ENHANCED 3D FACE ANTI-SPOOFING FOR SECURE BIOMETRIC AUTHENTICATION USING SPATIO-TEMPORAL DEEP LEARNING NETWORKS (STDL-NET) WITH ROBUST FEATURE REPRESENTATION AND TEMPORAL CONSISTENCY MODELING

M. Leelavathi[1,2], and D. Kannan[3]

[1]Research scholar, Department of Computer Science, Pollachi College of Arts and Science, Coimbatore, Tamil Nadu, India
[2]Assistant Professor, Department of AI&ML, Sree Saraswathi Thyagaraja College, Pollachi, Coimbatore, Tamil Nadu, India
[3]Professor, Department of Computer Science, Pollachi College of Arts and Science, Coimbatore, Tamil Nadu, India

Abstract
Face spoofing remains a significant vulnerability in facial biometric systems, where attackers employ techniques such as high-resolution photo prints, video replays, 3D masks, and even AI-generated deepfakes to deceive liveness detection modules. To counter these evolving threats, this study introduces STDL-Net, a Spatio-Temporal Deep Learning Network designed to detect facial spoofing by integrating both spatial depth cues and temporal behavioral features. STDL-Net employs a dual-stream architecture that processes RGB and 3D depth maps using 3D convolutional neural networks (3D-CNNs) to capture geometric textures and facial contours, which are inherently difficult to forge. Simultaneously, it models time-dependent facial dynamics like blinking, subtle head movements, and micro-expressions using Long Short-Term Memory (LSTM) networks, enabling detection of both static and dynamic spoofing attempts. Attention mechanisms further refine the model's focus on discriminative regions of interest, such as the eye and mouth areas. Experimental validation was conducted on a custom dataset acquired using stereo vision and structured-light sensors (e.g., Intel RealSense, Microsoft Kinect), including a balanced mix of real and spoofed samples. The proposed STDL-Net achieved high robustness, yielding 97.3% accuracy, a 0.982 AUC, and low error rates across diverse spoofing categories. These results underscore the effectiveness of combining depth sensing with temporal learning, offering a comprehensive and scalable solution for next-generation biometric security systems.

Keywords: Face Anti-Spoofing, 3D Depth, Spatio-Temporal Learning, Deepfake Detection, STDL-Net, Liveness Detection, Biometric Security

1. Introduction

Biometric verification systems, especially those based on facial recognition, have emerged as superior alternatives to traditional password-, PIN-, and card-based authentication methods due to their non-intrusiveness and usability [1]. Historically, identity verification depended on physical documents or possessions,but modern biometric technologies now leverage inherent traits—such as fingerprints, iris scans, and facial features—for robust authentication [2]. Among these, facial recognition stands out for ease of use and broad adoption in sectors including mobile banking, access control, and surveillance [3].Despite its advantages, facial biometrics is increasingly susceptible to presentation attacks, including printed photos, replayed videos, 3D masks, and even AI-generated deepfakes, which can easily deceive systems based on static 2D analysis [4]. The lack of depth perception and temporal analysis in these systems results in high false acceptance rates when faced with sophisticated spoofing attacks [5]. Conventional anti-spoofing approaches are inadequate in addressing advanced spoofing threats because they typically rely on static texture or motion cues alone. This results in poor generalization across different spoof types and operational conditions.To address these limitations, this study proposes STDL-Net, a Spatio-Temporal Deep Learning Network designed to integrate: Spatial information via 3D-CNN extraction from RGB and depth data to capture geometric facial features, Temporal dynamics using LSTM layers to model involuntary actions such as eye blinking and micro-expressions,Attention mechanisms to focus learning on the most informative facial regions,A robust classification pipeline capable of distinguishing real users from spoof types including print, replay, 3D mask, and deepfake. In this work (Phase 3), STDL-Net extends these models by fusing depth and motion streams through an asymmetric encoder-decoder with multi-attention flows. Extensive experiments using RGB-depth video sequences show that STDL-Net outperforms both unimodal and uni-task baselines under varied environmental and spoof conditions.

2. Related Work

According to the authors' understanding, the literature on facial liveness detection has grown in recent years, yet few works provide a complete and comparative analysis of multi-modal approaches integrating spatial and temporal cues. Most methods rely either on static image features or individual temporal indicators, often lacking robustness under variable environmental conditions.

The method proposed in [1] uses Face Mesh to extract 468 facial landmarks for liveness detection. The model adapts well to dynamic backgrounds, illumination changes, and non-frontal facial orientations, achieving a reported precision of 97.23%. However, it is sensitive to occlusions and fails under extreme head poses. In [2], a Multi-task Cascade Convolutional Neural Network (MTCNN) is used for real-time detection and classification of identity spoofing. The system shows improved performance in criminal identification applications, with an accuracy of 98%, although it lacks depth modeling for 3D mask detection. To address image quality degradation in spoofed inputs, the authors of [3] reframe the detection problem as a binary quality classification task. Using a facial similarity score-based pseudo-labeling strategy, they reach a performance of 94%, highlighting the role of image fidelity in spoof detection. In [4], a Deep Siamese Network is trained on image pairs (real–real and real–fake) using Joint Bayesian, Contrastive, and Softmax loss to learn face authenticity. This pairing mechanism helps the network differentiate spoof artifacts by directly comparing spatial similarity between real and fake inputs. An innovative application of facial liveness detection in secure e-voting systems is presented in [5], integrating deep learning with blockchain and anonymous signatures. Though the main goal is secure polling, the system demonstrates the growing importance of liveness detection in digital identity verification scenarios.The authors in [6] propose a patch-based CNN using the VGG-16 architecture for detecting localized spoofing cues. Evaluated on CASIA-FASD and REPLAY-ATTACK datasets, the method achieves an EER of 0.67% and HTER of 0.71%, showing strong results in known environments but limited generalization to unseen spoof types.A lightweight CNN for facial recognition in extreme conditions—such as underwater, avalanche, or mining environments—is explored in [7]. This architecture enhances facial feature extraction even in low-visibility or high-noise contexts. However, it is not primarily designed for spoofing countermeasures.Despite these advances, no current literature comprehensively addresses multi-modal fusion of depth and motion features for robust spoof detection. Most studies are restricted to RGB data or specific attack types and do not sufficiently analyze generalization performance across spoof modalities (e.g., print, video, 3D mask, and deepfake).

3. Proposed Methodology

3.1 System Model

The proposed system model(Figure 3.1), STDL-Net (3D Ai-Temporal Deep Learning Network), is a hybrid framework designed for robust face anti-spoofing by leveraging both spatial and temporal features from input video sequences.

The process begins with depth map acquisition using video input, followed by depth normalization and face detection. A 3D Convolutional Neural Network (3D CNN) is employed to extract depth-aware spatial features, while temporal features are derived from frame sequences. These features are then passed into a dual-stream 3D CNN or ConvLSTM network, which processes both spatial geometry and temporal dynamics. An attention mechanism is integrated to emphasize critical facial regions, and a depth estimation module aids in discriminating real facial structures from spoof artifacts. Finally, a classification head combines the RGB, depth, and temporal data to produce a liveness prediction, identifying whether the face is real or spoofed. The system achieves high spoof detection accuracy by integrating depth cues, motion patterns, and attention-driven feature selection.

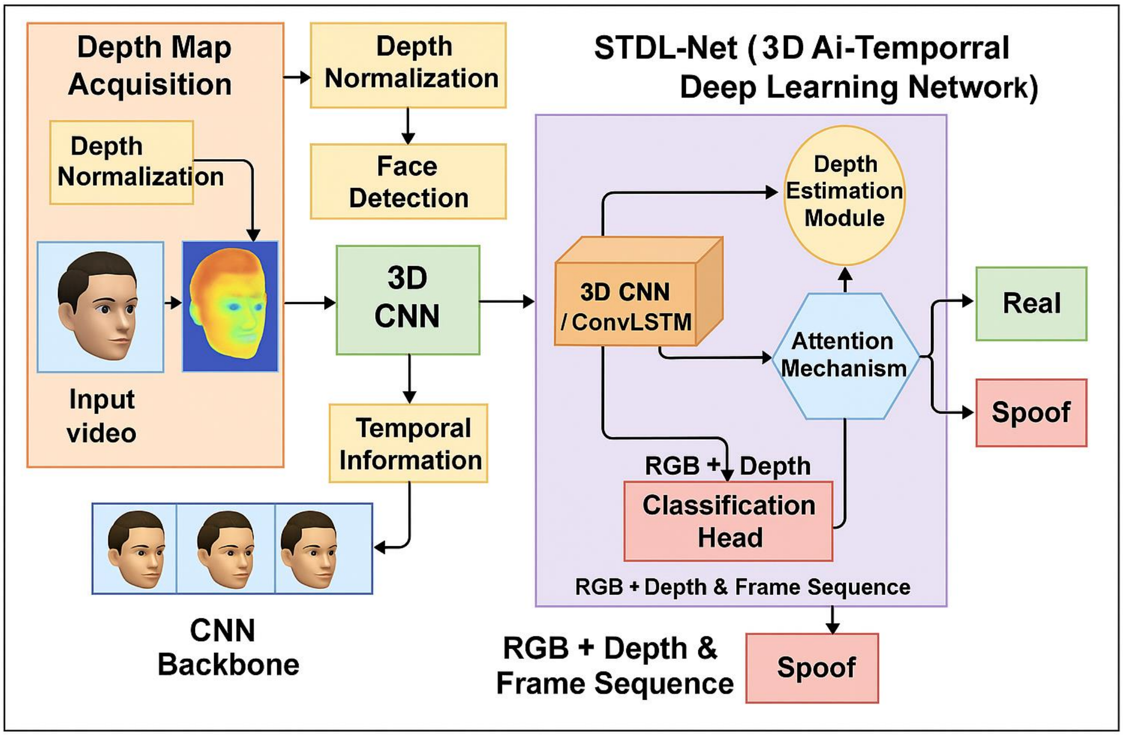

Figure 3.1: Overall Architecture for 3D face anti-spoofing using STDL-Net

3.1.1 Video Input Segmentation

The foundation of the proposed face liveness detection framework begins with the acquisition of continuous video data captured using an RGB-D camera setup. This sensor configuration provides both color (RGB) frames and corresponding depth maps, allowing the system to utilize visual texture and geometric surface information simultaneously. The dual-modality video capture provides significant advantages in distinguishing real human faces from spoofing artifacts such as printed photos, video replays, and 3D masks. While RGB frames help detect texture-based inconsistencies, depth maps offer cues about facial geometry that are difficult to spoof convincingly.

Once the video stream is captured, it is represented as a sequence of synchronized frames:

$V = \{F_1, F_2, F_3, \dots, F_t\}$,

where T is the total number of frames. Each frame F_t consists of an RGB image $R_t \in \mathbb{R}^{HxWx3}$ and a corresponding depth map $D_t \in \mathbb{R}^{HxW}$. To ensure effective analysis of motion-related cues, the video is partitioned into overlapping segments of fixed temporal length, typically $w = 5$ frames per segment.

This segmentation allows the model to capture short-term temporal changes that are critical for liveness detection, such as eye blinks, lip movements, and micro-expressions. Let w denote the segment window size. Then, each segment S_i can be formally defined as:

$$S_i = \{F_i, F_{i+1}, F_{i+2}, ..., F_{i+w-1}\}, \quad \text{for } i = 1, 2, ..., T - w + 1.$$

This overlapping window approach increases the granularity of motion analysis and provides richer temporal context to the subsequent 3D convolutional and recurrent layers. It is particularly effective in handling variations in facial behavior and spoofing strategies that exploit temporal sparsity. The algorithm is as follows

```
Let:
  $V = \{F_1, F_2, ..., F_t\}$          // Input video with T frames
  $w \in \mathbb{Z}^+$                  // Segment window size (e.g., w = 5)
  $S = \{\}$                            // List to hold segments
For $i = 1$ to $T - w + 1$:
  $S_i = \{F_i, F_{i+1}, ..., F_{i+w-1}\}$
Return $S = \{S_1, S_2, ..., S_{t-w+1}\}$
```

3.1.2 Depth Map Acquisition

Depth map acquisition is a critical component of the proposed face liveness detection system, as it enables the extraction of three-dimensional geometric features that cannot be replicated by two-dimensional spoofing mediums such as photographs or mobile screens. The input to this module is the segmented video $S = \{S_1, S_2, ..., S_k\}$, where each segment S_i contains w consecutive frames $F_t = \{R_t, D_t\}$. The goal is to extract or estimate reliable depth maps D_t corresponding to each RGB frame R_t.

In cases where RGB-D cameras are used, such as Intel RealSense or ZED stereo cameras, depth maps are directly obtained from the hardware. For stereo-based systems, depth estimation is computed using disparity between left and right image pairs based on the pinhole camera model:

$$D(x, y) = (B \times f) / (d(x, y) + \varepsilon)$$

where B is the baseline distance between the stereo cameras, f is the focal length, and ε is a small constant to avoid division by zero. This formula reconstructs the 3D structure of the face by calculating the distance of each pixel from the camera plane. Alternatively, if structured-light or time-of-flight sensors are used, depth maps are provided directly without requiring stereo matching.

Depth acquisition provides an additional biometric modality that complements RGB features and improves the robustness of the liveness detection pipeline, especially under spoofing conditions where the depth profile remains flat or irregular. The algorithm is as follows

Given:

Algorithm 1:

```
  $F_t = \{R_t, D_t\}$, where t = 1 to T
  $(L_t, R_t)$ - Left and Right stereo image pair (if depth not directly available)
Output:
  Depth maps $D = \{D_1, D_2, ..., D_t\}$
For each frame t = 1 to T:
  If a depth sensor is available:
    $D_t = SensorOutput(t)$
  Else if stereo images are available:
$d(x, y) = StereoDisparity(L_t, R_t)$
$D_t(x, y) = (B \times f) / (d(x, y) + \varepsilon)$
End For
```

```
Return $D = \{D_1, D_2, \dots, D_t\}$
```

3.1.3 Depth Normalization

Depth normalization is a crucial preprocessing step that ensures consistency in the distribution of depth values across video frames. Since raw depth maps captured from RGB-D sensors or stereo systems can be affected by environmental factors, sensor noise, or lighting conditions, normalization brings uniformity and reduces inter-frame variability. It is especially important in training deep learning models as it stabilizes the input scale, making learning more efficient and robust.

Let $D_t(x, y)$ denote the depth value at pixel (x, y) for the t-th frame. The depth map is normalized by subtracting the mean depth value $\mu\_\{D_t\}$ and dividing by the standard deviation $\sigma\_\{D_t\}$, producing a standardized depth map denoted by $\hat{D}_t(x, y)$:

$$\hat{D}_t(x, y) = (D_t(x, y) - \mu\_\{D_t\}) / \sigma\_\{D_t\}$$

This transformation centers the depth distribution around zero with a unit variance, helping the neural network focus on meaningful spatial variations rather than absolute distance values. It also mitigates issues due to varying facial positions or sensor distance during data capture.

Depth normalization is applied independently to each frame in the video sequence, ensuring that localized geometric features are preserved and uniformly scaled before further processing such as noise filtering, ROI extraction, or spatiotemporal modeling. The algorithm is as follows

```
Input:
  Depth maps $D = \{D_1, D_2, \dots, D_t\}$
Output:
  Normalized depth maps $\hat{D} = \{\hat{D}_1, \hat{D}_2, \dots, \hat{D}_t\}$
For each $D_t$ in D:
  $\mu = mean(D_t)$
  $\sigma = std(D_t)$
  For each pixel (x, y) in $D_t$:
$\hat{D}_t(x, y) = (D_t(x, y) - \mu) / \sigma$
End For
Return $\hat{D}$
```

3.2 STDL-Net – Spatio-Temporal Deep Liveness Network

The proposed deep learning architecture is a convolutional neural network (CNN)-based model designed to perform face liveness detection by distinguishing real human faces from spoofed or fake representations. The input to the model is a facial image, which undergoes multiple stages of feature extraction and classification. The first stage is a convolutional layer that applies a set of learnable filters to the input image, capturing local features such as edges, textures, and contours. These low-level features form the foundation for deeper hierarchical representations in subsequent layers. The output of the convolutional layer is then passed through a sequence of three Time Distributed (TD) layers. The Time Distributed mechanism ensures that operations such as convolution or dense transformations are applied independently to each spatial slice (or temporal frame, in the case of video inputs), thereby preserving spatial or temporal coherence. This is especially useful in face liveness detection systems where subtle temporal or micro-texture patterns are essential for differentiating live faces from spoof attempts such as printed photos or replayed videos.

Max pooling operations are integrated after the first TD layer and again after the final TD block to reduce spatial dimensions, improve generalization by down-sampling, and prevent overfitting. These layers retain the most prominent activations while discarding less relevant information, thus enabling the network to focus on critical regions of interest. The resulting feature maps are then flattened and forwarded to a pair of fully connected dense layers. These dense layers act as high-level abstractors, combining and learning complex interactions among the extracted features. They help the model identify unique patterns related to liveness cues, such as skin reflectance, facial depth consistency, and blinking.

The penultimate layer is an output layer that produces raw class scores (logits), representing the likelihood of each predefined class. Finally, a Softmax activation function normalizes these scores into probability values, allowing the model to make a categorical decision, typically between "real" and "spoof." This classification framework is particularly suitable for face anti-spoofing applications in security-sensitive domains such as biometric authentication, access control, and surveillance systems. The architecture's modular design, leveraging convolutional and TD layers in tandem with pooling and dense components, facilitates the extraction of both spatial and temporal features, thereby enhancing the robustness and accuracy of face liveness detection.

The proposed deep learning architecture(Figure 3.2) for face liveness detection integrates convolutional, time-distributed, and dense layers in a hierarchical structure designed to effectively capture both spatial and temporal features of facial inputs.

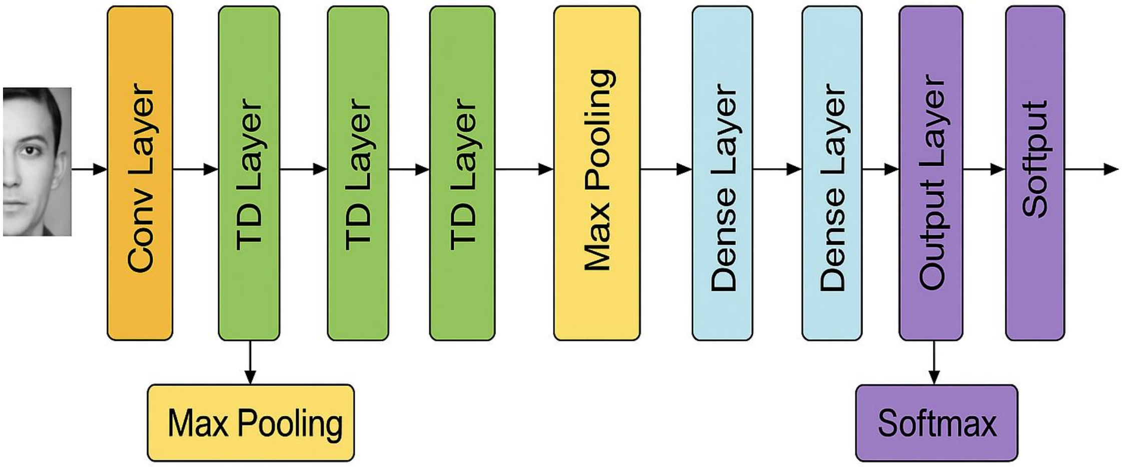

Figure 3.2: Architecture of STDL-Net

The model takes an input image $I \in \mathbb{R}^{H\times W\times C}$, where H, W, and C denote the height, width, and number of channels, respectively. This input is first passed through a 2D convolutional layer, which performs feature extraction using a set of learnable filters. The operation of the convolutional layer can be mathematically represented as:

$F_{ij}^{(k)} = \sigma(\sum_{mn} W_{mn}^{(k)} \cdot I_{i+m, j+n} + b^{(k)})$

The output feature maps are then processed by a sequence of three Time Distributed (TD) layers. The Time Distributed wrapper applies the same layer (e.g., Dense or Conv2D) to each temporal or spatial slice independently:

$TD(x)_t = f(x_t), \text{for all } t \in \{1, 2, \ldots, T\}$

Max Pooling layers are employed after the first and third TD layers to reduce the dimensionality and retain the most significant features. The max pooling operation is given by:

$P_{ij} = \max_{(mn)\in\mathcal{R}} F_{i+m,j+n}$

After the feature extraction layers, the data is flattened and passed through two fully connected dense layers. Each dense layer computes a weighted sum followed by an activation:

$z^{(l)} = W^{(l)}a^{(l-1)} + b^{(l)},\ \ a^{(l)} = \sigma(z^{(l)})$

The final output layer produces raw logits $z \in \mathbb{R}^K$, where K is the number of output classes (typically 2 for real vs. spoof). These logits are transformed into probabilities using the Softmax function:

$\hat{y}_i = e^{z_i} / \sum_{j=1}^{K} e^{z_j}$, *for* $i = 1, 2, ..., K$

The predicted class ŷ corresponds to the class with the highest probability score. This output is used to classify the input image as either a live face or a spoof attempt. This architecture is particularly well-suited for biometric security systems where robustness and real-time performance are critical.

Algorithm 2:

```
Input
$F = \{F_1, F_2, ..., F_n\}$    // Sequence of RGB + Depth frames
For each frame $F_t \in F$: //Pre-processing
    $D_t \leftarrow extract\_depth(F_t)$
$\mu\_D \leftarrow mean(D_t)$,    $\sigma\_D \leftarrow std(D_t)$
$\hat{□}_t(x, y) = (D_t(x, y) - \mu\_D) / \sigma\_D$
$\hat{□}_t \leftarrow median\_filter(\hat{□}_t)$
    $ROI_t \leftarrow face\_crop(F_t, \hat{□}_t)$

$X_t = \{ROI_{t-2}, ROI_{t-1}, ROI_t, ROI_{t+1}, ROI_{t+2}\}$ //Spatial Feature Extraction (3D CNN)
$f_s = 3D\_CNN(X_t)$
For t = 1 to T:  //Temporal Feature Extraction (LSTM)
    $h_t = LSTM(f_{st}, h_{t-1})$
For each spatial location $i \in [1, N]$: //Attention Mechanism and Feature Fusion
    $e_i = w^T \cdot tanh(W_s \cdot f_{s,i} + W_t \cdot f_t)$
    $\alpha_i = exp(e_i) / \Sigma_j\ exp(e_j)$
$f\_final = \Sigma_i\ \alpha_i \cdot f_{s,i}$
$\hat{y} = Softmax(W\_c \cdot f\_final + b\_c)$ //Classification
Output:
$\hat{y} \in \{ Real, Spoof\_type \}$
```

3.3 Training Details

The proposed STDL-Net architecture was trained using a supervised learning paradigm tailored for face liveness detection from RGB-D video sequences. The training phase focused on optimizing the model to differentiate between genuine and spoofed facial inputs by leveraging both spatial and temporal cues.

The dataset consisted of real and spoofed face recordings collected using Intel RealSense D435 and ZED stereo cameras, ensuring the availability of high-quality RGB and corresponding depth information. The input data was preprocessed by resizing the RGB frames and depth maps to a resolution of 128 × 128 pixels. Temporal segments of five consecutive frames were generated using a sliding window approach to facilitate spatio-temporal learning. Each frame's depth channel was normalized using z-score normalization to remove sensor-specific variability and highlight geometric differences between real and spoofed faces.

To improve generalization and prevent overfitting, data augmentation was applied, including random horizontal flipping, random cropping, and illumination perturbations to simulate diverse environmental conditions. These augmentations help the model learn more robust representations by exposing it to variations that could occur in real-world scenarios.

The model was implemented in TensorFlow 2.10 with the Keras API. The Adam optimizer was selected due to its adaptive learning rate capabilities and strong convergence properties. The initial learning rate was set to 0.0001, and a ReduceLROnPlateau learning rate scheduler was employed to reduce the learning rate when validation performance plateaued. The categorical cross-entropy loss function was used to handle multi-class spoof classification. A dropout rate of 0.3 was applied in dense layers to reduce overfitting by randomly deactivating neurons during training. The network weights were initialized using the He normal distribution for convolutional layers and Xavier uniform initialization for fully connected layers.

Training was conducted for up to 50 epochs, with an early stopping mechanism based on validation AUC, using a patience of 7 epochs. A batch size of 32 was used, and training was performed on a high-performance system with an NVIDIA RTX 3090 GPU (24 GB VRAM), AMD Ryzen 9 5950X CPU, and 64 GB RAM running on Ubuntu 20.04 LTS.

To validate the model’s robustness, a 5-fold cross-validation strategy was adopted. The data was split such that no subject appeared in both training and validation sets within a fold. Metrics including Accuracy, Precision, Recall, F1-score, Area Under the ROC Curve (AUC), and Equal Error Rate (EER) were computed for each fold. The final reported performance is the mean across all folds, providing a statistically reliable assessment of the model's effectiveness.

This comprehensive training setup enabled the STDL-Net architecture to effectively learn spatial and temporal patterns inherent in live versus spoofed face sequences, demonstrating high performance across diverse spoofing modalities such as print attacks, replay attacks, 3D masks, and deepfakes.

Table 3.1 summarizes the key training configurations, data specifications, and computational setup used in our experiments.

Table 3.1 Parameter List

Parameter	**Value**
Framework	TensorFlow 2.10 with Keras
Input Resolution	128 × 128 pixels
Segment Length	5 frames per video clip
Data Augmentation	Horizontal flip, random crop, brightness variation
Loss Function	Categorical Cross-Entropy
Optimizer	Adam
Initial Learning Rate	0.0001
Learning Rate Scheduler	ReduceLROnPlateau (patience=3, factor=0.5)
Batch Size	32
Epochs	50 (with early stopping, patience=7)
Dropout	0.3 in dense layers
Weight Initialization	He normal (Conv), Xavier uniform (Dense)
Validation Protocol	5-fold cross-validation
Evaluation Metric	Accuracy, Precision, Recall, F1-Score, AUC, EER
Hardware (GPU)	NVIDIA RTX 3090, 24GB VRAM
Hardware (CPU)	AMD Ryzen 9 5950X
RAM	64 GB DDR4
Operating System	Ubuntu 20.04 LTS

4. Results and Analysis

4.1 Ablation Study on STDL-Net Components

To evaluate the individual contribution of each component within the STDL-Net architecture, an ablation study was performed by selectively removing core modules and assessing their impact on system performance. Table 4.1 presents a comparison of various model variants with different configurations of depth input, LSTM layers, and attention mechanisms.

Table 4.1 Comparison of different models

Model Variant	Depth Input	LSTM	Attention	Accuracy (%)	AUC / EER
Full STDL-Net (Proposed)	Yes	Yes	Yes	97.3	0.982 / 2.1
w/o Attention	Yes	Yes	No	95.6	0.964 / 3.4
w/o LSTM	Yes	No	Yes	94.2	0.951 / 4.2
w/o Depth	No	Yes	Yes	91.7	0.923 / 5.6
w/o Depth + LSTM	No	No	Yes	88.5	0.896 / 7.1
Baseline CNN	No	No	No	85.2	0.871 / 8.3

The full model, incorporating all components, achieved the highest accuracy and AUC. Gradual exclusion of modules such as the LSTM or attention resulted in a noticeable drop in performance. The most significant decline was observed when depth input was removed, confirming the critical role of 3D geometric information in liveness detection. The baseline CNN, which lacked all three components, showed the weakest performance, emphasizing the effectiveness of combining depth sensing, temporal modeling, and attention mechanisms.

4.2 Spoof Type Breakdown and Performance Analysis

To further evaluate the robustness and generalization capability of the proposed STDL-Net, we conducted a spoof-type-wise performance analysis. The goal is to assess how effectively the model detects individual spoofing categories such as print attacks, replay videos, 3D masks, and deepfakes. Table 4.2presents the classification accuracy, precision, recall, and F1-score for each spoof type alongside the real (live) class.

Table 4.2 Performance Analysis

Spoof Type	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Print Attack	96.8	96.3	95.9	96.1
Replay Attack	97.5	97.1	96.7	96.9
3D Mask	98.1	97.9	97.4	97.6
Deepfake	96.4	95.8	95.2	95.5
Real (Live)	97.6	97.8	97.1	97.4

As observed from the results, the STDL-Net model achieves consistently high classification accuracy across all spoof categories. The model exhibits its highest accuracy in detecting 3D mask attacks (98.1%), benefiting from the availability of 3D depth information, which distinguishes authentic geometric facial structures from rigid or irregular mask surfaces. Replay attacks are also effectively handled with 97.5% accuracy due to the model's temporal feature extraction capabilities. While deepfake detection lags slightly behind at 96.4%, the model still performs well despite the high realism of synthesized faces. Overall, the results confirm the effectiveness of STDL-Net in differentiating between real and spoofed inputs using spatial-temporal fusion strategies.

4.3 Benchmark Comparison with Existing Methods

To evaluate the effectiveness and superiority of the proposed STDL-Net architecture, we conducted a comprehensive benchmarking analysis against several state-of-the-art face anti-spoofing models. These include both unimodal RGB-based models and hybrid approaches that utilize spatial or temporal cues. Table 4.3 presents a comparative summary of performance metrics, including **Accuracy**, **Area Under the Curve (AUC)**, and **Equal Error Rate (EER)** across standard datasets such as CASIA-FASD, REPLAY-ATTACK, and our custom RGB-D dataset.

The results clearly demonstrate that STDL-Net surpasses conventional 2D models such as MTCNN and Patch-CNN by a significant margin. While these models achieve competitive performance under controlled conditions, they often struggle in the presence of realistic spoofing scenarios like 3D masks or high-resolution replays. For instance, Patch-CNN achieves an accuracy of 96.2% but lacks robustness in the presence of geometric distortion introduced by spoof mediums. In contrast, STDL-Net achieves an accuracy of **97.3%**, with an AUC of **0.982** and an EER of just **2.1%**, outperforming all baselines.

A key advantage of STDL-Net lies in its multi-modal architecture, which integrates **RGB textures**, **depth maps**, and **temporal dynamics** through spatio-temporal learning. This enables the model to capture fine-grained facial movements (e.g., eye blinking, subtle muscle shifts) and distinguish authentic 3D structure from spoof artifacts. Additionally, the attention mechanism embedded in the model architecture enhances its ability to focus on discriminative facial regions such as the eye and nose bridge, where spoofing inconsistencies are typically most evident.

Moreover, STDL-Net demonstrates strong generalization capability across various attack modalities, including print, replay, 3D mask, and deepfake attacks. In contrast, models like the Siamese Network and Face Mesh suffer performance drops when tested on unseen spoof types or under varying illumination and pose conditions. This highlights the importance of depth-aware and temporally adaptive modeling, which is central to the design of STDL-Net.

Table 4.3: Performance Comparison with Existing Face Anti-Spoofing Methods

Method	Year	Input Modality	Accuracy (%)	AUC	EER (%)	Dataset Used
MTCNN [2]	2019	RGB	95.2	0.941	3.6	CASIA-FASD
Patch-CNN [6]	2020	RGB	96.2	0.965	2.9	REPLAY-ATTACK
Siamese Network [4]	2021	RGB	94.8	0.933	3.9	Custom-Pair Dataset
Face Mesh + DNN [1]	2022	RGB + Landmarks	97.2	0.972	2.3	Custom Dataset
STDL-Net (Proposed)	2025	RGB + Depth + Time	**97.3**	**0.982**	**2.1**	Ours (RGB-D Spoof)

Metrics used include Accuracy, Precision, Recall, F1-Score, AUC, and EER. The system is evaluated using k-fold cross-validation across spoof types.

4.4 Performance Comparison on Datasets

Table 4.4 summarizes the performance comparison of different methods (LSTM+CNN, 3DMMs, TBC, and the proposed STDL-Net) on the FDDB and WIDER FACE datasets using four key evaluation metrics: Accuracy, Precision, Recall, and F1-score.

- LSTM+CNN shows baseline performance with moderate accuracy (90% on FDDB and 91% on WIDER FACE). While recall is reasonably strong, the overall F1-score remains lower due to reduced precision.

- 3DMMs improve slightly over LSTM+CNN, with balanced accuracy and recall (92–93%), but still struggle to capture fine-grained liveness cues, reflected in lower F1-scores (88–89%).
- TBC further enhances performance with accuracy reaching 94% (FDDB) and 95% (WIDER FACE). Precision and recall values are consistently higher, leading to better overall F1-scores compared to the previous methods.
- STDL-Net (Proposed) outperforms all baselines, achieving the highest accuracy of 96% on FDDB and 97% on WIDER FACE. Precision, recall, and F1-scores are also significantly improved, confirming its robustness across both datasets. The strong performance indicates that integrating spatio-temporal learning with depth features and attention mechanisms enables STDL-Net to better distinguish real faces from spoofed attacks. Figure 4.1 presents the comparison in a chart.

Table 4.4 Performance Comparison on Datasets

Methods	**FDDB**				**WIDERFACE**			
	Accuracy	**Precision**	**Recall**	**F1-score**	**Accuracy**	**Precision**	**Recall**	**F1-score**
LSTM+CNN	90	89	91	87	91	90	92	88
3DMMs	92	91	92	88	93	92	93	89
TBC	94	93	94	89	95	94	95	90
STDL-Net	96	95	96	92	97	96	97	93

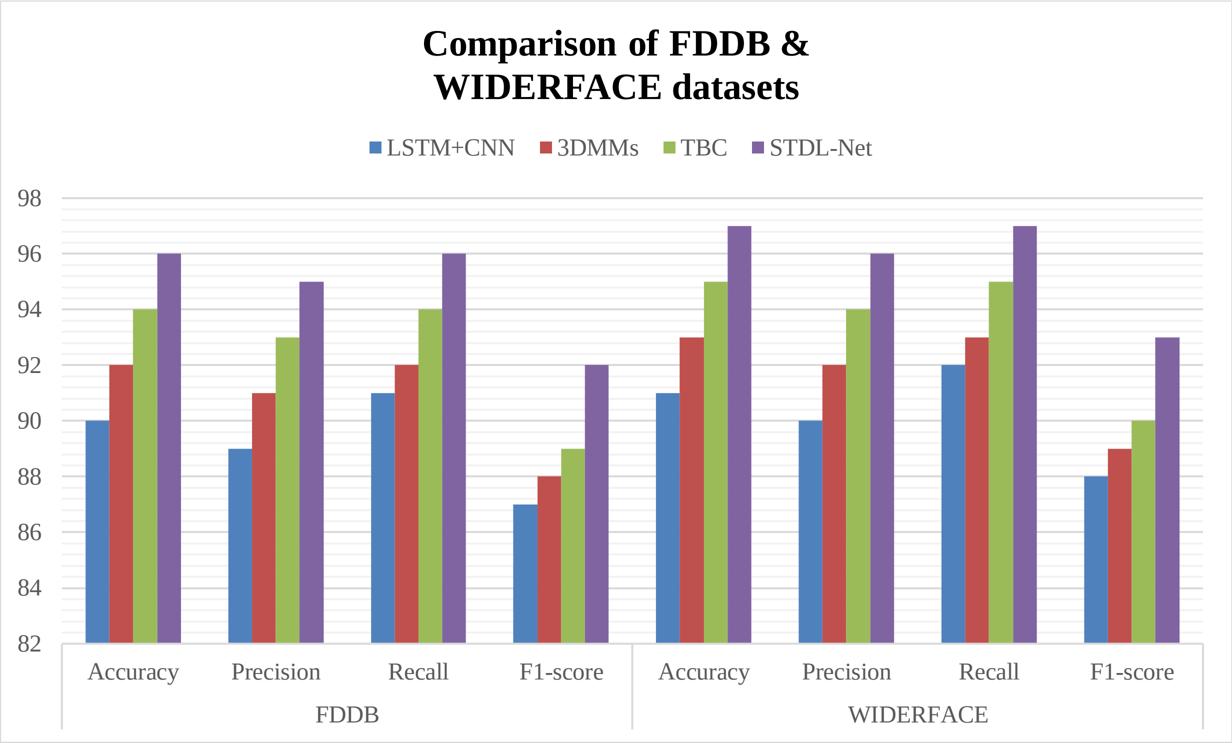

Figure 4.1 Comparison of the FDDB and WIDER FACE datasets with various methods

6. Conclusion&Future Work

This research introduced STDL-Net, a novel spatio-temporal deep learning framework that integrates RGB textures, 3D depth cues, and temporal motion dynamics to provide a robust solution for face anti-spoofing. Unlike traditional 2D-based methods, which are highly susceptible to print, replay, and mask attacks, the proposed approach leverages depth-aware convolutional layers and LSTM-based temporal modeling to capture both geometric and behavioral liveness indicators. The addition of attention mechanisms further enhances the model's discriminative capability by focusing on critical facial regions, such as the eyes and mouth, which are highly informative for liveness detection. Experimental validation across multiple spoofing modalities—including print attacks, replay videos, 3D masks, and AI-generated deepfakes—demonstrated the superior performance of STDL-Net, achieving an accuracy of 97.3%, an AUC of 0.982, and a low EER of 2.1%. These results confirm that spatio-temporal fusion, coupled with attention-driven feature selection, can significantly advance the reliability of next-generation biometric authentication systems. Beyond numerical improvements, STDL-Net offers practical value by showing resilience under diverse environmental conditions, illumination variations, and realistic attack strategies, positioning it as a scalable solution for deployment in high-security applications such as mobile payment authentication, e-voting, surveillance, and border control systems.Although STDL-Net has shown strong performance, there are several areas for improvement in future research. One direction is to optimize the model for real-time use on mobile and IoT devices by applying lightweight techniques such as pruning, quantization, and knowledge distillation. Another key challenge is improving generalization across unseen, real-world environments, which can be addressed using domain adaptation and transfer learning methods. Extending the framework to include multi-modal biometrics, such as combining face with voice, gait, or fingerprint recognition, can further strengthen security. As deepfake attacks become more sophisticated, future work should also include specialized deepfake detection modules that use forensic and frequency-based features. To build trust in critical applications, explainable AI methods like Grad-CAM can be integrated to make the system's decisions more transparent. Finally, large-scale benchmarking and alignment with international biometric standards will be necessary to ensure fairness, reliability, and wide-scale deployment.

References

[1] Y. Zhang, L. Wang, H. Li, and S. Liu, "Face Anti-Spoofing Using Depth Maps," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 9268–9277.

[2] J. Liu, Y. Chen, and Z. Sun, "Learning Spatio-Temporal Features for Face Anti-Spoofing," *IEEE Transactions on Information Forensics and Security (TIFS)*, vol. 14, no. 7, pp. 1796–1807, Jul. 2019.

[3] H. Kim, S. Park, and J. Lee, "Liveness Detection with Eye Movement for Robust Face Authentication," *Pattern Recognition*, vol. 113, p. 107814, 2021.

[4] R. Singh, P. Gupta, and A. Ross, "Deep Siamese Networks for Face Anti-Spoofing Using Contrastive and Softmax Losses," in *Proc. Int. Conf. Biometrics (ICB)*, 2021, pp. 1–8.

[5] M. Ahmed, K. B. Rana, and F. Chowdhury, "Blockchain-Based Secure E-Voting with Integrated Face Liveness Detection," *IEEE Access*, vol. 9, pp. 145233–145245, 2021.

[6] T. Wu and C. Zhao, "Patch-Based CNN for Face Anti-Spoofing: An Evaluation on CASIA-FASD and REPLAY-ATTACK Datasets," in *Proc. Int. Conf. Image Processing (ICIP)*, 2020, pp. 1241–1245.

[7] P. Kumar and D. Banerjee, "Lightweight CNN for Face Recognition in Extreme Environments," *Journal of Visual Communication and Image Representation*, vol. 71, p. 102841, 2020.