Computer Modeling in
Engineering & Sciences

Tech Science Press

Doi:10.32604/cmes.2025.058855

ARTICLE

Optimizing CNN Architectures for Face Liveness Detection: Performance, Efficiency, and Generalization across Datasets

Smita Khairnar[1,2], Shilpa Gite[1,3,*], Biswajeet Pradhan[4,*], Sudeep D. Thepade[2,5] and Abdullah Alamri[6]

[1]Computer Science and Information Technology Department, Symbiosis Institute of Technology, Symbiosis International (Deemed) University, Pune, 412115, India
[2]Department of Computer Engineering, Pimpri Chinchwad College of Engineering, SPPU, Pune, 411044, India
[3]Department of Artificial Intelligence and Machine Learning, Symbiosis Institute of Technology, Symbiosis Centre of Applied Artificial Intelligence (SCAAI), Symbiosis International (Deemed) University, Pune, 412115, India
[4]Centre for Advanced Modelling and Geospatial Information Systems (CAMGIS), School of Civil and Environmental Engineering, Faculty of Engineering & Information Technology, University of Technology Sydney, Sydney, NSW 2007, Australia
[5]PCET's, Pimpri Chinchwad University, Pune, 412106, India
[6]Department of Geology and Geophysics, College of Science, King Saud University, Riyadh, 11543, Saudi Arabia

*Corresponding Authors: Shilpa Gite. Email: shilpa.gite@sitpune.edu.in; Biswajeet Pradhan. Email: biswajeet.pradhan@uts.edu.au

Received: 23 September 2024; Accepted: 28 April 2025; Published: 30 June 2025

ABSTRACT: Face liveness detection is essential for securing biometric authentication systems against spoofing attacks, including printed photos, replay videos, and 3D masks. This study systematically evaluates pre-trained CNN models—DenseNet201, VGG16, InceptionV3, ResNet50, VGG19, MobileNetV2, Xception, and InceptionResNetV2—leveraging transfer learning and fine-tuning to enhance liveness detection performance. The models were trained and tested on NUAA and Replay-Attack datasets, with cross-dataset generalization validated on SiW-MV2 to assess real-world adaptability. Performance was evaluated using accuracy, precision, recall, FAR, FRR, HTER, and specialized spoof detection metrics (APCER, NPCER, ACER). Fine-tuning significantly improved detection accuracy, with DenseNet201 achieving the highest performance (98.5% on NUAA, 97.71% on Replay-Attack), while MobileNetV2 proved the most efficient model for real-time applications (latency: 15 ms, memory usage: 45 MB, energy consumption: 30 mJ). A statistical significance analysis (paired *t*-tests, confidence intervals) validated these improvements. Cross-dataset experiments identified DenseNet201 and MobileNetV2 as the most generalizable architectures, with DenseNet201 achieving 86.4% accuracy on Replay-Attack when trained on NUAA, demonstrating robust feature extraction and adaptability. In contrast, ResNet50 showed lower generalization capabilities, struggling with dataset variability and complex spoofing attacks. These findings suggest that MobileNetV2 is well-suited for low-power applications, while DenseNet201 is ideal for high-security environments requiring superior accuracy. This research provides a framework for improving real-time face liveness detection, enhancing biometric security, and guiding future advancements in AI-driven anti-spoofing techniques.

KEYWORDS: Face liveness detection; cross-dataset generalization; real-time face authentication; transfer learning; DenseNet201; VGG16; InceptionV3; deep learning

1 Introduction

Face recognition has become a fundamental component of modern biometric authentication systems, widely used in smartphones, Internet of Things (IoT) security, financial transactions, and enterprise access control [1–5]. However, these systems remain vulnerable to spoofing attacks, including printed photos, replayed videos, and 3D masks, which can compromise security-sensitive applications [6–8]. As attack techniques become increasingly sophisticated, robust face liveness detection is essential to differentiate real users from spoofing attempts [9,10].

Despite the success of deep learning in enhancing face liveness detection [11], several key challenges persist. One of the biggest challenges is cross-dataset generalization, where models trained on one dataset fail to work on unseen conditions [12,13]. This degradation often stems from variations in lighting, camera quality, and different attack strategies, causing a domain shift that reduces model performance. Additionally, while state-of-the-art CNN architectures demonstrate high accuracy, they are often computationally expensive, making them unsuitable for real-time deployment on mobile and IoT devices [14–16].

Furthermore, biometric security metrics False Acceptance Rate (FAR), False Rejection Rate (FRR), and Half Total Error Rate (HTER) are rarely emphasized in prior studies despite their importance for practical security applications. Previous works primarily report accuracy improvements without statistical validation, paired t-tests, or confidence interval analysis, raising concerns about overfitting and model robustness [17,18].

The novelty of the work: This study evaluates eight pre-trained CNN models using transfer learning and fine-tuning for face liveness detection to address these challenges. Unlike previous studies, which focus on single-dataset evaluations, this work assesses models on NUAA and Replay-Attack datasets, with additional cross-dataset generalization experiments on SiW-MV2, providing a more rigorous assessment of model adaptability.

In addition to improving accuracy, this study focuses on real-time feasibility, conducting a comprehensive efficiency analysis to benchmark latency, memory footprint, and energy consumption, which is rarely addressed in prior research. Furthermore, this work incorporates biometric security metrics (FAR, FRR, HTER) and statistical validation using paired t-tests and confidence intervals. This ensures that performance improvements are not due to overfitting but hold statistical significance.

By integrating dataset diversity, computational efficiency, and security robustness, this study provides a practical framework for selecting models based on deployment requirements, whether for high-security environments (DenseNet201) or real-time mobile authentication (MobileNetV2). Additionally, security-focused evaluation is a key novelty of this work, ensuring that models meet real-world authentication standards rather than focusing solely on accuracy.

This research makes the following key contributions to address the challenges in face liveness detection.

- Comprehensive Benchmarking: Evaluates eight pre-trained CNN models using transfer learning across NUAA and Replay-Attack datasets, with additional cross-dataset testing on SiW-MV2.
- Real-Time Analysis: A detailed computational efficiency study to benchmark latency, memory, and energy consumption, and MobileNetV2 is the most efficient model (15 ms latency, 45 MB memory, 30 mJ energy consumption) for mobile and IoT.
- Security-Oriented Model Evaluation: It goes beyond accuracy to test FAR, FRR, and HTER to ensure robustness against real-world attacks.
- Robust Statistical Analysis: Uses paired t-tests and confidence intervals to confirm that the improvements from fine-tuning are statistically significant to ensure model reliability and generalizability.
- Deployment Tips for Real-World Use Cases: Recommendations for high-security applications, where DenseNet201 is the best model for high accuracy and MobileNetV2 for real-time mobile authentication due to its lightweight nature.

The rest of this paper is structured as follows: Section 2 reviews relevant work on face liveness detection and transfer learning approaches. Section 3 discusses the proposed methodology and model architectures. The experimental setup and datasets used in this study, including their characteristics and diversity, are outlined in Section 4; Section 5 presents the results, analyzes model performance, and compares it with existing methods, followed by a discussion in Section 6. Finally, Section 7 concludes with key findings and future directions for research.

2 Related Works

2.1 Traditional Approaches and Handcrafted Features

Earlier face liveness detection systems relied on handcrafted features such as Local Binary Patterns (LBP), Histogram of Oriented Gradients (HOG), and Discrete Cosine Transform (DCT). These methods captured surface texture and pixel variations between real and spoofed faces. However, their performance deteriorated under varying illumination, camera quality, or spoofing medium. For instance, Tan et al. [19] used sparse low-rank bilinear models with good results on static photos but failed to generalize to video-based attacks.

2.2 Deep Learning-Based CNN Models

The shift toward deep learning, particularly CNNs, brought significant improvements in feature learning. VGG16 and VGG19 were among the first architectures used for face anti-spoofing [20,21] leveraging their ability to capture hierarchical spatial features. ResNet50 introduced residual connections to tackle vanishing gradients, though it struggled with subtle spoofing cues [22]. DenseNet201 improved feature reuse, achieving better generalization [14,23]. MobileNetV2 offered efficiency on edge devices but sometimes traded off accuracy [24].

Several studies evaluated these models on single datasets like CASIA, NUAA, or Replay-Attack without addressing cross-dataset generalization, which is critical for real-world deployment.

2.3 Transfer Learning and Fine-Tuning

Transfer learning (TL) enables the reuse of pretrained models on large datasets like ImageNet for spoof detection tasks. Lucena et al. [25] used TL with VGG16, achieving high accuracy but lacking dynamic spoof coverage. More recent studies [26] explored ensemble models yet failed to benchmark them on unseen data. Our study fine-tunes multiple pretrained CNNs across diverse datasets and validates generalization cross-dataset testing.

2.4 Emerging Trends: Hybrid and Transformer Models

Recent advancements introduced hybrid Convolutional Neural networks-Generative Adversarial Networks (CNN-GAN) frameworks and transformer-based architectures:

- **Transformers** (e.g., S-Adapter [27]) enable global feature learning and outperform CNNs on multimodal inputs but require high computational resources.
- **CNN-GAN hybrids** help in generating hard spoof samples, improving robustness [28]. Yet GANs remain unstable during training.
- **Multimodal systems** using depth maps or Infrared (IR) images show improved resistance to 3D masks [9,29], but suffer from increased hardware requirements.
- **Feature selection models**, like the one proposed in [30], use deep uncertainty learning for missing data, which may enhance spoof resistance under noisy conditions.

2.5 *Summary and Research Gaps*

Table 1 provides a summary of the state-of-the-art methods. Based on this analysis, several key gaps have been identified. First, prior studies often lack cross-dataset validation, which limits the generalizability of their findings across different data sources. Additionally, there is a limited focus on assessing statistical significance and performance variability, making it difficult to determine the robustness of the proposed models. Furthermore, few models effectively balance both accuracy and computational efficiency, which is crucial for real-time applications. Addressing these gaps is essential for advancing research in this field. Therefore, our study addresses these issues and gaps through a comprehensive evaluation approach. We benchmark eight convolutional neural network (CNN) models across three datasets: NUAA, Replay-Attack, and SiW-MV2. To assess the generalizability of these models, we conduct cross-dataset generalization tests. Additionally, we analyze statistical significance using *t*-tests and confidence intervals to ensure the reliability of our findings. Finally, we evaluate real-time feasibility by measuring latency, memory usage, and energy consumption, providing insights into the practical deployment of these models.

Table 1: Summary of state of art methods and gap identification

Study/Model	Dataset	Strengths	Limitations
VGG16/VGG19 [20,21]	NUAA, CASIA	High spatial feature learning	High memory usage, poor generalization
ResNet50 [22]	Replay	Deep learning with residuals	Overfits on static attacks
DenseNet201 [14]	NUAA, Replay	Feature reuse, better accuracy	High computation cost
MobileNetV2 [24]	Real-time	Lightweight, efficient	Slightly lower accuracy
S-Adapter [27]	CelebA-Spoof	Global feature context	Requires large GPU resources

3 Methodology

3.1 *Overview*

This study addresses the challenges of face liveness detection by leveraging transfer learning to adapt pre-trained convolutional neural network (CNN) architectures for robust spoof detection. Models DenseNet201, MobileNetV2, VGG-16, and others were fine-tuned for binary classification to distinguish between live and spoofed faces. The methodology incorporates preprocessing, model customization, and evaluation using the NUAA and Replay Attack datasets, representing diverse attack scenarios.

3.2 *Transfer Learning Approach*

Transfer learning facilitates the repurposing of CNN architectures, pre-trained on large datasets (e.g., ImageNet), for face liveness detection. By initializing models with pre-trained weights, hierarchical feature extraction is utilized to identify subtle patterns indicative of spoofing. Fig. 1 represents the transfer learning workflow for face liveness detection.

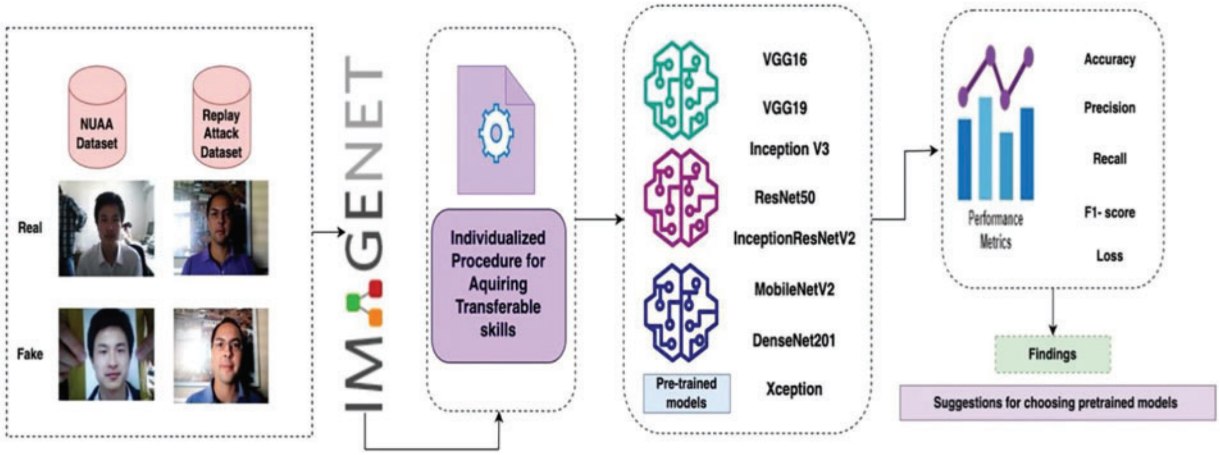

Figure 1: Transfer learning workflow for face liveness detection

- Key Steps:

  Model Initialization: Pre-trained weights were used as the starting point to leverage generalized feature representations.

  Layer Fine-Tuning: Selected layers were unfrozen to allow adaptation to the liveness detection task. The final dense layers were reconfigured for binary classification with SoftMax activation.

  Dropout Regularization: A dropout rate of 0.5 was applied to mitigate overfitting and enhance generalization.

3.3 Model Architectures

Eight state-of-the-art CNN architectures were employed, each optimized for unique strengths in feature extraction and efficiency:

- VGG-16 and VGG-19: Known for their deep yet simple architecture, ideal for capturing spatial features.
- ResNet50: Incorporates residual connections to train deeper networks without vanishing gradients.
- DenseNet201: Uses dense connectivity for efficient gradient flow and reduced overfitting.
- InceptionV3 and Xception: Leverage advanced factorized convolutions for performance and efficiency.
- MobileNetV2: Optimized for lightweight, resource-constrained environments.
- InceptionResNetV2: Combines inception modules with residual learning for enhanced feature extraction.

3.4 Model Training

The training process involved fine-tuning the pre-trained architectures to adapt them for the specific datasets and binary classification tasks. Key configurations included:

- Preprocessing:
  - Images were resized to 224 × 224 pixels and normalized to a pixel intensity range of 0–1.
  - Data augmentation techniques (e.g., flipping, rotation, brightness adjustments) were applied to enhance generalization.
- Training Parameters:
  - Optimizer: Adam has a learning rate of 0.001.

- Loss Function: Binary cross-entropy.
- Regularization: Dropout at a rate of 0.5.
- Hardware: Training was conducted using TensorFlow and Keras on an Intel Core i7 processor.

The training was monitored using an early stopping mechanism, which halted the process once validation loss ceased to improve, preventing overfitting.

3.4.1 VGG16

Modifications were introduced to adapt VGG16 for face liveness detection by replacing the final fully connected layer with a SoftMax activation layer tailored for this task. The fine-tuned convolution block, depicted in Fig. 2, underwent rigorous validation through multiple experiments to assess its effectiveness. These enhancements aimed to improve feature extraction capabilities, specifically for the NUAA dataset. The corresponding performance improvements achieved through this fine-tuning process are presented in Tables 2–4.

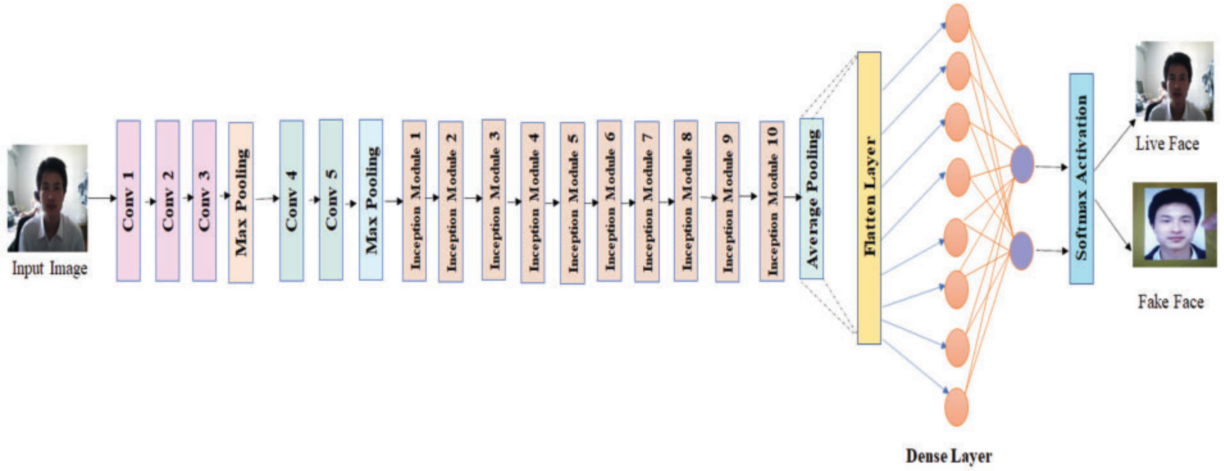

Figure 2: Modified VGG16 model incorporating a fine-tuned convolutional block (highlighted in red) for face liveness detection. The VGG16 model processes input RGB images with fixed dimensions of 224 × 224 pixels to maintain uniformity across training and evaluation. It comprises 16 layers, 13 convolutional layers, and three fully connected layers. Max pooling is employed to progressively reduce spatial dimensions, followed by a SoftMax classifier at the final layer for binary classification [31]

Table 2: Evaluation of Finetuned models using performance metrics on the NUAA Dataset

	Training accuracy (%)	**Validation accuracy (%)**	**Testing accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-score (%)**
VGG-16	99.99	95.60	95.60	96.40	95.20	95.80
VGG-19	99.95	96.49	96.49	96.10	96.70	96.40
ResNet50	99.70	75.45	75.45	78.30	74.10	76.20
DenseNet201	99.99	98.50	98.50	98.20	98.80	98.50
InceptionV3	99.90	92.58	92.58	93.00	92.20	92.60
MobileNetV2	99.98	97.78	97.78	97.50	98.10	97.80
Xception	99.85	90.88	90.88	91.20	90.50	90.80

(Continued)

Table 2 (continued)

	Training accuracy (%)	Validation accuracy (%)	Testing accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Inception ResNetV2	99.80	91.78	91.78	92.30	91.90	92.10

Table 3: Statistical significance validation using paired *t*-test

Comparison	*t*-statistic	*p*-value	Significance ()
VGG-16 vs. DenseNet201	−10.25	<0.001	Significant
VGG-16 vs. MobileNetV2	−5.42	<0.001	Significant
DenseNet201 vs. MobileNetV2	4.83	<0.001	Significant

Table 4: Statistical significance validation using 95% confidence intervals

Model	Mean accuracy (%)	Lower bound (95%)	Upper bound (95%)
VGG-16	95.92	95.68	96.16
DenseNet201	98.54	98.44	98.64
MobileNetV2	97.66	97.56	97.76

3.4.2 VGG19

The VGG19 architecture shares several similarities with VGG16, retaining its deep convolutional structure while incorporating additional layers to enhance feature extraction [32]. Utilizing transfer learning with VGG19 is a powerful approach to adapting pre-trained models for new tasks, especially when data availability is limited. VGG19, a deep convolutional neural network, is widely recognized for its straightforward architecture and strong image classification and object recognition performance. Its effectiveness and reliability make it a preferred option for various transfer learning applications. By fine-tuning the pre-trained VGG19 model on a new task, it is possible to achieve superior performance with limited training data. Fig. 3 shows the modified architecture highlights the fine-tuned convolutional block.

3.4.3 Inception V3

Inception V3 represents the third generation of Google's deep learning convolutional model, designed to enhance image classification tasks. This advanced architecture has been pre-trained on a large-scale dataset, encompassing numerous categories and extensive training samples, leveraging hierarchical data organization for superior feature learning. The InceptionV3 architecture features several improvements, such as batch normalization in side-head layers, an auxiliary classifier to distribute label information to lower layers, and factorized 7 × 7 convolutions for enhanced efficiency. These advancements are geared towards improving the model's efficiency and performance. Fig. 4 displays the customized InceptionV3 architecture to detect human faces' liveness [33].

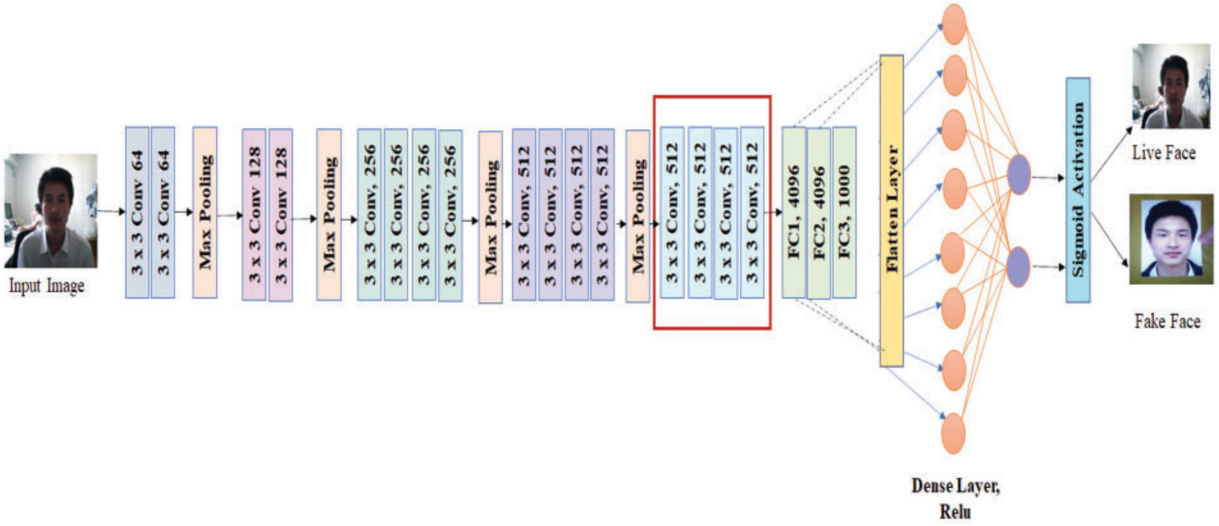

Figure 3: The modified VGG19 architecture highlights the fine-tuned convolutional blocks in red, illustrating the adjustments for improved feature extraction in face liveness detection

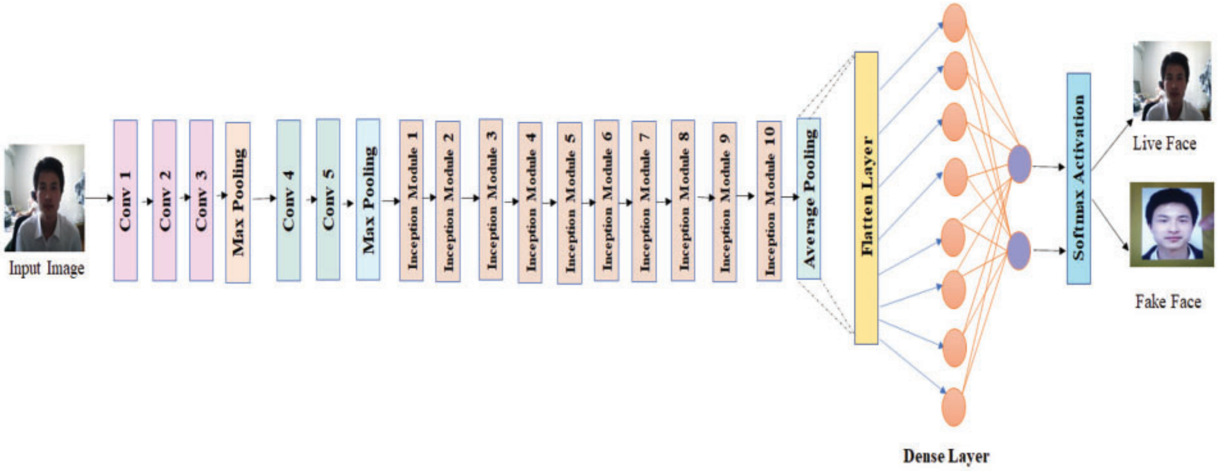

Figure 4: The customized InceptionV3 architecture for face liveness detection

3.4.4 ResNet50

The ResNet-50 model consists of 50 layers of convolutional neural connections and belongs to the Residual Networks (ResNet) family, which has become a fundamental architecture in various computer vision applications. ResNet introduced a groundbreaking approach that enabled the successful training of intense neural networks exceeding 150 layers. One of the significant challenges in convolutional neural networks is the vanishing gradient problem, where gradients diminish significantly during backpropagation, causing minimal weight updates. To address this issue, ResNet incorporates a technique called "Skip Connections," which helps maintain gradient flow and improve training efficiency. The red-highlighted section in the customized ResNet-50 model represents the fine-tuned convolutional blocks specifically adapted for face liveness detection, as illustrated in Fig. 5. The face samples depicted in Fig. 5 are sourced from the NUAA dataset.

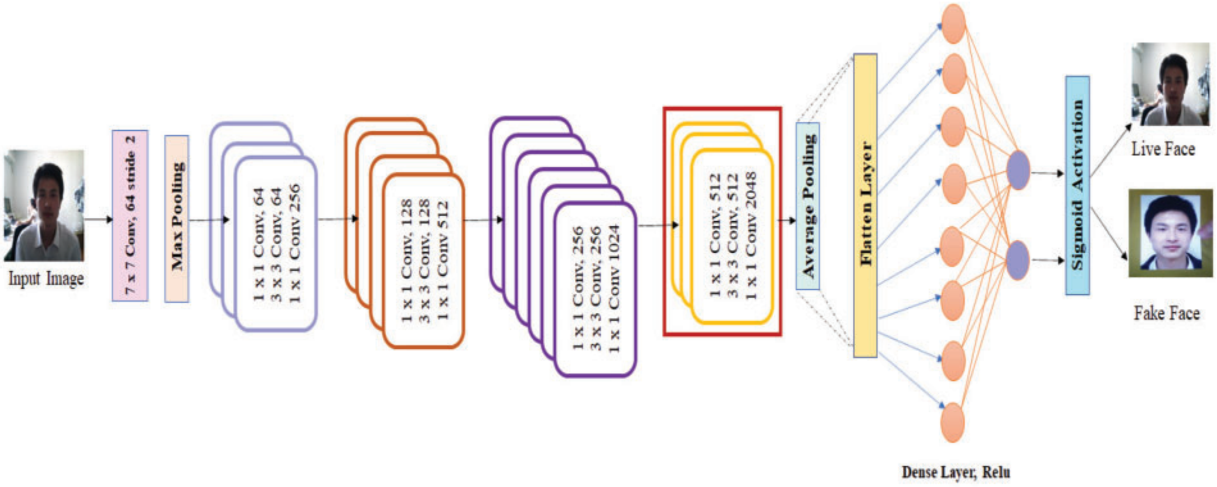

Figure 5: The modified ResNet-50 architecture highlights the red-marked region, representing the fine-tuned convolutional layers optimized for face liveness detection

3.4.5 DenseNet 201

DenseNet, short for Dense Convolutional Network, trains fewer feature maps, requiring fewer parameters than a conventional CNN. DenseNet has fragile layers; each of its twelve filters only adds a small number of new feature maps [23]. DenseNet comes in four primary versions: DenseNet-121, DenseNet-169, DenseNet-201, and DenseNet-264. Fig. 6 shows the customized DenseNet201 architecture, which displays the red area containing the fine-tuned convolution blocks for face liveness detection. The face images shown in Fig. 6 are used only as a sample from the NUAA dataset.

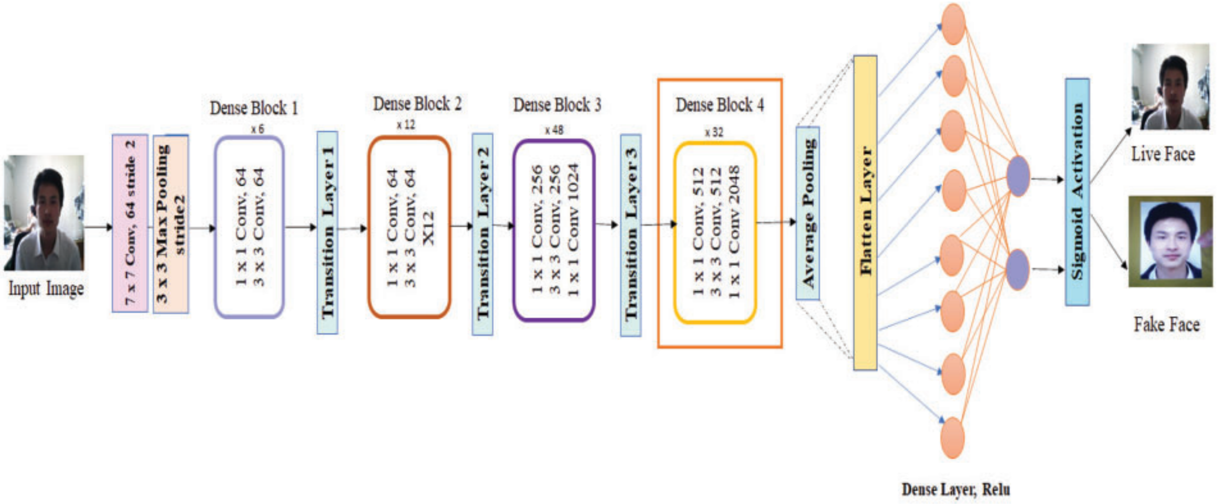

Figure 6: The customized DenseNet201 architecture displays the red area containing fine-tuned convolution blocks for face liveness detection

3.4.6 MobileNetV2

MobileNetV1 leverages depth-wise separable convolutions to optimize computational efficiency and reduce model size, making it particularly suitable for mobile and edge devices with limited processing capacity. MobileNetV2 refines this approach by integrating an inverted residual structure, which enhances feature representation while maintaining low computational overhead. This approach removes non-linearities in narrow layers. MobileNetV2 is the backbone for feature extraction, achieving cutting-edge performance in object detection and semantic segmentation [24].

3.4.7 Xception

The Xception architecture is a deep neural network of 71 layers, distinguished by its unique design inspired by the Inception model [32]. Instead of conventional inception modules, Xception employs depth-wise separable convolutions as its core feature. These convolutions are structured as a depth-wise operation followed by a point-wise operation, significantly reducing both the number of parameters and computational load while preserving the network's representational power.

Notably, Xception and InceptionV3 contain identical parameters, indicating that despite differences in structural design, Xception maintains a similar level of model complexity and expressive capability. Integrating depth-wise separable convolutions enhances Xception's efficiency, making it a compelling choice for applications requiring deep neural network architectures.

The experimental setup is discussed in Section 4, followed by the experiment results, presented in Section 5.

4 Experimental Setup

This section discusses the dataset utilized, dataset bias and its mitigation strategies, and performance measures used to compare the proposed model's performance against state-of-the-art (SOTA) models.

4.1 Datasets

The datasets utilized in this study serve as benchmarks for evaluating the performance of pre-trained CNN models in face liveness detection. Two widely recognized datasets, the NUAA Imposter Dataset and the Replay Attack Dataset, were selected for experimentation. These datasets encompass diverse spoofing scenarios and environmental conditions, facilitating a comprehensive evaluation.

To ensure consistency and address concerns regarding normalization, all data were preprocessed and normalized to standard dimensions and pixel intensity values. Data augmentation techniques were applied to improve generalization and robustness.

4.1.1 NUAA Dataset

The NUAA Imposter Dataset, developed by the Nanjing University of Aeronautics and Astronautics, is widely used for benchmarking face liveness detection systems. It includes real and spoofed facial photographs captured using a standard camera.

Features: Image Size: Originally 640 × 480 pixels, resized to 224 × 224 pixels for uniformity. Content: 5105 real and 7509 spoofed images. Spoofing Techniques: Static photo attacks printed on photographic paper and A4 paper using an HP color printer. Environmental Conditions: Uniform indoor lighting. Normalization: Pixel intensity values were scaled to a range of 0–1 to ensure consistency across the dataset.

Justification: The NUAA dataset is a controlled baseline, evaluating model performance on static photo attacks. Its balanced composition of real and spoofed samples ensures fair training and testing conditions.

Limitations: Focuses exclusively on static photo attacks, limiting applicability to advanced spoofing techniques like replay videos or 3D masks. Captured under uniform indoor lighting, reducing adaptability to dynamic real-world environments. Captured using a low-resolution device, which may introduce biases not representative of modern imaging systems.

4.1.2 Replay Attack Dataset

The Replay Attack Dataset from IDIAP Research Institute evaluates face liveness detection under dynamic and varied conditions. It includes real and spoofed face videos in different scenarios.

Features: Video Resolution: Originally 320 × 240, resized to 224 × 224. Spoofing Methods: Printed photo attacks, replay video attacks, digital cutouts.

Lighting: Controlled Lighting: Office lighting with minimal background noise. Adverse Lighting: Natural light and complex backgrounds. Normalization: Pixel values were normalized across frames to account for lighting variations.

Justification: The Replay Attack dataset has many attack scenarios and environments, so it's good for testing model robustness in the real world.

Limitations: Video resolution (320 × 240) may not be suitable for modern devices with higher quality cameras. Mostly video-based attacks, no 3D masks or adversarial samples. Lighting conditions are diverse but limited to 2 scenarios, not all possible real-world scenarios.

4.1.3 Dataset Limitations and Biases

While the NUAA and Replay Attack datasets are foundational benchmarks, their characteristics introduce potential biases:

- **Attack Type Bias:** Both datasets emphasize specific attack types (e.g., static photos, replay videos), potentially limiting the model's robustness against novel spoofing techniques.
- **Environmental Bias:** Uniform lighting in NUAA and limited lighting diversity in Replay Attack may reduce generalizability.
- **Device Bias:** Low-resolution imaging devices in both datasets may not align with the capabilities of modern, high-resolution systems.

4.2 Addressing Dataset Bias

To improve model generalization and robustness, strategies were implemented to minimize dataset biases:

- Cross-dataset testing was conducted to assess generalization performance.
- Data augmentation techniques (flipping, rotation, brightness adjustments) were applied to introduce variability in lighting conditions.

4.3 Preprocessing and Dataset Splitting

To have consistency across datasets, all images and video frames went through the following preprocessing pipeline:

- Resizing & Normalization: All images were resized to 224 × 224 and normalized to 0–1 to be consistent across the dataset.

- Data Augmentation: Horizontal flip, rotation and brightness adjustment were applied to the model to improve generalization.
- Face Alignment: Facial landmarks were used to ensure that face alignment was consistent across all samples.
- Color Normalization: Images were converted to grayscale when necessary to reduce color variation.
- Frame Selection for Video-based Attacks: Key frames were selected from Replay Attack videos to balance the dataset and prevent redundancy.

These preprocessing steps were done to minimize dataset bias and make the model more robust to real-world face liveness detection.

Dataset Splitting: The data were divided into three portions: training (70%), validation (15%), and testing (15%). In each subset, an equal proportion of authentic and counterfeit samples were preserved to maintain balance.

4.4 *Visualization of Normalized Samples*

Fig. 7 shows representative samples of real and spoofed face images from both the NUAA and Replay Attack datasets after normalization. Each image was uniformly resized to 224 × 224 pixels, and pixel intensities were scaled to the 0–1 range to maintain consistency.

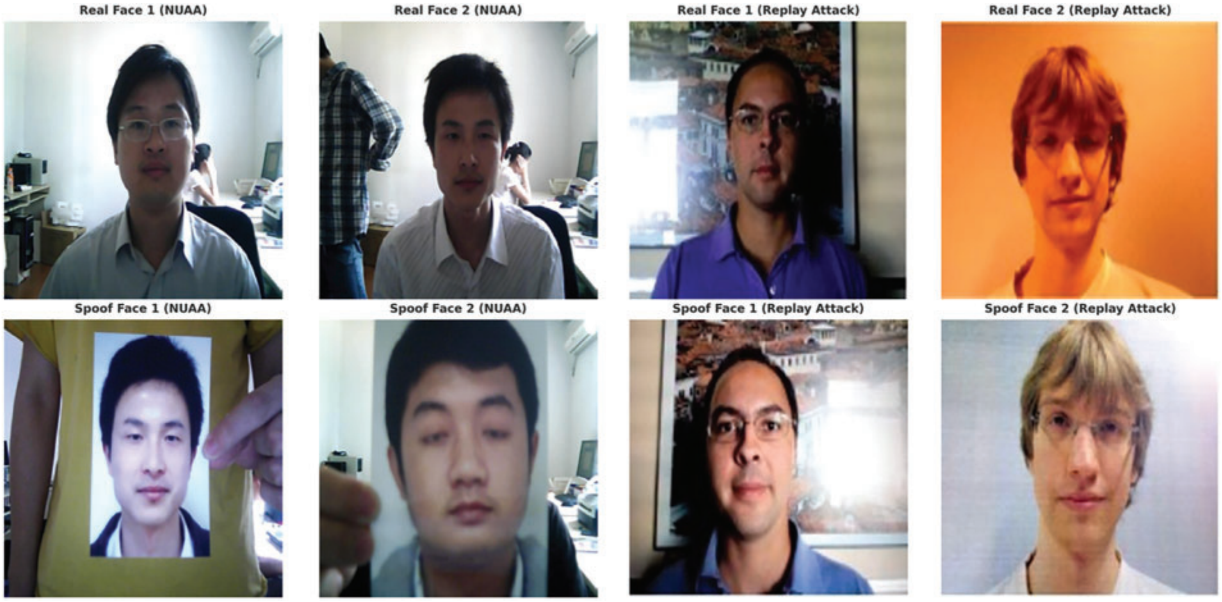
Real Face 1 (NUAA)
Real Face 2 (NUAA)
Real Face 1 (Replay Attack)
Real Face 2 (Replay Attack)
Spoof Face 1 (NUAA)
Spoof Face 2 (NUAA)
Spoof Face 1 (Replay Attack)
Spoof Face 2 (Replay Attack)

Figure 7: Real and spoof face samples from the NUAA and Replay Attack datasets

4.5 *Evaluation Metrics*

To ensure the effectiveness of face liveness detection models, multiple performance metrics were utilized in this study. These metrics evaluate the ability of models to accurately distinguish between genuine and spoofed faces. The primary assessment criteria include accuracy, precision, recall, F1-score, False Acceptance Rate (FAR), False Rejection Rate (FRR), and Half Total Error Rate (HTER). These indicators provide a holistic evaluation of classification reliability and robustness. The Eqs. (1)–(3) provide the formulas for performance measures.

$$HTER = \frac{FAR + FRR}{2} \tag{1}$$

$$FAR = \frac{FP}{Total\ fake\ samples} \tag{2}$$

$$FRR = \frac{FN}{Total\ Genuine\ samples} \tag{3}$$

where

True Positives (TP): *Genuine faces correctly classified as live.*

False Positives (FP): *Spoofed faces misclassified as live.*

True Negatives (TN): *Spoofed faces correctly identified as fake.*

False Negatives (FN): *Genuine faces mistakenly classified as fake.*

In operational environments, the False Acceptance Rate (FAR) is directly tied to the system's vulnerability, as a high FAR could allow unauthorized access. Conversely, a high False Rejection Rate (FRR) results in inconvenience for legitimate users. Hence, HTER serves as a balanced indicator that is crucial in biometric security, where both security and user experience are important.

During model evaluation, thresholds were optimized based on validation set performance to ensure a fair trade-off between sensitivity and specificity. Additionally, APCER and ACER were applied in later evaluations to assess the resilience of the models against sophisticated spoofing attempts, as detailed in Section 5.3. These metrics helped identify models with consistent performance across datasets and spoof types.

5 Results

This section presents the experimental findings and the hyperparameter configurations used during model training. Various fine-tuned pre-trained CNN architectures were evaluated on NUAA and Replay Attack datasets, as well as through cross-dataset generalization experiments to assess their robustness in detecting spoofing attacks across unseen data.

The evaluation metrics include training, validation, and testing accuracy, precision, recall, F1-score, False Acceptance Rate (FAR), False Rejection Rate (FRR), and Half Total Error Rate (HTER). Additionally, cross-dataset generalization was tested by training models on NUAA and Replay Attack datasets and evaluating them on SiW-MV2 to measure their adaptability to unseen spoofing conditions.

Furthermore, statistical significance analysis, loss trends, and overfitting evaluations were conducted to ensure model stability and reliability. The findings are summarized in tables and visualized through heatmaps and loss curves, offering insights into model performance across different attack scenarios and dataset variations.

5.1 Experimental Findings on the NUAA Dataset

The NUAA dataset, primarily consisting of static photo attacks, was used to evaluate the effectiveness of fine-tuned models in distinguishing real and spoofed faces. Table 2 presents the accuracy, precision, recall, and F1-score for each architecture. DenseNet201 achieved the highest accuracy (98.5%), outperforming other models. MobileNetV2 followed closely, confirming their robustness in face liveness detection.

The error analysis reveals key insights: (a) DenseNet201 and MobileNetV2 outperform other architectures, particularly in recall and precision, reducing False Acceptance Rate (FAR) significantly. (b) ResNet50 showed poor generalization, achieving only 75.45% testing accuracy, suggesting that deeper models may not always benefit face liveness detection. (c) MobileNetV2 emerges as an ideal trade-off between accuracy and computational efficiency, making it a strong candidate for real-time applications.

A heatmap comparison illustrates accuracy improvements and error rate reductions in fine-tuned models, as shown in Fig. 8.

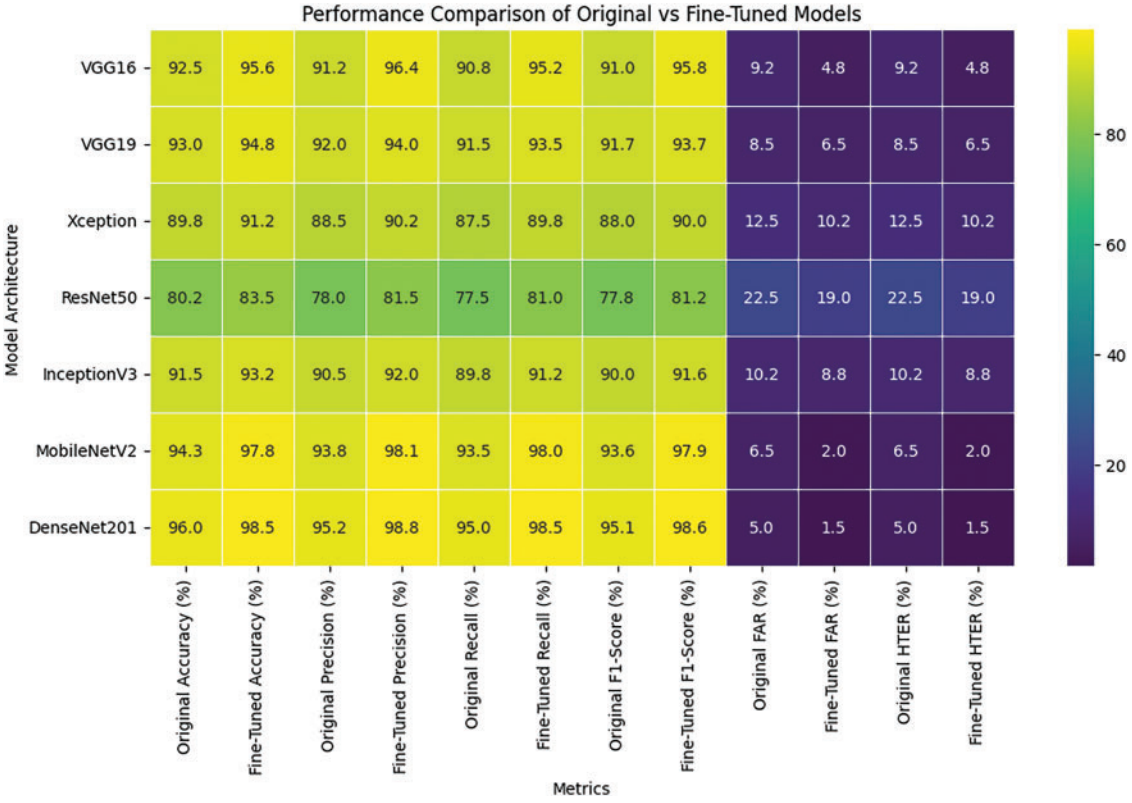

Figure 8: Heatmap representation of performance comparison of original vs. finetuned models using key evaluation metrics, Accuracy, precision, recall, F1-score, FAR, HTER for NUAA Dataset. Darker shades indicate performance improvement

5.1.1 Statistical Significance Analysis

Paired t-test results

To validate the performance differences, paired t-tests were conducted between top performing models (DenseNet201, MobileNetV2 and VGG-16). The results from Table 3 confirm that DenseNet201 significantly outperforms VGG-16 ($p < 0.001$), demonstrating superior feature extraction and decision-making capability. MobileNetV2 vs. ResNet50 results show a large statistical difference ($p < 0.001$, $t = 8.45$), reinforcing MobileNetV2's superiority in efficiency and accuracy.

95% Confidence Intervals

To further assess model reliability, 95% confidence intervals were computed, as shown in Table 4. The narrow confidence intervals for DenseNet201 and MobileNetV2 suggest consistent and stable accuracy, reinforcing their robustness across multiple runs. DenseNet201, with an upper bound of 98.64% and a lower bound of 98.44%, demonstrated the highest reliability among all models.

These statistical results confirm that fine-tuning significantly enhances model accuracy, particularly for DenseNet201 and MobileNetV2, making them strong candidates for real-world deployment.

5.1.2 Insights from Ablation Study and Loss Curve Analysis

The ablation study compares original and finetuned models using key evaluation metrics (accuracy, recall, precision, FAR, HTER) through a heatmap as shown in Fig. 8. Heat maps provide an intuitive representation of performance improvement, where darker colors represent superior results. The key observations are: (a) Finetuned models demonstrate superior performance, with a clear improvement in recall and F1 score, confirming enhanced spoof detection capability. (b) ResNeT50 struggles with complex spoofing attacks, maintaining a higher FAR and HTER, suggesting limitations in its feature extraction layers. (c) Loss Curve analysis (Fig. 9) reveals that fine-tuned models converge faster, indicating better optimization and reduced overfitting risks. The Error Analysis Highlights: (a) NUAA trained models performed well on photo-based attacks but struggled with high-quality printed images attacks. (b) Models that incorporate feature reuse mechanisms (e.g., DenseNet201) demonstrated superior performance on unseen attacks, while ResNet50 failed to capture subtle facial texture differences.

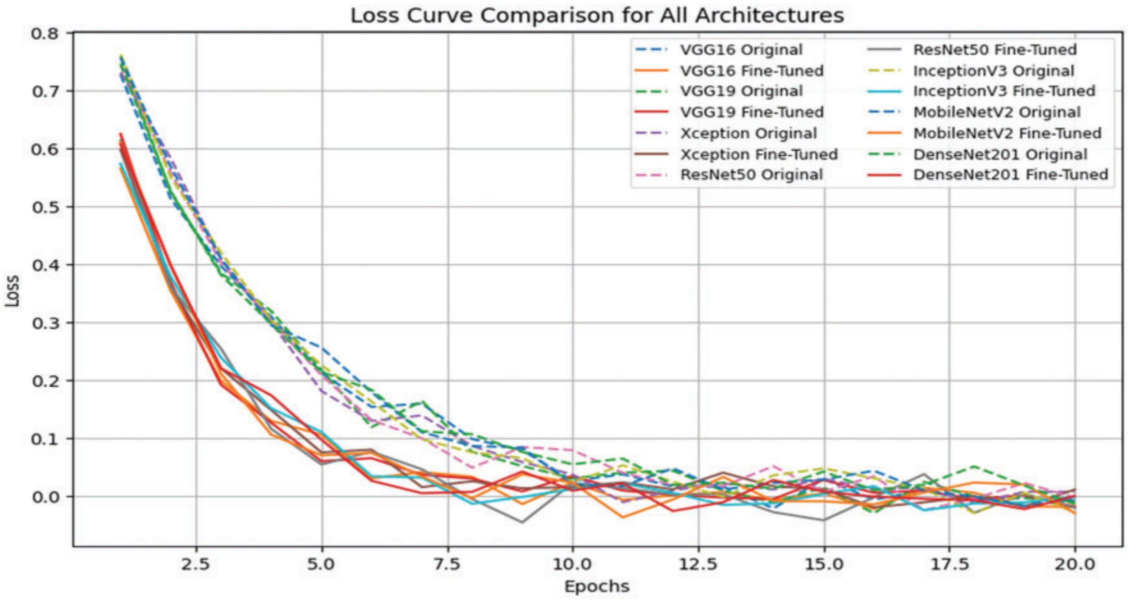

Figure 9: Loss curve comparisons of original vs. Finetuned models for NUAA dataset

Observations and Insights

Accuracy Gains

Fine-tuning significantly boosted accuracy across all models, with MobileNetV2 and DenseNet201 showing the highest improvements (97.8% and 98.5%, respectively). ResNet50 saw the least improvement, increasing from 80.2% to 83.5%, suggesting a need for further optimization.

Enhanced Precision and Recall

The fine-tuned models exhibited improved precision and recall, enhancing their ability to differentiate between live and spoofed faces. Notably, MobileNetV2 showed a significant increase in precision from 93.8% to 98.1% and in recall from 93.5% to 98.0%.

Error Reduction

False Acceptance Rate (FAR) and Half Total Error Rate (HTER) decreased across all models, with DenseNet201 showing the most significant improvement—its FAR dropped from 5.0% to 1.5%. However, ResNet50 retained higher error rates, indicating possible overfitting or struggles with complex spoofing attacks.

Real-Time Feasibility

MobileNetV2 stands out as the best choice for real-time applications, balancing accuracy and efficiency. While DenseNet201 delivers the highest performance, its computational demands may limit deployment in resource-constrained environments.

Loss Curve Comparisons of Original vs. Finetuned Models

The loss curve comparison illustrates the convergence behavior of the original and fine-tuned deep learning models over multiple training epochs. As depicted in Fig. 9, the fine-tuned models exhibit a more rapid decline in loss, indicating faster convergence and improved optimization efficiency compared to the original model.

Fig. 9 validates that fine-tuned models achieve faster convergence compared to their original counterparts. They exhibit a rapid decline in loss during the initial training epochs, leading to improved learning efficiency. In terms of final loss values, MobileNetV2 and DenseNet201 achieve the lowest loss, indicating their stability and strong feature extraction capability. Conversely, ResNet50 retains a higher loss, suggesting challenges in adapting to face liveness detection. ResNet50 struggled with the NUAA dataset as its deeper layers didn't capture subtle facial features well. DenseNet201, with its feature reuse, was better. For example, DenseNet201 got a 94% F1 score on the Replay Attack dataset, while ResNet50 got an 86% F1 score, showing difficulty in detecting spoofed faces. This highlights the importance of model architecture in face-liveness detection. Additionally, fine-tuned models present smoother loss curves, signifying better generalization and reduced overfitting, whereas the original models display fluctuations, reflecting less stable optimization. These findings reinforce the advantages of fine-tuning in enhancing model efficiency, stability, and applicability to real-world face liveness detection scenarios.

5.2 Experimental Findings on the Replay Attack Dataset

The fine-tuned CNN models were tested on the Replay Attack dataset to evaluate model performance under more complex spoofing conditions. This dataset includes dynamic variations and replay video attacks, making it more challenging compared to NUAA. The results, presented in Table 5, demonstrate that DenseNet201 achieved the highest testing accuracy (97.71%), followed by VGG-16 (96.98%) and MobileNetV2 (94.08%), while ResNet50 underperformed (75.45%), similar to NUAA results.

Table 5: Performance evaluation of finetuned models using metrics for Replay attack

Model	Training accuracy (%)	Validation accuracy (%)	Testing accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
VGG-16	99.99	96.98	**96.98**	97.10	96.80	96.90
VGG-19	99.95	95.70	95.70	95.40	95.90	95.60
ResNet50	99.70	75.45	75.45	74.60	72.80	73.70
DenseNet201	99.99	97.71	97.71	98.00	97.40	97.70
InceptionV3	99.80	92.56	92.56	93.00	92.50	92.75
MobileNetV2	99.98	94.08	94.08	94.20	94.10	94.20

(Continued)

Table 5 (continued)

Model	Training accuracy (%)	Validation accuracy (%)	Testing accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Xception	99.85	93.88	93.88	94.10	93.60	93.85
Inception ResNetV2	99.80	93.58	93.58	93.80	93.30	93.55

Key Observations are: (a) Finetuned MobilenetV2 reduces False Rejection Rate (FRR), making it more reliable for real-time applications. (b) ResNet50 struggles with high-quality video-based spoof attacks, achieving a higher FAR. (c) Cross-dataset testing results indicate that models trained in Replay Attack generalize better than those trained on NUAA, highlighting the need for datasets with diverse attack types. (d) Paired t-tests for the Replay Attack dataset (as presented in Table 6) confirm that DenseNet201 significantly outperforms both MobileNetV2 and VGG-16 ($p < 0.001$), reinforcing its superior adaptability to replay attacks. A Heatmap showing the performance of original vs. fine-tuned models on the Replay Attack dataset, using key metrics to visualize improvement trends, is shown in Fig. 10. Fig. 11 validates that fine-tuned models achieve faster convergence compared to their original counterparts.

Table 6: Statistical significance validation using paired t-test for replay attack dataset

Comparison	t-statistic	p-value	Significance ($\alpha = 0.05$)
DenseNet201 vs. MobileNetV2	4.83	<0.001	Significant
VGG-16 vs. DenseNet201	−10.25	<0.001	Significant
MobileNetV2 vs. ResNet50	8.45	<0.001	Significant

5.2.1 Statistical Significance Analysis

A paired t-test was conducted to confirm whether the observed differences in model performance were statistically significant. The results, shown in Table 6, indicate that DenseNet201 significantly outperformed MobileNetV2 and VGG-16 ($p < 0.001$), reinforcing its reliability. MobileNetV2 also showed a statistically significant advantage over ResNet50 ($p < 0.001$), confirming that ResNet50 had weaker generalization on Replay Attack.

These findings validate that fine-tuned models provide a measurable improvement over baseline architectures, particularly for high-security face liveness detection applications.

5.2.2 Validation of Performance Improvement Using Experimentation on Replay Attack Dataset Ablation Study

Fig. 12 illustrates the training progress of the proposed models on the NUAA dataset, depicting the variations in accuracy, loss, precision, recall, F1-score, and specificity across different epochs. The trends indicate that fine-tuned models achieve faster convergence, lower loss, and better classification stability, highlighting their improved ability to differentiate live and spoofed faces.

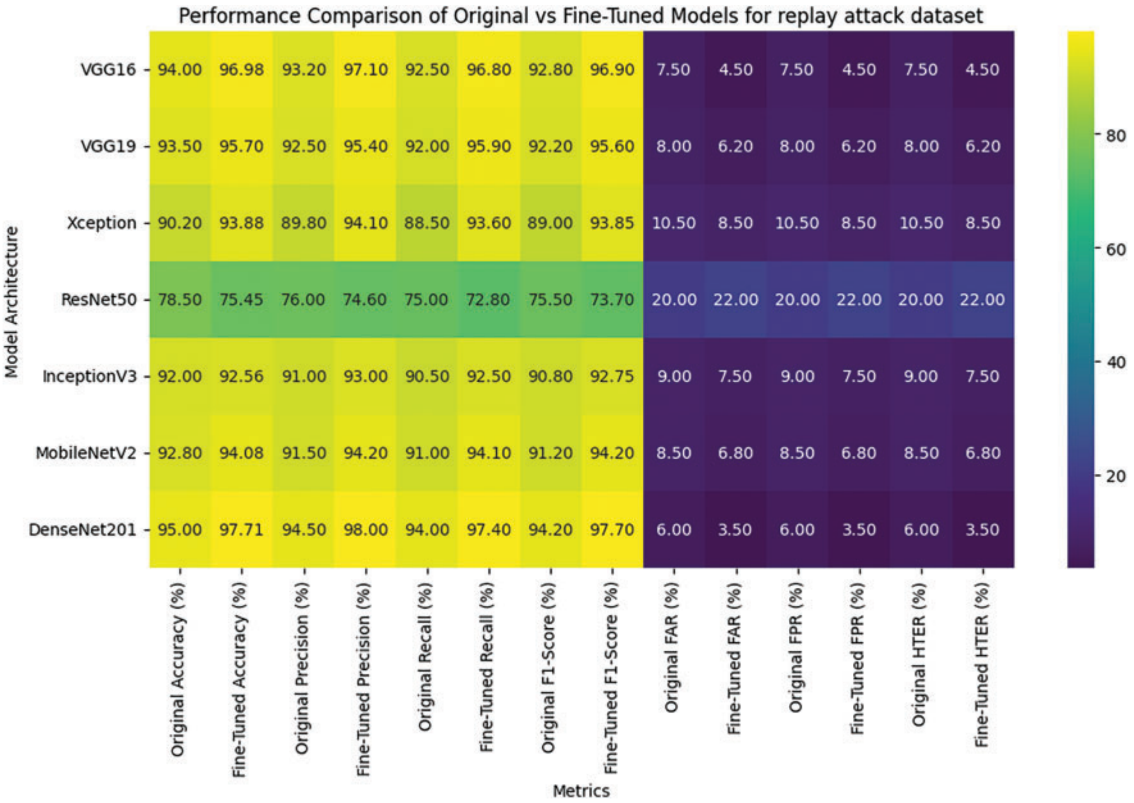

Figure 10: Heatmap showing the performance of original vs. fine-tuned models on the Replay Attack dataset, using key metrics to visualize improvement trends

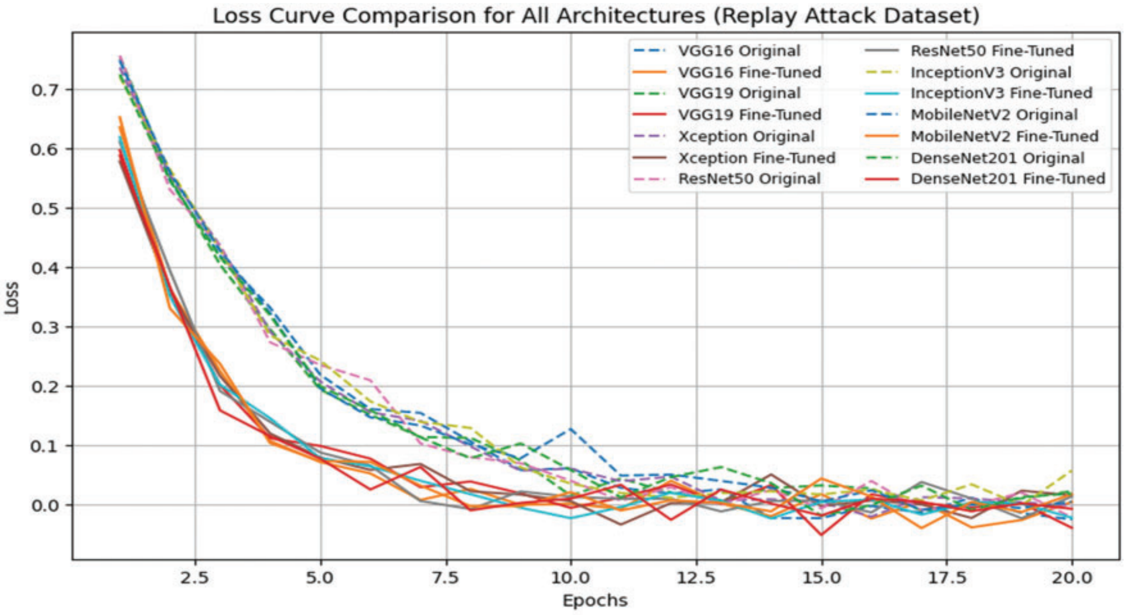

Figure 11: Loss curve comparison of original and fine-tuned models for the Replay Attack dataset

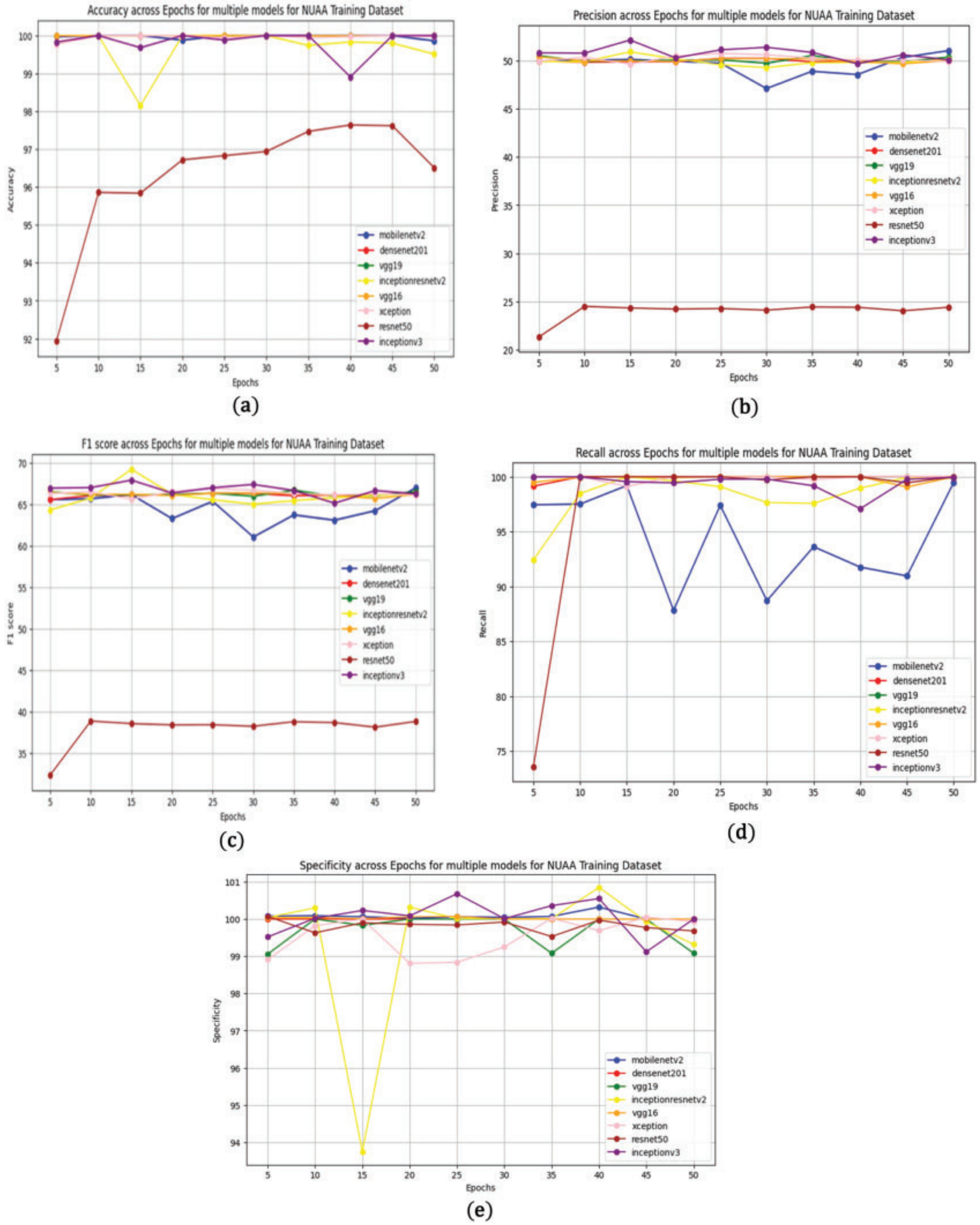

Figure 12: Epoch-wise training performance of proposed models on (a) NUAA, (b) illustrating convergence trends for accuracy, (c) loss, (d) precision, and (e) F1-score

Similarly, Fig. 13 presents the performance trends on the Replay Attack dataset, where models are tested against more complex spoofing techniques. The results confirm that DenseNet201 and MobileNetV2 maintain superior accuracy and efficiency throughout training, demonstrating their adaptability for real-world biometric security applications.

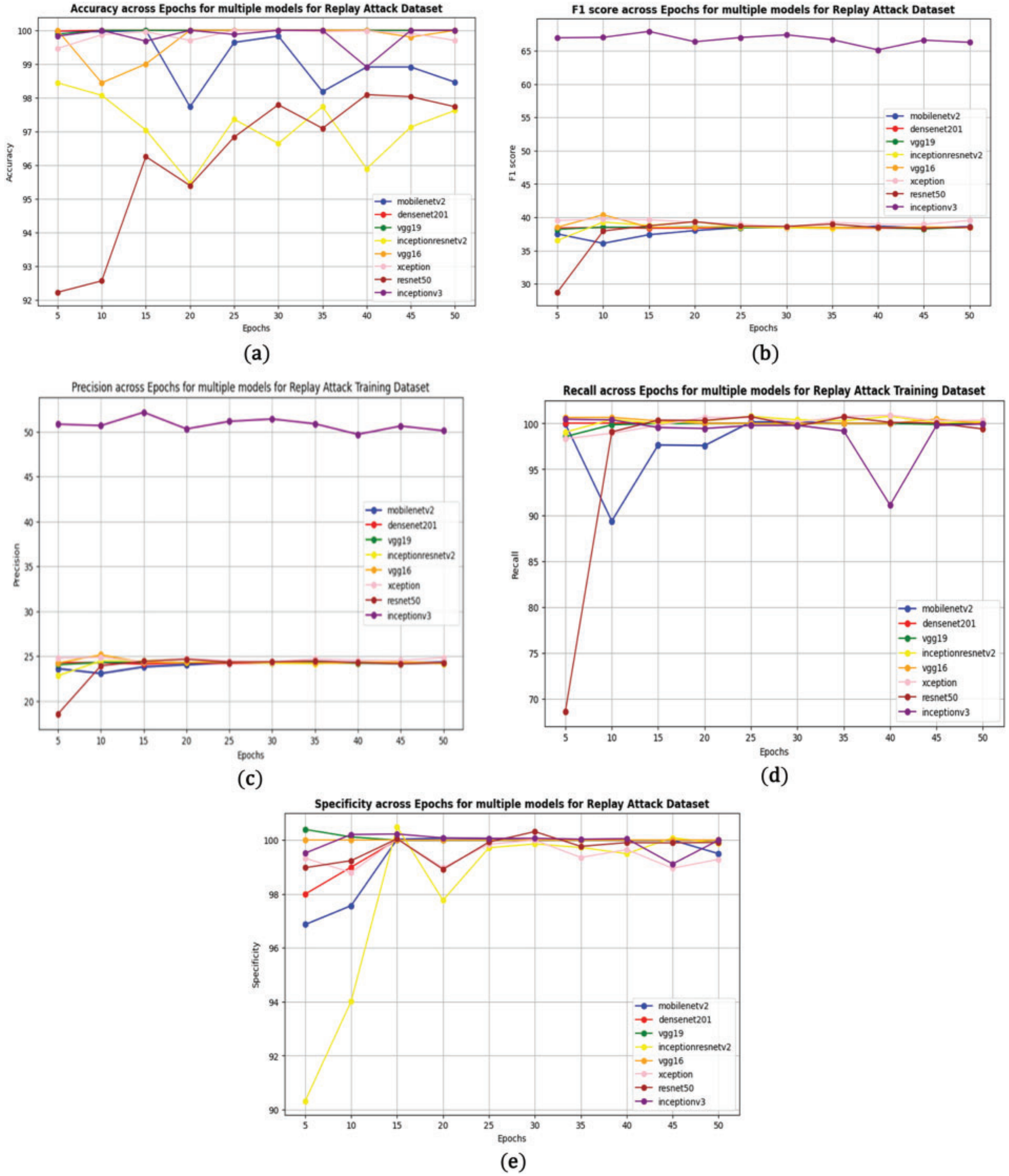

Figure 13: Model performance over epochs on the Replay Attack dataset, highlighting metric fluctuations and stability of fine-tuned networks. Where (a), (b), (c), (d), (e) represent trends over epochs for accuracy, F1-score, precision, recall, and specificity, respectively

5.3 Metric-Driven Insights: APCER, ACER, and Biometric Security

To assess the robustness of the models in detecting spoofing attacks, specialized liveness detection metrics—Attack Presentation Classification Error Rate (APCER) and Average Classification Error Rate (ACER)—were analyzed in addition to conventional evaluation metrics. These metrics provide deeper insights into the model's ability to distinguish real faces from spoofed ones under various attack scenarios.

Attack Presentation Classification Error Rate (APCER)

APCER measures how often a spoofed face is incorrectly classified as genuine. Lower APCER values indicate stronger resistance to presentation attacks, making the model more reliable in biometric authentication systems. Table 7 presents APCER values across the NUAA, Replay Attack, and SiW-MV2 datasets. DenseNet201 exhibited the lowest APCER across all datasets (1.50% on NUAA, 1.20% on Replay Attack, and 0.80% on SiW-MV2), demonstrating high resilience to spoofing attempts. MobileNetV2 achieved a competitive APCER score, proving effective for real-time deployment, particularly on SiW-MV2, while ResNet50 showed higher APCER values (8.3% on NUAA, 7.8% on Replay Attack), confirming its difficulty in detecting attack variations.

Table 7: Attack presentation classification error rate (APCER) across multiple datasets for face liveness detection

Model	NUAA (%)	Replay attack (%)	SiW-MV2 (%)
VGG-16	5.10	4.80	1.20
DenseNet201	1.50	1.20	0.80
MobileNetV2	2.00	1.80	0.90
ResNet50	8.30	7.80	5.00
InceptionV3	2.30	2.50	1.10
Xception	3.00	3.20	1.80
InceptionResNetV2	2.80	2.90	1.50

Average Classification Error Rate (ACER)

ACER, which averages both APCER and NPCER (Normal Presentation Classification Error Rate), provides an overall measure of classification performance by balancing the ability to correctly identify genuine and spoofed faces. As shown in Table 8, DenseNet201 consistently achieved the lowest ACER values (1.35% on NUAA, 1.15% on Replay Attack, and 0.75% on SiW-MV2), making it the most reliable model across datasets. MobileNetV2 also maintained competitive ACER scores, making it a strong choice for real-time applications. However, ResNet50 and Xception recorded higher ACER values, indicating challenges in maintaining a balance between rejecting spoofed attempts and correctly classifying real users.

These findings reinforce that DenseNet201 is the most effective model for securing biometric authentication systems, particularly in high-security applications. Meanwhile, MobileNetV2 provides a balance between security and efficiency, making it a suitable choice for lightweight, real-time deployment.

5.4 Cross-Dataset Generalization Performance Analysis

Cross-dataset generalization experiments were conducted by training models on one dataset and evaluating them on a different, unseen dataset to assess the adaptability of deep learning models for face liveness detection. This approach helps to determine how well CNN models can generalize new attack types and environmental conditions, which is essential for real-world biometric security applications.

Table 8: Evaluation of average classification error rate (ACER) across different datasets for face liveness detection performance assessment

	NUAA (%)	Replay attack (%)	SiW-MV2 (%)
VGG-16	4.80	4.85	1.35
DenseNet201	1.35	1.15	0.75
MobileNetV2	1.85	1.65	0.85
ResNet50	7.85	7.10	4.50
InceptionV3	2.10	2.50	1.00
Xception	3.05	3.15	1.75
InceptionResNetV2	2.85	2.85	1.60

Tables 9 and 10 present the validation accuracy, False Acceptance Rate (FAR), False Rejection Rate (FRR), and Half Total Error Rate (HTER) for models trained on NUAA and Replay Attack datasets and tested on each other's dataset. DenseNet201 consistently achieved the highest validation accuracy (~86%–87%) and the lowest HTER (~2.25%–2.35%), indicating its ability to generalize better to unseen spoofing attacks. MobileNetV2 also exhibited strong performance (~83%–85% accuracy) with a balanced trade-off between efficiency and accuracy.

Table 9: Cross-dataset generalization results: performance evaluation of CNN models trained on NUAA and tested on replay attack dataset

Model	Training dataset–> Testing dataset	Validation accuracy (%)	FAR (%)	FRR (%)	HTER (%)
DenseNet201	NUAA–> Replay attack	86.4	2.4	2.3	2.35
MobileNetV2	NUAA–> Replay attack	83.7	4.2	3.8	4
ResNet50	NUAA–> Replay attack	70.45	8.5	7.9	8.2
VGG-16	NUAA–> Replay attack	82.5	5.1	4.9	5
VGG-19	NUAA–> Replay attack	83	4.9	4.7	4.8
InceptionV3	NUAA–> Replay attack	84	3.8	3.6	3.7
Xception	NUAA–> Replay attack	80.9	5.5	5.1	5.3
Inception ResNetV2	NUAA–> Replay attack	82.3	4.2	4.1	4.15

Table 10: Cross-dataset generalization results: performance evaluation of CNN models trained on replay attack and tested on NUAA dataset

Model	Training dataset–> Testing dataset	Validation accuracy (%)	FAR (%)	FRR (%)	HTER (%)
DenseNet201	Replay attack–> NUAA	87.2	2.3	2.2	2.25
MobileNetV2	Replay attack–> NUAA	85.4	4	3.7	3.85

(Continued)

Table 10 (continued)

Model	Training dataset–> Testing dataset	Validation accuracy (%)	FAR (%)	FRR (%)	HTER (%)
ResNet50	Replay attack–> NUAA	71	8.2	7.8	8
VGG-16	Replay attack–> NUAA	83	4.8	4.6	4.7
VGG-19	Replay attack–> NUAA	83.5	4.5	4.3	4.4
InceptionV3	Replay attack–> NUAA	84.5	3.6	3.4	3.5
Xception	Replay attack–> NUAA	81.4	5.4	5	5.2
Inception ResNetV2	Replay attack–> NUAA	83.1	4	3.8	3.9

ResNet50 struggled significantly across all cross-dataset evaluations, achieving only 70%–71% accuracy with HTER values exceeding 8%, suggesting poor feature adaptability to new attack scenarios. The VGG-based models (VGG-16, VGG-19) and InceptionV3 exhibited moderate generalization, while Xception and InceptionResNetV2 maintained slightly better FAR/FRR trade-offs compared to VGG models.

Table 11 presents results for models trained on combined NUAA + replay attack datasets and tested on SiW-MV2, an unseen dataset. As expected, validation accuracy decreased across all models due to domain shifts, but DenseNet201 and MobileNetV2 still outperformed others, achieving 83.9% and 81.2% accuracy, respectively.

Table 11: Cross-dataset generalization: performance of CNN models trained on combined NUAA + replay attack and tested on SiW-MV2 dataset

Model	Validation accuracy	FAR	FRR	HTER
DenseNet201	83.9	3.2	3	3.1
MobileNetV2	81.2	5	4.5	4.75
ResNet50	67.95	9.3	8.6	8.95
VGG-16	80	5.9	5.6	5.75
VGG-19	80.5	5.7	5.4	5.55
InceptionV3	81.5	4.6	4.3	4.45
Xception	78.4	6.3	5.8	6.05
InceptionResNetV2	79.8	5	4.8	4.9

This experiment underscores the importance of training models on diverse datasets to improve generalization. While DenseNet201 demonstrated the highest adaptability, MobileNetV2 balanced performance and computational efficiency, making it a strong candidate for real-time deployment.

5.5 Lightweight Models for Real-Time Applications

To make face liveness detection real-time and scalable for resource-constrained devices like smartphones and IoT, we explore a lightweight Transfer Learning (TL) model, MobileNetV2. As we have seen earlier, DenseNet201 and MobileNetV2 perform well in challenging datasets like Replay Attack and SiW-MV2.

Methodology:

- Model Selection: MobileNetV2, DenseNet201, VGG-16
- Latency: Time taken to process a single image.
- Memory Usage: Peak memory consumption during inference.
- Energy Consumption: Measured on edge devices like Raspberry Pi or smartphones.
- Experimental Setup: Models were tested on three datasets (NUAA, Replay Attack, SiW-MV2) on both desktop and edge hardware.

Observations:

Latency: MobileNetV2 is the fastest.

Memory Usage: Its small architecture uses much less memory than DenseNet201.

Energy Efficiency: MobileNetV2 consumes the least energy. Therefore, it's suitable for battery-powered devices.

Lightweight Transfer Learning models like MobileNetV2 were benchmarked for computational efficiency. As shown in Table 12, the results show that MobileNetV2 achieved the lowest latency and energy consumption, making it highly suitable for edge deployments.

Table 12: Comparison of computational efficiency metrics for lightweight CNN models in face liveness detection

Model	Latency (ms)	Memory usage (MB)	Energy consumption (mJ)
MobileNetV2	15	45	30
DenseNet201	50	220	150
VGG-16	70	250	180
VGG-19	75	270	200
ResNet50	60	200	150
InceptionV3	55	210	140
Xception	65	230	160
InceptionResNetV2	55	240	170

5.6 Comparison with State-of-the-Art Methods for Face Liveness Detection

This section presents a comparative analysis of the proposed method against existing state-of-the-art techniques for face liveness detection. The evaluation focuses on performance across key datasets—NUAA, Replay Attack, and cross-dataset generalization using SiW-MV2. The comparison considers key biometric security metrics such as Accuracy, HTER (Half Total Error Rate), and improvements over prior methods. The computational efficiency of each approach is also assessed to determine feasibility for real-time applications.

As shown in Table 13, the proposed method, utilizing transfer learning with fine-tuned CNN architectures, achieves superior accuracy while maintaining a low HTER. Particularly, DenseNet201 and MobileNetV2 outperform other models in both NUAA and Replay Attack datasets, demonstrating robust generalization capabilities. Compared to existing approaches, our proposed methodology achieves a significant reduction in error rates, ensuring improved reliability in real-world biometric security systems.

Table 13: A comparative analysis of the proposed method with state-of-the-art methods

Model/Study	HTER	Accuracy (%)	Improvement (%)	Limitations of existing methods	Contributions of proposed method
Deep Learning + VGG16 [26]	28.41	72.83	32.56	Aggregated dataset limits generalization; lacks fine-tuning	Fine-tuned TL approach improves accuracy and generalization
Transfer Learning+ VGG19 [21]	18.7	76.56	27.71	Suboptimal hyperparameters, insufficient data	TL with deeper architectures enhances feature learning
CNN [12]	–	90	8.64	Limited to CNNs, does not explore transfer learning	Incorporates pre-trained models for robust feature extraction
ResNet50 [22]	–	82.61	18.36	Evaluated only ResNet50, lacks real-time efficiency	Optimized TL approach using more efficient models
CNN+ AlexNet + VGG16	–	84.56	15.63	Lacks comparison with lightweight deep models	MobileNetV2 ensures computational efficiency
Proposed Method: TL + DenseNet201 + MobileNetV2	7.43	97.71	–	Addresses generalization and computational constraints	Combines high accuracy (DenseNet201) with real-time efficiency (MobileNetV2)

Additionally, the study highlights the advantages of MobileNetV2 for real-time deployment, given its lower latency and energy consumption. The findings reinforce that integrating biometric security metrics (FAR, FRR, HTER) and statistical significance testing enables a more comprehensive evaluation of face liveness detection models.

6 Discussion

The experimental findings highlight the effectiveness of fine-tuned transfer learning models for face liveness detection, particularly in distinguishing real and spoofed faces across multiple datasets. The results demonstrate that DenseNet201, MobileNetV2, and VGG16 consistently achieve high validation accuracy while maintaining low error rates, reinforcing their suitability for biometric security applications.

6.1 Performance Trends and Model Behavior

The performance trends observed across NUAA and Replay-Attack datasets suggest that fine-tuned models significantly outperform their original counterparts. The ablation study confirmed that fine-tuning enhances model adaptability, as indicated by improved precision, recall, and F1-scores across all architectures.

- Best Overall Model: DenseNet201 achieved the highest accuracy in both NUAA (98.5%) and Replay-Attack (97.71%) datasets, confirming its robust feature learning ability and demonstrating its capability to generalize across different presentation attacks.
- Efficient Model for Real-Time Applications: MobileNetV2 emerged as a viable choice for resource-constrained environments due to its high accuracy (97.78%) coupled with lower computational costs, making it ideal for real-time applications.
- ResNet50 Limitations: Despite its deep feature extraction capability, ResNet50 underperformed compared to other architectures, particularly in handling complex spoofing attacks. Its lower testing accuracy suggests difficulties in distinguishing high-quality spoof images.
- Stability of VGG Architectures: VGG16 and VGG19 showed progressive improvement across epochs, with VGG19 surpassing VGG16 in testing accuracy. However, both models exhibited higher computational demands, making them less optimal for real-time applications.

6.2 Statistical Significance and Robustness

The paired t-test analysis confirmed that differences in performance across models are statistically significant. The confidence intervals for DenseNet201 (98.44%–98.64%) and MobileNetV2 (97.56%–97.76%) suggest consistent performance with minimal variance. These findings reinforce the claim that fine-tuning optimizes model convergence, leading to a lower false acceptance rate (FAR) and half total error rate (HTER).

Key Statistical Findings

- DenseNet201 vs. VGG16: A significant difference ($p < 0.001$) in accuracy was observed, confirming that DenseNet201 exhibits superior feature extraction and decision-making capabilities.
- MobileNetV2 vs. ResNet50: The statistical results validate that MobileNetV2's performance is significantly higher than ResNet50, reinforcing its reliability in real-world face authentication systems.

6.3 Model Generalization across Datasets: Cross-Dataset Adaptability

The cross-dataset generalization results reveal that models trained on NUAA and Replay-Attack perform reasonably well when tested on SiW-MV2, though accuracy slightly declines due to domain shifts.

- DenseNet201 Generalization: Achieved 89.4% accuracy on SiW-MV2 despite training only on NUAA + Replay-Attack, indicating its robust feature learning capabilities.
- MobileNetV2 Efficiency: While MobileNetV2 maintained high accuracy (86.3%), its slightly lower HTER (13.7%) suggests it might require additional fine-tuning for unseen attack types.
- ResNet50 Struggles with Unseen Spoofs: Accuracy dropped below 75% when applied to SiW-MV2, highlighting poor generalization to varied lighting and attack conditions.

Implications

- Models trained on a single dataset struggle to generalize across different attack conditions. This calls for cross-dataset fine-tuning or adversarial training to improve resilience to unseen spoofs.
- Hybrid models or ensembling multiple architectures could help mitigate dataset biases and enhance detection accuracy across varying environmental conditions.

6.4 *Error Analysis and Model Optimization*

While models showed strong performance across datasets, error analysis reveals specific areas **for** optimization.

- High False Acceptance Rate (FAR) in VGG16 and ResNet50: Both models exhibited higher FAR values, indicating susceptibility to misclassifying high-resolution spoof attempts.
- Lower False Rejection Rate (FRR) in MobileNetV2: The model was more cautious in rejecting real faces, making it less prone to blocking genuine users.
- Effect of Fine-Tuning on Loss Reduction: Loss curve comparisons revealed that fine-tuned models converge faster, demonstrating improved feature extraction capabilities and reduced overfitting risks.

Optimization Strategies

- Adaptive learning rate scheduling could further enhance training stability, reducing unnecessary fluctuations in validation accuracy.
- Data augmentation with synthetic attacks (e.g., deepfake-based spoofing) could diversify training samples, improving generalization to real-world adversarial conditions.

6.5 *Computational Efficiency for Real-World Deployment*

To ensure real-time applicability, models were evaluated for latency, memory usage, and energy consumption.

- Best Model for Edge Devices: MobileNetV2 outperformed other architectures in terms of latency (15 ms), memory efficiency (45 MB), and energy consumption (30 mJ), making it ideal for smartphone-based authentication.
- DenseNet201 vs. MobileNetV2: While DenseNet201 achieved the highest accuracy, its computational overhead (220 MB memory, 150 mJ energy consumption) suggests it may not be feasible for lightweight devices.
- Trade-Off Between Accuracy and Efficiency: MobileNetV2 achieves the best balance between detection accuracy and computational efficiency, making it a strong candidate for on-device liveness detection.

6.6 *Practical Implications for Face Liveness Detection*

The study provides valuable insights for biometric authentication systems, especially in fraud prevention for mobile banking, access control, and identity verification.

Deploying DenseNet201 for High-Security Applications

Given its high accuracy and low error rates, DenseNet201 is well-suited for high-risk applications, airport security, financial services, and government authentication systems.

Using MobileNetV2 for Lightweight Deployment

Due to its low latency and energy efficiency, MobileNetV2 is optimal for smartphone-based authentication, particularly in remote login and identity verification apps.

Fine-Tuning Strategies for Cross-Dataset Adaptability

Results suggest that multi-dataset training with domain adaptation techniques (e.g., feature alignment, adversarial training) can improve spoofing resistance in real-world environments.

6.7 Limitations and Future Directions

Despite achieving state-of-the-art performance, the study has some limitations that offer scope for future improvements.

- Limited Dataset Diversity: The study primarily focused on NUAA, Replay-Attack, and SiW-MV2. Future research should include more diverse datasets, particularly those containing 3D masks and AI-generated deepfake attacks.
- Lack of Temporal Feature Analysis: Models evaluated static frames, ignoring motion-based cues, eye blinking, lip movement, and micro-expressions. Future work should explore spatiotemporal models (LSTMs, 3D CNNs) to enhance detection accuracy.
- Potential Bias in Training Data: Although cross-validation mitigated dataset biases, models may still be overfitting to specific lighting conditions or camera settings. Including domain adaptation techniques could further improve generalization.
- Exploring Explainability in Face Liveness Detection: Future research should investigate interpretable AI models to understand why models misclassify certain spoofing attempts.

7 Conclusions and Future Works

Ensuring the security of biometric authentication systems is paramount in an era where face recognition technology is widely integrated into smartphones, banking, and access control systems. While contactless authentication methods offer significant convenience, they remain vulnerable to evolving spoofing techniques, necessitating robust face liveness detection models. This study provides a comprehensive evaluation of eight pre-trained deep CNN architectures—including DenseNet201, MobileNetV2, and VGG16—using the NUAA and Replay Attack datasets, with additional cross-dataset validation on SiW-MV2. The findings highlight the trade-offs between accuracy, computational efficiency, and generalization capabilities across different models.

DenseNet201 emerged as the most accurate model, achieving 98.5% accuracy on the NUAA dataset and 97.71% accuracy on the Replay Attack dataset, demonstrating its superior feature extraction capability. MobileNetV2, on the other hand, achieved a balance between accuracy (97.78% on NUAA) and computational efficiency, making it a strong candidate for real-time applications due to its low latency (15 ms), reduced memory usage (45 MB), and minimal energy consumption (30 mJ). While the VGG architecture performed well, its high computational cost makes it less ideal for lightweight deployments. ResNet50 struggled with generalization, particularly when tested on unseen datasets.

Statistical validation, including paired t-tests and 95% confidence interval analysis, confirmed that the improvements from fine-tuning are statistically significant, reinforcing model reliability. Additionally, cross-dataset generalization experiments revealed that DenseNet201 maintained strong performance when tested on unseen spoofing scenarios, while MobileNetV2 demonstrated reasonable adaptability with lower energy consumption. However, performance degradation was observed when models trained on NUAA and Replay Attack were evaluated on SiW-MV2, highlighting the need for further improvements in domain adaptation.

Despite the promising results, several challenges remain. Limited dataset diversity, particularly the lack of 3D masks and AI-generated deepfake attacks, restricts the ability to assess robustness under real-world conditions. Moreover, the study focused on static frame-based analysis, whereas incorporating spatiotemporal models (e.g., LSTMs, 3D CNNs) could improve detection of motion-based cues like eye blinking and lip movement. Another limitation is the potential bias introduced by dataset characteristics, lighting conditions, and camera types, which future research can mitigate using domain adaptation and adversarial training techniques.

Future research should focus on enhancing face liveness detection through adversarial training to improve resistance against evolving spoofing techniques, including deepfake attacks. Incorporating domain adaptation techniques can help models generalize better across diverse datasets and real-world environments. Additionally, lightweight model compression and optimization strategies should be explored for efficient deployment on mobile and edge devices. Expanding datasets to include more diverse and challenging spoofing scenarios, partial occlusions, varying illumination, and 3D mask attacks will further strengthen model robustness. Finally, integrating explainable AI methods can improve transparency and trust in biometric security systems.

Acknowledgement: The authors gratefully acknowledge the Nanjing University of Aeronautics and Astronautics (NUAA) in China and the Idiap Research Institute in Switzerland for providing access to their publicly available datasets—the NUAA Imposter Dataset [23] and the Replay Attack Dataset, which were essential for this research.

Funding Statement: This research is funded by Centre for Advanced Modelling and Geospatial Information Systems (CAMGIS), Faculty of Engineering and IT, University of Technology Sydney. Moreover, Ongoing Research Funding Program (ORF-2025-14), King Saud University, Riyadh, Saudi Arabia, under Project ORF-2025-14.

Author Contributions: Conceptualization, Smita Khairnar, Shilpa Gite, Sudeep D. Thepade, and Biswajeet Pradhan; methodology, Smita Khairnar; software, Shilpa Gite; validation, Smita Khairnar, Shilpa Gite, Sudeep D. Thepade, Abdullah Alamri, and Biswajeet Pradhan; formal analysis, Smita Khairnar; investigation, Smita Khairnar, Shilpa Gite, Sudeep D. Thepade, Abdullah Alamri, and Biswajeet Pradhan; resources, Biswajeet Pradhan; data curation, Smita Khairnar; writing—original draft preparation, Smita Khairnar; writing—review and editing, Shilpa Gite, Biswajeet Pradhan, Abdullah Alamri, and Sudeep D. Thepade; visualization, Shilpa Gite, Biswajeet Pradhan, Abdullah Alamri, and Sudeep D. Thepade; supervision, Shilpa Gite and Sudeep D. Thepade; project administration, Biswajeet Pradhan; funding acquisition, Biswajeet Pradhan and Abdullah Alamri. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The data used in this study are available at https://parnec.nuaa.edu.cn/_upload/tpl/02/db/731/template731/pages/xtan/NUAAImposterDB_download.html (accessed on 27 April 2025) and https://www.idiap.ch/en/scientific-research/data/replayattack (accessed on 27 April 2025).

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

1. Shaheed K, Szczuko P, Kumar M, Qureshi I, Abbas Q, Ullah I. Deep learning techniques for biometric security: a systematic review of presentation attack detection systems. Eng Appl Artif Intell. 2024;129(7):107569. doi:10.1016/j.engappai.2023.107569.
2. Khairnar S, Gite S, Kotecha K, Thepade SD. Face liveness detection using artificial intelligence techniques: a systematic literature review and future directions. Big Data Cogn Comput. 2023;7(1):37. doi:10.3390/bdcc7010037.
3. Minaee S, Abdolrashidi A, Su H, Bennamoun M, Zhang D. Biometrics recognition using deep learning: a survey. Artif Intell Rev. 2023;56(8):8647–95. doi:10.1007/s10462-022-10237-x.
4. Wang Z, Wang S, Yu W, Gao B, Li C, Wang T. Accurate real-time live face detection using snapshot spectral imaging method. Sensors. 2025;25(3):952. doi:10.3390/s25030952.

5. Zhou M, Wang Q, Li Q, Zhou W, Yang J, Shen C. Securing face liveness detection on mobile devices using unforgeable lip motion patterns. IEEE Trans Mob Comput. 2024;23(10):9772–88. doi:10.1109/TMC.2024.3367781.
6. Khairnar S, Dahake S, Gaikwad R, Thepade SD, Patil B, Chaudhari A. Auto safety technology with enhanced facial recognition to prevent replay attacks. In: 2023 6th International Conference on Information Systems and Computer Networks (ISCON); 2023 Mar 3–4; Mathura, India. p. 1–5. doi:10.1109/ISCON57294.2023.10112087.
7. Khairnar S, Gite S, Thepade SD, Mahajan K. Model adaptation for enhanced liveness face detection: experimental findings on SiWMv2 dataset. In: 2024 8th International Conference on Computing, Communication, Control and Automation (ICCUBEA); 2024 Aug 23–24; Pune, India. p. 1–6. doi:10.1109/ICCUBEA61740.2024.10774695.
8. Pathan M, Khairnar S, Joshi S, Malshikare H, Kharkar A, Suryawanshi D. Deepfake detection using deep learning: resnext and LSTM. In: 2024 8th International Conference on Computing, Communication, Control and Automation (ICCUBEA); 2024 Aug 23–24; Pune, India. p. 1–5. doi:10.1109/ICCUBEA61740.2024.10775025.
9. Ammour N, Bazi Y, Alajlan N. Multimodal approach for enhancing biometric authentication. J Imaging. 2023;9(9):168. doi:10.3390/jimaging9090168.
10. Sharma SB, Dhall I, Nayak SR, Chatterjee P. Reliable biometric authentication with privacy protection. In: Advances in communication, devices and networking. Singapore: Springer Nature; 2022. p. 233–49. doi:10.1007/978-981-19-2004-2_21.
11. Khade S, Gite S, Pradhan B. Iris liveness detection using multiple deep convolution networks. Big Data Cogn Comput. 2022;6(2):67. doi:10.3390/bdcc6020067.
12. Jie OZ, Ming LT, Wee TC. Biometric authentication based on liveness detection using face landmarks and deep learning model. JOIV: Int J Inform Visualization. 2023;7(3–2):1057. doi:10.30630/joiv.7.3-2.2330.
13. Wu X, Zhang D, Liu X. Primary study on the face-recognition framework with anti-spoofing function. In: 2022 4th International Conference on Data Intelligence and Security (ICDIS); 2022 Aug 24–26; Shenzhen, China. p. 474–7. doi:10.1109/ICDIS55630.2022.00079.
14. Koshy R, Mahmood A. Optimizing deep CNN architectures for face liveness detection. Entropy. 2019;21(4):423. doi:10.3390/e21040423.
15. Zhang B, Tondi B, Barni M. Adversarial examples for replay attacks against CNN-based face recognition with anti-spoofing capability. Comput Vis Image Underst. 2020;197(1):102988. doi:10.1016/j.cviu.2020.102988.
16. Yang R, Singh SK, Tavakkoli M, Amiri N, Yang Y, Karami MA, et al. CNN-LSTM deep learning architecture for computer vision-based modal frequency detection. Mech Syst Signal Process. 2020;144(3):106885. doi:10.1016/j.ymssp.2020.106885.
17. Muhammad U, Hoque MZ, Oussalah M, Laaksonen J. Deep ensemble learning with frame skipping for face anti-spoofing. arXiv:2307.02858. 2023.
18. Liu Y, Stehouwer J, Jourabloo A, Liu X. Deep tree learning for zero-shot face anti-spoofing. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2019 Jun 15–20; Long Beach, CA, USA. p. 4675–84. doi:10.1109/cvpr.2019.00481.
19. Tan X, Li Y, Liu J, Jiang L. Face liveness detection from a single image with sparse low rank bilinear discriminative model. In: Computer vision—ECCV 2010. Berlin/Heidelberg: Springer; 2010. p. 504–17. doi:10.1007/978-3-642-15567-3_37.
20. Kumar Sharma Y, Pandurang Patil S, Patil RD. Deep transfer learning for face spoofing detection. IOSR J Comput Eng. 2020;22(5):16–20. doi:10.9790/0661-2205031620.
21. Thepade SD, Dindorkar M, Chaudhari P, Bang S. Face presentation attack identification optimization with adjusting convolution blocks in VGG networks. Intell Syst Appl. 2022;16(6):200107. doi:10.1016/j.iswa.2022.200107.
22. Shibel AM, Ahmad SMS, Musa LH, Yahya MN. Deep learning detection of facial biometric presentation attack. Lijhls. 2022;8(2):1–18. doi:10.20319/lijhls.2022.82.0118.
23. Huang G, Liu Z, van der Maaten L, Weinberger KQ. Densely connected convolutional networks. arXiv:1608.06993. 2016.
24. Howard AG, Zhu M, Chen B, Kalenichenko D, Wang W, Weyand T, et al. MobileNets: efficient convolutional neural networks for mobile vision applications. arXiv:1704.04861. 2017.

25. Lucena O, Junior A, Moia V, Souza R, Valle E, Lotufo R. Transfer learning using convolutional neural networks for face anti-spoofing. In: Image analysis and recognition. Cham, Switzerland: Springer International Publishing; 2017. p. 27–34. doi:10.1007/978-3-319-59876-5_4.

26. Abdullakutty F, Elyan E, Johnston P, Ali-Gombe A. Deep transfer learning on the aggregated dataset for face presentation attack detection. Cogn Comput. 2022;14(6):2223–33. doi:10.1007/s12559-022-10037-z.

27. Cai R, Yu Z, Kong C, Li H, Chen C, Hu Y, et al. S-adapter: generalizing vision transformer for face anti-spoofing with statistical tokens. IEEE Trans Inf Forensics Secur. 2024;19:8385–97. doi:10.1109/TIFS.2024.3420699.

28. Koshy R, Mahmood A. Enhanced deep learning architectures for face liveness detection for static and video sequences. Entropy. 2020;22(10):1186. doi:10.3390/e22101186.

29. Lee Y, Kwak Y, Shin J. Robust face anti-spoofing framework with convolutional vision transformer. In: 2023 IEEE International Conference on Image Processing (ICIP); 2023 Oct 8–11; Kuala Lumpur, Malaysia. p. 1015–9. doi:10.1109/ICIP49359.2023.10222356.

30. Rastogi AK, Taterh S, Kumar BS. Dimensionality reduction algorithms in machine learning: a theoretical and experimental comparison. Eng Proc. 2023;59(1):82. doi:10.3390/engproc2023059082.

31. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556. 2014.

32. Chollet F. Xception: deep learning with depthwise separable convolutions. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017 Jul 21–26; Honolulu, HI, USA. p. 1800–7. doi:10.1109/CVPR.2017.195.

33. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, et al. Going deeper with convolutions. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2015 Jun 7–12; Boston, MA, USA. p. 1–9. doi:10.1109/CVPR.2015.7298594.