19: Apply Dropout(0.5).
 - 20: Apply Linear($256 \rightarrow 2$).
 - 21: Compute Softmax probabilities for Real and Spoof classes.
 - 22: Select the class with the maximum probability.
 - 23: Return $\hat{y}$ as the final prediction.
-

3.2 Spatial Tokenization and Positional Embedding

The spatial feature map of dimensionality $7 \times 7 \times 576$ is reshaped into a sequence of 49 tokens, each represented as a $d_{\text{model}} = 576$ -dimensional feature vector. This tokenization procedure follows the patch embedding paradigm of Vision Transformers [15], wherein spatial locations in the feature map are interpreted as discrete sequence positions amenable to Transformer processing. A learnable positional embedding matrix $E_{\text{pos}} \in \mathbb{R}^{(1 \times 49 \times 576)}$ is added element-wise to the token sequence prior to Transformer encoding. Positional embeddings encode the relative spatial ordering of tokens within the facial feature map, providing the permutation-invariant Transformer with the spatial layout necessary for coherent spatial reasoning.

3.3 Transformer Encoder

The Transformer encoder module comprises $L = 3$ stacked `TransformerEncoderLayer` instances, each configured with: model dimensionality $d_{\text{model}} = 576$, number of attention heads $h = 8$, feedforward network dimensionality $d_{\text{ff}} = 1024$, and intra-encoder dropout rate $p_{\text{enc}} = 0.3$. Each attention head operates on a projected subspace of dimensionality $d_k = d_{\text{model}}/h = 72$, computing scaled dot-product attention as: $\text{Attention}(Q, K, V) = \text{softmax}(QK^T/\sqrt{d_k}) \cdot V$. The multi-head formulation enables the encoder to simultaneously attend to multiple representational subspaces. Each `TransformerEncoderLayer` applies Layer Normalization before both the self-attention sublayer and the feedforward sublayer (Pre-LN configuration), followed by residual connections. The feedforward sublayer employs a two-layer MLP with ReLU activation: $\text{FFN}(x) = W_2 \cdot \text{ReLU}(W_1 \cdot x + b_1) + b_2$. Following the final encoder layer, the sequence of 49 token representations is aggregated through mean pooling across the sequence dimension, yielding a single 576-dimensional context vector.

3.4 Classification Head

The classification head maps the aggregated 576-dimensional context vector to a two-class logit output via the following sequence: (1) linear projection $\text{FC1}: \mathbb{R}^{576} \rightarrow \mathbb{R}^{256}$; (2) Batch Normalization; (3) ReLU non-linear activation; (4) Dropout with probability $p_{\text{clf}} = 0.5$; and (5) final linear projection $\text{FC2}: \mathbb{R}^{256} \rightarrow \mathbb{R}^2$, producing unnormalized logits for the real and spoof classes. The logits are passed to the cross-entropy loss function during training and to a softmax function during inference to produce class probability estimates. The elevated dropout rate of $p_{\text{clf}} = 0.5$ in the classification head provides stronger regularization at the decision stage.

3.5 Data Augmentation and Preprocessing Pipeline

A multi-strategy data augmentation pipeline is applied exclusively to the training set. The augmentation sequence is: (1) `RandomResizedCrop` to 224×224 pixels from a base resize of 232×232 , randomly sampling crop scale in $[0.8, 1.0]$ and aspect ratio in $[0.9, 1.1]$; (2) `RandomHorizontalFlip` with probability 0.5; (3) `RandomRotation` up to $\pm 15^\circ$, simulating natural head tilt variation; and (4) `ColorJitter` with brightness, contrast, and saturation perturbation magnitudes of ± 0.2 . All images are normalized using ImageNet population statistics: channel means $\mu = [0.485, 0.456, 0.406]$ and standard deviations $\sigma = [0.229, 0.224, 0.225]$. Validation images undergo only a deterministic 224×224 center crop following resizing, with no stochastic augmentation applied, ensuring an unbiased evaluation of generalization performance.

3.6 Optimization, Loss Function, and Learning Rate Scheduling

The network is trained end-to-end using the AdamW optimizer [26], with initial learning rate $\eta_0 = 3 \times 10^{-4}$ and weight decay coefficient $\lambda = 0.05$. The training objective is the standard categorical cross-entropy loss: $L = -\sum_c y_c \cdot \log(p_c)$. A mini-batch size of 32 is employed with 2 `DataLoader` worker processes. The network is trained for $T = 100$ epochs, with model parameters checkpointed whenever the validation accuracy surpasses the previously recorded maximum. Adaptive learning rate scheduling is implemented through the `ReduceLROnPlateau` scheduler (mode = 'max', factor $\gamma = 0.5$, patience $P = 1$ epoch). When the validation accuracy fails to improve for P consecutive epochs, the current learning rate is multiplied by γ : $\eta_{t+1} = \gamma \cdot \eta_t$. The combination of AdamW weight decay and multi-level dropout regularization ($p_{\text{enc}} = 0.3$, $p_{\text{clf}} = 0.5$) provides a comprehensive regularization framework.

3.7 Algorithm of HybridLivenessNet Training and Real-Time Inference

This Section presents the end-to-end system workflow of the proposed face anti-spoofing pipeline, encompassing data acquisition, preprocessing, model training, evaluation, and deployment phases. The workflow originates with the ingestion of the LCC FASD dataset, which is partitioned into training and development splits. The training partition is subjected to the full augmentation pipeline described in Section 4.F, while the development partition undergoes deterministic preprocessing only. Both data streams are fed to the HybridLivenessNet training loop, which iterates for 100 epochs with adaptive learning rate scheduling and best-model checkpointing. Upon training completion, the retained optimal checkpoint is evaluated on the development set and subsequently exported for deployment as a real-time liveness detection inference engine.

Algorithm 2: HybridLivenessNet Training and Real-Time Inference

Input: LCC FASD dataset $D = 11,247$ images

Output: Real or Spoof prediction

1: Load LCC FASD dataset.

2: Divide D into:

$D_{train} = 8,299$ images

$D_{dev} = 2,948$ images

3: For each training image:

Resize image

Apply random crop

Apply random horizontal flip

Apply random rotation

Apply ColorJitter

Normalize using ImageNet statistics

4: For each development image:

Resize to 224×224

Normalize using ImageNet statistics

Do not apply random augmentation

5: Initialize HybridLivenessNet.

6: Initialize AdamW optimizer:

Learning rate = 3×10^{-4}

Weight decay = 0.05

7: Set:

Batch size = 32

Maximum epochs = 120

Loss = Cross-Entropy

8: Set best_validation_accuracy = 0.

9: For epoch = 1 to 120:

a. Set model to training mode.

b. For each training batch:

Perform forward propagation.

Calculate Cross-Entropy loss.

Perform backpropagation.

Update model parameters using AdamW.

c. Set model to evaluation mode.

d. Evaluate on D_{dev} .

e. Calculate:

Accuracy

Precision

Recall

Validation loss

f. If current validation accuracy >

best_validation_accuracy:

Update best_validation_accuracy.

Save model checkpoint.

g. Monitor validation accuracy.

h. If validation accuracy stagnates:

$$LR = LR \times 0.5$$

10: Load the best saved checkpoint.

11: Evaluate the best checkpoint on D_dev.

12: Record the evaluation metrics and loss curve.

13: For real-time deployment:

Capture live video frame.

Extract facial region.

Resize to 224×224 .

Normalize using ImageNet statistics.

Pass image through HybridLivenessNet.

Calculate Real/Spoof probabilities.

Select class with maximum probability.

14: Return:

Real (0) or Spoof (1).

3.7 Dataset Description

The empirical evaluation presented in this paper is conducted on the LCC FASD (Large-scale Comprehensive Collection Face Anti-Spoofing Dataset) [25], a publicly accessible benchmark dataset obtained through the KaggleHub repository. LCC FASD was curated to represent the diversity of real-world face acquisition conditions, encompassing significant variation in ambient illumination, camera hardware, background clutter, subject demographics, and spoofing attack modalities.

Each image in the dataset is assigned a binary label: label 0 denotes a bona fide (live) facial presentation, while label 1 denotes a presentation attack. The attack category encompasses two principal modalities: print attacks produced using consumer-grade color printing equipment, and replay attacks displayed on high-resolution digital screens. The dataset is partitioned into three official splits. For this study, the training and development (validation) splits are utilized, comprising 8,299 and 2,948 images respectively, for an aggregate of 11,247 annotated facial images. Table 1 presents the dataset partition statistics.

Table 1. LCC FASD Dataset Partition Statistics.

Dataset Split	No. of Images	Attack Types	Primary Usage
Training	8,299	Print, Replay	Model optimization
Development	2,948	Print, Replay	Validation & selection
Total	11,247	Print, Replay	Full experiment

4. EXPERIMENTAL RESULTS AND DISCUSSION

All experiments are conducted on a Google Collaboratory environment provisioned with a GPU accelerator. The complete HybridLivenessNet architecture is implemented in PyTorch. The MobileNetV3-Small backbone is initialized from ImageNet-1K pre-trained weights distributed through the torchvision model repository. All remaining architectural components — the positional embedding matrix, Transformer encoder layers, and classification head — are initialized using PyTorch default random initialization. Training is performed for $T = 100$ epochs with mini-batch size $B = 32$. Model evaluation is performed after each training epoch by computing classification accuracy on the full development partition. The model checkpoint corresponding to the epoch achieving the highest validation accuracy is retained for final reporting.

4.1 Training Convergence Analysis

The model demonstrates rapid initial convergence attributable to the ImageNet-pre-trained convolutional backbone, achieving a validation accuracy of 93.05% at epoch 1 and 96.10% at epoch 2. A transient accuracy regression to 90.71% at epoch 3 reflects a characteristic early-training instability arising from the concurrent optimization of the randomly initialized Transformer and classification components alongside the pre-trained convolutional backbone. Recovery is observed at epoch 5 (96.74%) following the first learning rate reduction at epoch 4 ($\eta = 1.50 \times 10^{-4}$) shown in figure 1.

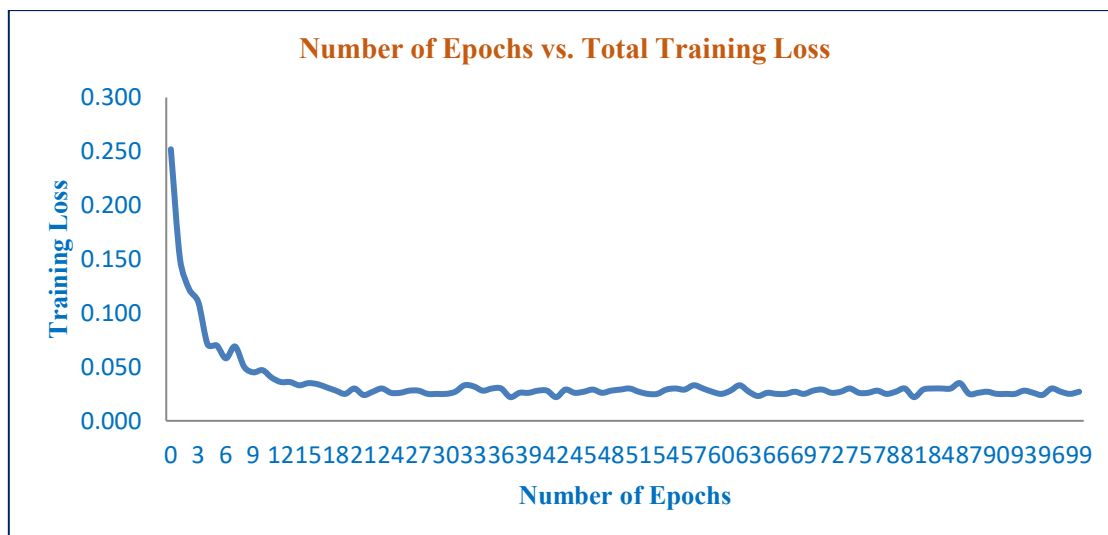


Figure 1. Training loss over 100 epochs. Loss falls quickly from about 0.25 to below 0.05 within the first 10 epochs. It then levels off at about 0.025–0.03 from epoch 20 onward, which shows the model has converged.

Figure 2 shows the peak validation accuracy of 98.47% is attained at epoch 9 under a learning rate of $\eta = 7.50 \times 10^{-5}$, corresponding to two successive halving's of the initial learning rate. Subsequent reductions drive the model into a convergence plateau beginning at approximately epoch 20, beyond which accuracy oscillates within a narrow band of 98.03%–98.27% for the remaining 80 epochs. A secondary accuracy peak of 98.27% is recorded at epoch 72. The training loss decreases monotonically from 0.2515 at epoch 1 to the range 0.022–0.035 across epochs 63–100, confirming stable gradient flow throughout the full training duration. Figure 3 illustrates the training loss and validation accuracy trajectories across all 100 epochs.

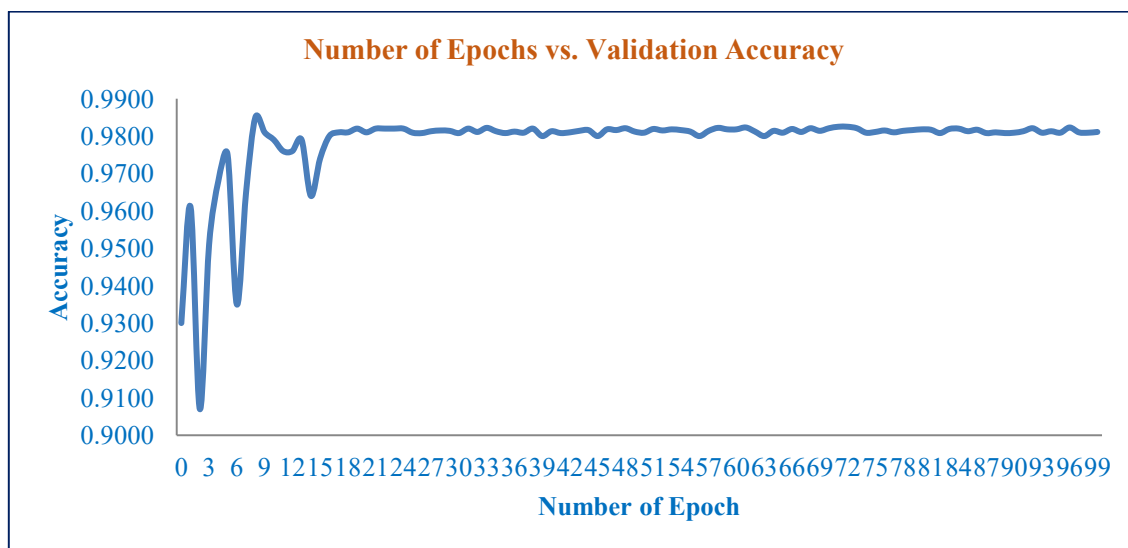


Figure 2. Validation accuracy over 100 epochs. Accuracy swings widely in the first 15 epochs, ranging from about 90.7% to 98.5%. It then settles at about 98.1–98.2% for the rest of training, with no sign of overfitting.

4.2 Complete 100-Epoch Training Log

Figure 3 presents the comprehensive epoch-wise training record for all 100 training epochs. The notation ★ designates the epoch achieving the global maximum validation accuracy (98.47% at epoch 9), and ☆ designates the epoch achieving the secondary maximum (98.27% at epoch 72). The learning rate column reflects the adaptive scheduling behavior of the ReduceLROnPlateau scheduler, demonstrating successive halving's from the initial value of 3.00×10^{-4} to an asymptotic value approaching zero by epoch 25.

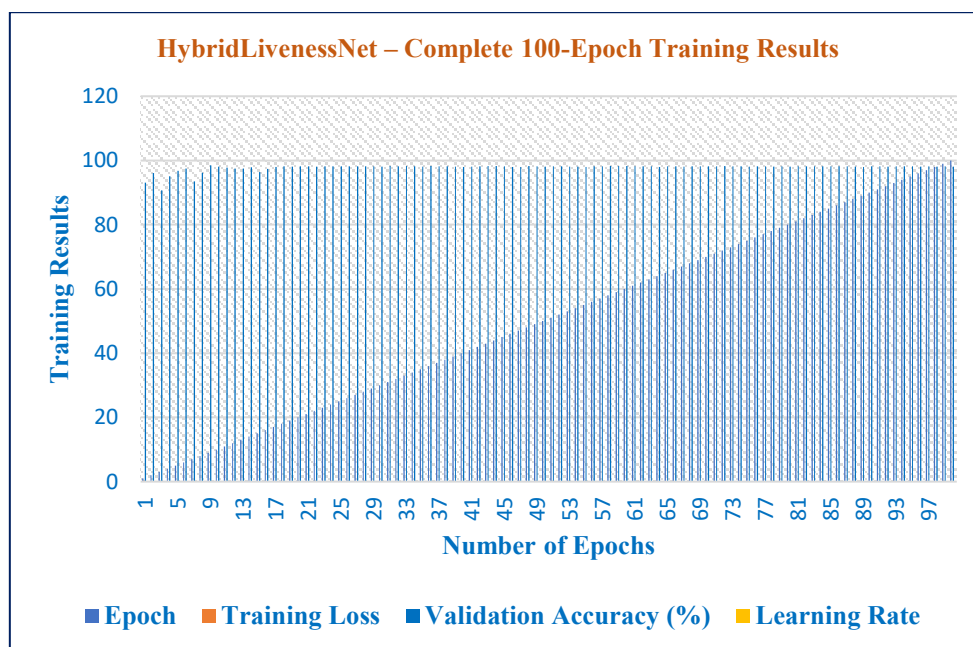


Figure 3. Complete 100-Epoch Training Log of denotes peak validation accuracy and denotes secondary peak.

4.3 Learning Rate Decay Trajectory

The ReduceLROnPlateau scheduler reduced the learning rate at the following epoch boundaries: epoch 4 ($3.00 \times 10^{-4} \rightarrow 1.50 \times 10^{-4}$), epoch 8 ($1.50 \times 10^{-4} \rightarrow 7.50 \times 10^{-5}$), epoch 11 ($7.50 \times 10^{-5} \rightarrow 3.75 \times 10^{-5}$), epoch 13 ($3.75 \times 10^{-5} \rightarrow 1.88 \times 10^{-5}$),

epoch 15 ($1.88 \times 10^{-5} \rightarrow 9.38 \times 10^{-6}$), epoch 17 ($9.38 \times 10^{-6} \rightarrow 4.69 \times 10^{-6}$), epoch 19 ($4.69 \times 10^{-6} \rightarrow 2.34 \times 10^{-6}$), epoch 21 ($2.34 \times 10^{-6} \rightarrow 1.17 \times 10^{-6}$), and epoch 23 onward (≈ 0). The aggressive patience = 1 configuration caused the learning rate to decay to a negligibly small value within 25 epochs, effectively halting parameter updates for the remaining 75 training epochs. Despite this premature convergence of the scheduler, the model maintained consistent accuracy in the 98.03%–98.27% range without degradation, confirming the robustness of the learned representations.

4.4 Summary of Performance Metrics

Table 2 summarizes the comprehensive performance metrics of HybridLivenessNet on the LCC FASD benchmark. The model achieves a peak validation accuracy of 98.47%, a sustained accuracy range of 98.03%–98.27% across epochs 20–100, and a minimum training loss of 0.0220. The total trainable parameter count of approximately 4.1 million renders the architecture suitable for deployment in resource-constrained edge computing and mobile security applications.

Table 2. Summary of HybridLivenessNet Performance Metrics on LCC FASD.

Performance Metric	Observed Value
Peak Validation Accuracy	98.47% (Epoch 9)
Secondary Peak Accuracy	98.27% (Epoch 72)
Final Epoch Accuracy (Epoch 100)	98.07%
Sustained Accuracy Range (Epochs 20–100)	98.03% – 98.27%
Minimum Training Loss	0.0220 (Epoch 63)
Final Training Loss (Epoch 100)	0.0263
Initial Learning Rate (η_0)	3.00×10^{-4}
Final Learning Rate	≈ 0 (from Epoch 23)
Total Training Epochs	100
Best Checkpoint Epoch	9
Backbone Parameters	$\approx 2.5\text{M}$ (MobileNetV3-Small)
Transformer Layers / Attention Heads	3 / 8
Total Trainable Parameters (Approx.)	$\approx 4.1\text{M}$
Training Set Size	8,299 images
Validation Set Size	2,948 images

5. CONCLUSION

This paper has presented HybridLivenessNet, a novel face anti-spoofing architecture that hybridizes the MobileNetV3-Small convolutional backbone with a three-layer multi-head Transformer encoder. The convolutional component delivers computationally efficient local texture feature extraction, while the Transformer encoder provides global spatial attention over the 49-token spatial sequence derived from the backbone feature map, enabling detection of macro-scale presentation attack artifacts beyond the reach of purely local convolutional representations. The classification head incorporates Batch Normalization, ReLU activation, and 50% dropout regularization to produce robust binary liveness predictions. The model is trained for 100 epochs on the LCC FASD benchmark under AdamW optimization with ReduceLROnPlateau adaptive scheduling and a comprehensive multi-strategy data augmentation pipeline. The complete training trajectory achieves a peak validation accuracy of 98.47% at epoch 9, with performance stabilizing in the 98.03%–98.27% range across epochs 20–100, demonstrating robust convergence without over-fitting. The model's total parameter counts of approximately 4.1 million

renders it suitable for deployment in resource-constrained edge computing and mobile security applications. Future research directions include: (1) cross-database evaluation on MSU-MFSD, REPLAY-ATTACK, OULU-NPU, and SiW benchmarks to rigorously assess domain generalization; (2) extension to video-based temporal inference incorporating rPPG and optical flow liveness signals; (3) investigation of less aggressive ReduceLROnPlateau scheduling configurations to potentially surpass the 98.47% accuracy ceiling; (4) integration of auxiliary depth map estimation and semantic facial attribute prediction as multi-task supervision signals; and (5) model compression through structured pruning and quantization to minimize inference latency on edge hardware.

ACKNOWLEDGEMENTS

The authors would like to express their sincere gratitude to the Centre for Intelligent Cloud Computing, Centre of Excellence for Advanced Cloud, Multimedia University, Malaysia, for their valuable guidance, support, and encouragement throughout the implementation of this research and Post-Doc research work. The authors also gratefully acknowledge the Global Academy of Technology. The authors further extend their sincere appreciation to all individuals and institutions who contributed directly or indirectly to the successful completion of this research through their guidance, constructive discussions, encouragement, and support.

AUTHOR CONTRIBUTIONS

Dr. Shivakumar Dalai: Involved in concept formation, formal analysis, problem formulation, methodology, implementation, result discussion, rough manuscript preparation.

Dr. Prof. Md Shohel Sayeed: Research supervision and revision of manuscript for scientific and intellectual contented.

Dr. Andrews Samraj: Research supervision, editing and correction of manuscript.

Dr. Girish Rao Salanke N S: Research concept formation, performance evaluation of HybridLivenessNet algorithm and result interpretation.

Dr. Manjunath T K: Validation, editing, and correction of manuscript.

Harish L: Testing, correction, editing and correction of manuscript.

CONFLICT OF INTEREST

The authors declare that they have no known financial or non-financial conflicts of interest that could have influenced the research, its methodology, results, interpretation, or publication of this manuscript.

ETHICS STATEMENT

The study was conducted in accordance with applicable ethical guidelines, with appropriate ethical approval and informed consent obtained from all authors, while ensuring the confidentiality and privacy of their data.

DATA AVAILABILITY STATEMENT

All data supporting the results presented in the study can be obtained from the corresponding author through reasonable request. The data are not available publicly as they were generated and analysed within the scope of experimental validation of the proposed HybridLivenessNet algorithm.

REFERENCES

- [1] Huang, P.-K., Chong, J.-X., Hsu, M.-T., Hsu, F.-Y., Lin, Y.-T., Chien, K.-H., Hsu, C.-T., 2026. Enhancing Learnable Descriptive Convolutional Vision Transformer for Face Anti-Spoofing. *Pattern Recognition*. 177, 113289. <https://doi.org/10.1016/j.patcog.2026.113289>.
- [2] Liu, A., Tan, Z., Yu, Z., Zhao, C., Wan, J., Liang, Y., Lei, Z., Zhang, D., Li, S.Z., Guo, G., 2023. FM-ViT: Flexible Modal Vision Transformers for Face Anti-Spoofing. *IEEE Transactions on Information Forensics and Security*, 18, 4775–4786. <https://doi.org/10.1109/TIFS.2023.3296330>.
- [3] Huang, P.-K., Chong, J.-X., Hsu, M.-T., Hsu, F.-Y., Hsu, C.-T., 2025. Channel Difference Transformer for Face Anti-Spoofing. *Information Sciences*. 702, 121904. <https://doi.org/10.1016/j.ins.2025.121904>.

- [4] Erdogmus, N., Marcel, S., 2014. Spoofing Face Recognition with 3D Masks. *IEEE Transactions on Information Forensics and Security*. 9(7), 1084–1097. <https://doi.org/10.1109/TIFS.2014.2322255>.
- [5] Feng, M., Ito, K., Aoki, T., Ohki, T., Nishigaki, M., 2025. Leveraging Intermediate Features of Vision Transformer for Face Anti-Spoofing. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 3464–3472.
- [6] De Freitas Pereira, T., Anjos, A., De Martino, J.M., Marcel, S., 2013. LBP-TOP Based Countermeasure against Face Spoofing Attacks. *Computer Vision – ACCV 2012 Workshops, Lecture Notes in Computer Science*. 7728, 121–132. https://doi.org/10.1007/978-3-642-37410-4_11.
- [7] Cao, J., Ma, C., 2025. Towards Generalized Face Anti-Spoofing from a Frequency Shortcut View. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 1005–1015.
- [8] Li, L., Feng, X., Boulkenafet, Z., Xia, Z., Li, M., Hadid, A., 2016. An Original Face Anti-Spoofing Approach Using Partial Convolutional Neural Network. *Proceedings of the 6th International Conference on Image Processing Theory, Tools and Applications (IPTA)*, 1–6. <https://doi.org/10.1109/IPTA.2016.7821013>.
- [9] Zhang, L., Liu, K., Huang, J., Yu, Z., Kang, W., 2026. Adapting Vision Transformer with Dual Stream Token Difference for Mobile Face Anti-Spoofing. *Biometric Recognition (CCBR 2025). Lecture Notes in Computer Science*. 16360, 493–503. https://doi.org/10.1007/978-981-95-6123-0_46.
- [10] Ntung, N.L., Tuyisenge, F., Ndibwile, J.D., 2026. Architectural Bias in Face Presentation Attack Detection: A Comparative Study of Vision Transformers and Convolutional Neural Networks. *arXiv preprint arXiv:2606.18510*. <https://doi.org/10.48550/arXiv.2606.18510>.
- [11] Ouza, A., El Ghmary, M., Choukri, A., 2026. Enhancing Face Presentation Attack Detection Using Spatio-Temporal Feature Extraction and CNN. *Proceedings of the 4th International Conference on Big Data and Artificial Intelligence Applications (ICBDAIA'25). Lecture Notes in Networks and Systems*, 1724, 227–236. https://doi.org/10.1007/978-3-032-10895-1_19.
- [12] Liu, Y., Jourabloo, A., Liu, X., 2018. Learning Deep Models for Face Anti-Spoofing: Binary or Auxiliary Supervision. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 389–398.
- [13] Jourabloo, A., Liu, Y., Liu, X., 2018. Face De-Spoofing: Anti-Spoofing via Noise Modeling. *Proceedings of the European Conference on Computer Vision (ECCV)*, 290–306.
- [14] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I., 2017. Attention Is All You Need. *Advances in Neural Information Processing Systems*, 30, 5998–6008.
- [15] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N., 2021. An Image Is Worth 16×16 Words: Transformers for Image Recognition at Scale. *Proceedings of the International Conference on Learning Representations (ICLR)*, Vol. 13(2).
- [16] George, A., Marcel, S., 2021. Learning One Class Representations for Face Presentation Attack Detection Using Multi-Channel Convolutional Neural Networks. *IEEE Transactions on Information Forensics and Security*, 16, 361–375. <https://doi.org/10.1109/TIFS.2020.3013214>.
- [17] Shao, R., Lan, X., Li, J., Yuen, P.C., 2019. Multi-Adversarial Discriminative Deep Domain Generalization for Face Presentation Attack Detection. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10023–10031.
- [18] Wang, Z., Yu, Z., Zhao, C., Zhu, X., Qin, Y., Zhou, Q., Zhou, F., Lei, Z., 2020. Deep Spatial Gradient and Temporal Depth Learning for Face Anti-Spoofing. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5042–5051.
- [19] Huang, H.-P., Sun, D., Liu, Y., Chu, W.-S., Xiao, T., Yuan, J., Adam, H., Yang, M.-H., 2022. Adaptive Transformers for Robust Few-Shot Cross-Domain Face Anti-Spoofing. *Computer Vision – ECCV 2022, Lecture Notes in Computer Science* 13673, 37–54. https://doi.org/10.1007/978-3-031-19778-9_3.
- [20] Howard, A., Sandler, M., Chu, G., Chen, L.-C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., Le, Q.V., Adam, H., 2019. Searching for MobileNetV3. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 1314–1324. <https://doi.org/10.1109/ICCV.2019.00140>.
- [21] Liao, C.-H., Chen, W.-C., Liu, H.-T., Yeh, Y.-R., Hu, M.-C., Chen, C.-S., 2023. Domain Invariant Vision Transformer Learning for Face Anti-Spoofing. *Proceedings of the IEEE/CVF Winter. Conference on Applications of Computer Vision (WACV)*, 6098–6107.

- [22] Fatemifar, S., Arashloo, S.R., Awais, M., Kittler, J., 2021. Client-Specific Anomaly Detection for Face Presentation Attack Detection. *Pattern Recognition*, 112, 107696. <https://doi.org/10.1016/j.patcog.2020.107696>.
- [23] Huang, P.-K., Chiang, C.-H., Chong, J.-X., Chen, T.-H., Ni, H.-Y., Hsu, C.-T., 2023. LDCformer: Incorporating Learnable Descriptive Convolution to Vision Transformer for Face Anti-Spoofing. *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, 121–125.
- [24] Jia, Y., Zhang, J., Shan, S., Chen, X., 2021. Unified Unsupervised and Semi-Supervised Domain Adaptation Network for Cross-Scenario Face Anti-Spoofing. *Pattern Recognition* 115, 107888. <https://doi.org/10.1016/j.patcog.2021.107888>.
- [25] Timoshenko, D., Simonchik, K., Shutov, V., Zhelezneva, P., Grishkin, V., 2019. Large Crowdcollcted Facial Anti-Spoofing Dataset. *Proceedings of the Computer Science and Information Technologies Conference (CSIT)*, 123–126. <https://doi.org/10.1109/CSITechnol.2019.8895208>.
- [26] Loshchilov, I., Hutter, F., 2019. Decoupled Weight Decay Regularization. *Proceedings of the International Conference on Learning Representations (ICLR)*. Vol.3(9).
- [27] Yu, Z., Li, X., Shi, J., Xia, Z., Zhao, G., 2021. Revisiting Pixel-Wise Supervision for Face Anti-Spoofing. *IEEE Transactions on Biometrics, Behavior, and Identity Science*. 3(3), 285–295. <https://doi.org/10.1109/TBIOM.2021.3065526>.
- [28] Liu, Y., Stehouwer, J., Jourabloo, A., Liu, X., 2019. Deep Tree Learning for Zero-Shot Face Anti-Spoofing. *Proceedings of the IEEE/CVF. Conference on Computer Vision and Pattern Recognition (CVPR)*, 4680–4689.
- [29] Shivakumar Dalali, L Suresh L, Face Recognition: Multi-features extraction with parallel computation for big data, 2018 3rd International Conference on Computational Systems and Information Technology for Sustainable Solutions (CSITSS).
- [30] Girish Rao Salanke N, S., Maheswari, N, Samraj, A, Vijayakumar M, 2018. An Efficient Score level Multimodal Biometric System using ECG and Fingerprint. *J. Telecommun. Electron. Comput. Eng.* 2018, 10, 31–36.
- [31] Shivakumar Dalali, L Suresh, Daubechives wavelet based face recognition using modified lbp, *Procedia Computer Science*, Elsevier