Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A Hybrid LBP-CNN Feature Fusion Framework Using SVM for Robust Face Anti-Spoofing Attack Detection

NAIMISHA AWASTHI
Dr. A.P.J. Abdul Kalam Technical University
Usha Sharma
Babu Banarasi Das Institute of Technology and Management
Yudhishthir Raut
yudhishthir.raut@sitpune.edu.in
Symbiosis Institute of Technology
Deepak Kumar Jain
Dalian University of Technology

Article

Keywords: Monocular RGB 2D photo print attack, Anti-Spoofing, Liveness Detection, Cross data Generalization, cross/multi source evaluation and Generalization, Robust Biometric Security

Posted Date: November 19th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11511764/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

A Hybrid LBP-CNN Feature Fusion Framework Using SVM for Robust Face Anti-Spoofing Attack Detection

Naimisha Awasthi[1], Usha Sharma[2], Yudhishthir Raut[3*], Deepak Kumar Jain[4*]

[1]Faculty of Engineering and Technology, Dr.APJ Abdul Kalam Technical University, Uttar Pradesh, India.
[2]Department of Information Technology, Babu Banarasi Das Institute of Technology and Management, Uttar Pradesh, India.
[3]Department of Robotics and Automation, Symbiosis Institute of Technology, Maharashtra, India.
[4]School of Artificial Intelligence, Dalian University of Technology, Dalian, China.

*Corresponding author(s). E-mail(s): yudhishthir.raut@sitpune.edu.in; dkj@dlut.edu.cn;
Contributing authors: naimishaawasthi56@gmail.com; usha19792023@gmail.com;

Abstract

Technologies working behind recognition and biometric authentication systems are increasing day by day so as the attacks posed by them. The security is a major concern that makes these systems vulnerable to presentation or anti spoofing attacks like 2D attacks such as printed photo attacks. Existing Face Anti Spoofing research has improved detection capability, but practical deployment still faces challenge such as obtaining sufficiently rich discriminative representations, controlling computational complexity, avoiding dependence on specialized sensing hardware, and maintaining performance when the data distribution changes. To counter these challenges this paper proposes a lightweight monocular RGB anti spoofing attack detection framework that can extract complementary features from a single conventional camera. Multi-Scale Local Binary Pattern provides multi-scale local texture information hierarchical fusing Convolutional Neural Network providing hierarchical learned semantic features and Self-Attention

emphasizing on informative spatial regions. Global Average Pooling providing deep semantic information. Classification performed using a hyperparameter-optimized Support Vector Machine (SVM). The monocular architecture enables feature extraction, feature fusion, and classification components to be independently enhanced while maintaining computational efficiency. Experimental evaluation demonstrates an overall accuracy of 97.26%, with Precision of 99.21%, Recall of 93.52%, Specificity of 99.57%, F1-score of 96.28%, and Balanced Accuracy of 96.54%. The architecture also attains an MCC of 0.9431 and Cohen's Kappa of 0.9421. Cross-dataset evaluation obtains approximately 93% accuracy, demonstrating promising generalization to an independent dataset. The results indicate that combining complementary local texture and hierarchical deep representations provides an effective and computationally efficient solution for RGB-based detection of planar presentation attacks.

Keywords: Keywords: Monocular RGB 2D photo print attack, Anti-Spoofing, Liveness Detection, Cross data Generalization, cross/multi source evaluation and Generalization, Robust Biometric Security

1 Introduction

There is an increase of face recognition and face biometric authentication because of its non contact operation, user convenience, as well as seamless integration with commodity imaging devices like cameras, smartphones, surveillance, laptops, and Internet of Things (IoT) platforms. The main reason of the technology being widely used is because of the widespread advancements in Artificial Intelligence and Computer Vision. The facial authentication system has been deployed in diverse applications including digital bank, healthcare, online examination, intelligent access control systems, mobile device locking. Despite these advantages the biometric systems are a concern of security and vulnerability to presentation attacks or face anti spoofing attack where an attacker impersonates itself as someone else through their face by surpassing the face biometric detection. This impersonation done by attackers is achieved through 2D attacks such as printed photo attacks, cutout print attack, cut-out mask of eyes and mouth wrapped or object mounted print attacks and 3D mask and artifacts [1]. Therefore, Presentation Attack Detection (PAD), also known as face anti-spoofing or face liveness detection, is a vital component for improving the security of face recognition systems. Traditionally PAD methods have mainly relied on handcrafted image as well as texture descriptors, whereas recent approaches increasingly employ deep learning along with multimodal information to learn more discriminative representations. Recent surveys as well as studies have highlighted the continuing importance of robust feature learning, attack diversity, cross-dataset generalization, and practical deployment in face anti-spoofing systems.

Although there have significant developments in face presentation attack detection, several challenges remain. Firstly, some approaches exploit depth, infrared/near-infrared, stereo, or other multimodal sensing technologies to obtain additional information for distinguishing genuine faces from presentation attacks; however, these

works may require additional sensors, specialized acquisition systems, as well as expensive modality-specific hardware, which can increase system complexity and deployment requirements. Secondly, many recent deep learning methods [2] rely on increasingly large or computationally intensive architectures, and a higher cost which can increase money and memory, computational cost, and inference requirements and may limit their aptness for resource-constrained or real-time applications [3], [4]. Thirdly, handcrafted texture descriptors can capture useful local surface as well as micro-texture characteristics, but they generally have limited ability to learn higher-level representations, whereas deep-net models can learn hierarchical features but may not explicitly preserve complementary handcrafted texture information. The limited integration of complementary handcrafted with hierarchical deep representations therefore remains an important research opportunity. In addition, high performance on a source dataset does not necessarily ensure reliable performance on an independent dataset because variations in acquisition environments, illumination, cameras, subjects, image quality, and attack characteristics can introduce distribution shifts [5]. Recent face anti-spoofing research has consequently identified cross-domain along with cross-dataset generalization as an important open challenge. These limitations show the need for a lightweight, monocular RGB-based PAD framework that can obtain complementary discriminative information from a simple monocular camera while reducing dependence on dedicated sensing hardware and computationally heavy architectures. To address these gaps, this study proposes a lightweight and monocular hybrid face presentation attack detection framework based entirely on monocular RGB facial imagery. Unlike approaches that require additional depth or infrared sensing, the proposed framework uses a conventional RGB camera therefore does not depend on specialized multimodal acquisition hardware. The framework integrates Multi-Scale Local Binary Pattern (MS-LBP) descriptors with hierarchical features extracted from a custom convolutional neural network incorporating Self-Attention with Global Average Pooling (GAP). MS-LBP is used to capture complementary multi-scale local texture semantics and traits associated with facial presentation artifacts, while the custom CNN learns hierarchical deep representations from RGB facial images. Self-Attention enhances the representation of informative spatial regions, and GAP produces compact deep feature representations. Features from different CNN levels are subsequently fused with the MS-LBP representation to combine local texture with higher-level information, and the resulting representation is classified using a hyperparameter-optimized Support Vector Machine (SVM). A key contribution of the proposed design is its modularity, in which feature extraction, feature fusion, and classification components can be independently modified or enhanced without redesigning the complete framework. The framework is evaluated using multiple performance measures and independent cross-dataset testing.

The remainder of this paper is organized as follows. Section 2 presents the related work on face presentation attack detection, including handcrafted texture-based methods, deep learning approaches, hybrid feature-learning methods, multimodal approaches, and recent studies on cross-dataset and cross-domain generalization. Section 3 describes the proposed methodology, including MS-LBP feature extraction, the custom CNN architecture, Self-Attention, Global Average Pooling, hierarchical

feature extraction, feature fusion, and SVM classification, working, evaluation of models, accuracy, Confusion matrix, Precision, Recall, F1- score, Balanced Accuracy, Matthews Correlation Coefficient, Cohen's Kappa, ROC and AUC graph. Section 4 presents and discusses the experimental results, including quantitative performance, confusion-matrix analysis, comparative evaluation, and cross-dataset results. Finally, Section 5 concludes the paper, discusses the limitations of the proposed framework, and presents directions for future research.

2 Literature Review

Existing face anti-spoofing methods can be broadly categorized into six major research directions based on their feature representation, learning strategy, and deployment objective.

2.1 Handcrafted Texture-Based and Classical Machine Learning Approaches

Researches such as [6], [7],[8],[9] studies rely mainly on handcrafted texture descriptors for distinguishing live and spoof faces. Common descriptors include LBP, PCA(Principle Component Analysis) -reduced LBP descriptors, and enhanced handcrafted features extracted after image preprocessing such as DoG(Difference of Gaussians) and Beltrami filtering. Classification is generally performed using SVM or fuzzy-SVM. All this contributes to demonstrate that handcrafted texture descriptors effectively capture spoof artifacts, produce reliable performance with limited training samples. maintain relatively low computational complexity, SVM-based classifiers provide stable decision boundaries. These researches depend heavily on manually engineered features unable to learn high-level semantic facial representations, performance decreases under unseen spoof conditions, most studies report only intra-dataset evaluation, limited robustness against illumination and camera variations.

2.2 Deep CNN and Pretrained deep models Based End-to-End Learning Approaches

Researches such as [10], [11],[12] presented methods that employ end-to-end CNN architectures for automatic feature learning. Some focus on lightweight CNNs, whereas others utilize pretrained deep networks such as DenseNet201 for real-time biometric authentication. These works contributes to learn hierarchical semantic representations automatically, achieve high accuracy on benchmark datasets, reduce dependency on handcrafted feature engineering, suitable for real-time biometric applications, these researches lacks large annotated datasets are required, CNNs often overfit dataset-specific characteristics limited interpretability, reduced cross-dataset robustness, pure semantic learning may overlook fine spoof texture artifacts.

2.3 Hybrid Texture Deep Learning Feature Fusion

Studies such as [4], [13],[14],[15],[16] combined handcrafted texture descriptors alongside CNN-based semantic features. Texture information (LBP, LTP (Local Ternary

Pattern.), grayscale structural features, color-texture descriptors) is fused with CNN embeddings before classification. Hybrid features outperform individual handcrafted or CNN features. Texture descriptors complement semantic representations. These researches lacks most frameworks still rely on end-to-end neural classifiers, fusion strategies are rarely optimized, feature learning and classification are not explicitly separated.

2.4 Deep Ensemble Learning and Domain Generalization

These papers [17], [18],[19],[20],[21] worked to improve robustness through ensemble learning, adversarial learning, synthetic data generation, adaptive fusion, and domain generalization strategies. Multiple CNN architectures or adversarial frameworks are employed to improve unseen-domain performance. There work contributed to improved cross-domain performance, ensemble learning increases classification robustness. These researches lacks large annotated datasets are required, CNN often overfit dataset-specific characteristics, limited interpretability, reduced cross-dataset robustness, pure semantic learning may overlook fine spoof texture artifacts.

2.5 Hybrid Machine Learning and Ensemble Classification

The approaches such as [22], [23] studied combine handcrafted descriptors and CNN features while comparing multiple machine learning classifiers including SVM, Decision Trees, KNN (K-Nearest Neighbors), Random Forest, and ensemble classifiers to determine optimal classification performance which contributed to demonstrate that classifier selection significantly influences spoof detection, ensemble classifiers reduce false positives and false negatives. These researches lacks focus mainly on classification accuracy, limited attention to feature representation learning. Cross-dataset validation remains insufficient and heavy dependence on dataset-specific optimization.

2.6 Application-Oriented Face Anti-Spoofing Systems

The frameworks such as [24], [25] integrated face anti-spoofing into practical biometric applications such as secure banking systems and authentication platforms. They combine texture analysis, CNN models, and behavioral verification to improve deployment capability that contributed to demonstrate practical deployment feasibility and improve security against presentation attacks. These researches lacks primarily application-specific, limited generalized evaluation. dependence on pretrained deep models.

As seen in over years, face anti-spoofing research has progressed from traditional texture-based methods to deep learning, also recently, hybrid approaches that combine handcrafted and learned features. Although these approaches differ in their underlying principles, they all aim to improve the reliability of biometric authentication against presentation attacks. A clearer understanding of these developments can be obtained by grouping the existing studies according to their feature representation as well as learning strategy rather than discussing them individually. The given table 1 presents a categorized summary of the representative face anti-spoofing approaches discussed

in this section, together with their learning strategies, key strengths, and remaining limitations.

Although the reviewed studies differ in methodology and application, many of them encounter similar challenges. Instead of treating these limitations separately, it is useful to examine them collectively to understand where further improvements are required. The common research gaps identified from the literature are summarized in table 2, together with the corresponding design considerations that motivated the proposed framework.

Hybrid face anti-spoofing methods have become increasingly popular because they combine the strengths of handcrafted texture analysis and deep representation learning. However, existing frameworks differ considerably in how features are extracted, fused, and classified. A direct comparison therefore helps clarify the architectural differences between previous studies and the proposed framework. Rather than comparing only recognition accuracy,3 highlights differences in feature extraction strategy, feature fusion, classifier selection, and evaluation methodology.

The literature reviewed in this section shows that considerable progress has been made in face presentation attack detection. However, opportunities still exist to improve the way texture and semantic information are combined, to separate feature learning from classification, and to evaluate models under more challenging testing conditions. Guided by these observations, the next section presents the proposed framework, which integrates Multi-Scale Spatial LBP, a custom CNN with hierarchical semantic feature extraction, complementary multi-level feature fusion, and a hyperparameter-optimized linear SVM for robust face anti-spoofing.

3 Proposed Methodology

This segment demonstrate the proposed lightweight hybrid face anti-spoofing framework designed to effectively distinguish genuine facial images from face anti spoofing attacks using only conventional RGB images and camera. The proposed methodology integrates MS-LBP descriptors, a custom attention-enhanced CNN, multi-level feature fusion, along with a hyperparameter-optimized SVM classifier into a unified presentation attack detection pipeline. From a Security perspective this section presents the complete methodology of proposed framework curated. The fig. 1 showing the methodology behind this research and fig. 2 represent the working and block view of the proposed model. Also a mathemetical representation of the same proposed model is given in table 4

3.1 Datasets

The Datasets that has been used in this research have been taken from kaggle and are used for Anti-Spoofing Detection and Liveness detection. The following datasets have been considered.

3.1.1 NUAA DATASET

The NUAA database has been picked up for carrying out this research from kaggle. NUAA, or Nanjing University of Aeronautics and Astronautics, discusses datasets for

Table 1: Categorization of Existing Face Anti-Spoofing Approaches

Category	Representative Studies	Primary Features	Learning Strategy	Major Strength	Primary Limitation
Handcrafted Texture-Based Methods	Mohan et al. (2020), Pandey and Tiwari (2025), El-Rashidy et al. (2025), Chakole and Dixit (2025)	LBP, PCA, CCP, handcrafted descriptors	SVM, Fuzzy-SVM, Classical ML	Efficient texture characterization with low computational cost	Unable to learn high-level semantic facial representations and limited cross-dataset robustness
CNN-Based End-to-End Methods	Shinde et al. (2025), Kuznetsov et al. (2025), Hiremath et al. (2025)	Deep CNN semantic features	End-to-end CNN classification	Automatic hierarchical feature learning with high recognition accuracy	Dataset-specific overfitting, limited interpretability, and large training data requirements
Hybrid Texture-Deep Learning Methods	Radad et al. (2025), Rusia and Singh (2022), Singh et al. (2023), Jha et al. (2024), Gupta and Kaur (2025)	LBP, LTP, grayscale features + CNN	Feature-level or decision-level fusion	Combines texture and semantic information for improved spoof detection	Fusion strategies remain relatively simple and generally retain end-to-end neural classification
Deep Ensemble and Domain Generalization Methods	Liu et al. (2023), Lee et al. (2023), Vannurswamy et al. (2024), Muhammad et al. (2024), Qi et al. (2025)	Multiple CNN representations	Ensemble learning, adversarial learning, domain adaptation	Strong cross-domain and unseen attack performance	High computational complexity and limited suitability for lightweight deployment
Hybrid ML and Ensemble Classification Methods	Prasad et al. (2024), Singh et al. (2023)	Handcrafted + CNN features	Multiple ML classifiers and ensemble learning	Improved classification reliability through classifier comparison	Primarily focused on classification accuracy rather than representation learning
Application-Oriented Face Anti-Spoofing Systems	Hamad et al. (2025), Chavan and Salunkhe (2025)	CNN and texture features	Application-specific biometric systems	Practical deployment for authentication systems	Limited standardized evaluation and weak cross-dataset validation

Table 2: Research Gaps Identified from Existing Literature and Motivation for the Proposed Framework

Research Gap	Observation from Existing Literature	Proposed Framework
Limited semantic learning in handcrafted approaches	Texture descriptors effectively capture spoof artifacts but fail to model global facial semantics.	Integrates CNN-based semantic representations with Multi-Scale LBP texture descriptors.
Over-reliance on end-to-end CNN classification	CNN simultaneously performs feature extraction and classification, increasing dataset-specific overfitting.	Decouples feature learning and classification using a hyperparameter-optimized Linear SVM.
Limited complementary feature interaction	Existing hybrid methods often concatenate handcrafted and CNN features without explicitly exploiting complementary relationships.	Employs complementary multi-level feature fusion between handcrafted texture descriptors and hierarchical CNN representations.
Insufficient exploitation of hierarchical CNN representations	Most hybrid methods utilize only the final CNN embedding while discarding intermediate semantic representations.	Preserves intermediate and deep semantic representations together with handcrafted texture information.
High computational complexity	Recent ensemble and domain adaptation methods rely on multiple pretrained networks and computationally expensive optimization.	Uses a lightweight hybrid architecture suitable for practical deployment.
Limited cross-dataset validation	Most studies evaluate models using samples originating from the same dataset.	Evaluates the framework under a strict cross-dataset protocol using independent genuine and spoof datasets.
Poor generalization to unseen spoof conditions	Existing models frequently learn dataset-specific characteristics.	Improves robustness through complementary feature learning and cross-dataset evaluation.

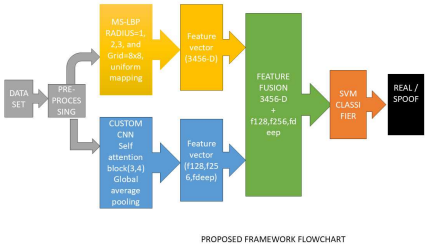


Fig. 1: Methodology of Proposed Architecture

research in computer vision with regard to anti-spoofing through facial recognition. The NUAA Photograph Imposter Database was developed by this university specifically for research into facial liveness detection. The database includes images of actual printed attack attempts and access attempts made by 15 unique individuals using standard web cameras. The NUAA Photograph Imposter Database consists of 5105 cases of authentic faces from individuals as well as 7509 instances of assigned "imposter" attacks made against individuals[26]. Some of the samples can be seen in fig.3.

Table 3: Comparison of Representative Hybrid Face Anti-Spoofing Frameworks

Study	Handcrafted Features	Deep Features	Feature Extraction Strategy	Classifier	Cross-Dataset Evaluation	Main Limitation
Radad et al. (2025)	LBP + LTP	Shallow ECNN	Parallel feature fusion	SVM	No	Limited semantic representation and shallow CNN
Singh et al. (2023)	LBP	CNN	Feature-level fusion	CNN	Limited	CNN performs both feature extraction and classification
Jha et al. (2024)	Texture descriptors	CNN	Feature-level fusion	Deep classifier	No	Mainly optimized for intra-dataset performance
Gupta and Kaur (2025)	LBP	CNN	Hybrid feature extraction	Deep classifier	Limited	End-to-end neural classification with limited generalization analysis
Proposed Framework	**Multi-Scale Spatial LBP**	**Custom CNN**	**Complementary parallel feature extraction with multi-level feature fusion**	**Optimized Linear SVM**	**Yes**	**Designed for lightweight deployment with improved cross-dataset robustness**

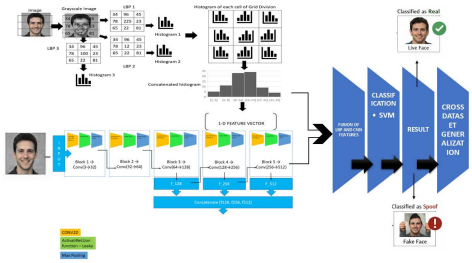


Fig. 2: Working of the proposed model

3.1.2 YALE FACE DATASET

The Yale face database has been picked up for carrying out this research from website named kaggle. The database contains 15 subjects of which total number is 165 GIF

Table 4: Mathematical Representation of the Proposed Hybrid LBP-CNN-SVM Pipeline.

Step / Operation	Mathematical Representation
1. Input & Preprocessing	
Input: Convert to Grayscale and Resize	$\mathbf{I}_r = \text{Resize}(\mathbf{I}, 64 \times 64)$
Preprocessing	$\mathbf{I}_g = \text{Gray}(\mathbf{I})$
2. Feature Extraction	
LBP Extraction: Multi-Scale LBP	$\mathbf{L} = \left[\text{LBP}_{r_1}(\mathbf{I}_g), \ldots, \text{LBP}_{r_3}(\mathbf{I}_g)\right]$
CNN Block 1: Conv $\rightarrow$ Max-Pool	$\mathbf{a}_2 = \text{MaxPool}(\text{LeakyReLU}(\mathbf{C}_1(\mathbf{I}_r; W_1) + b_1))$
CNN Block 2: Conv $\rightarrow$ Max-Pool	$\mathbf{a}_4 = \text{MaxPool}(\text{LeakyReLU}(\mathbf{C}_2(\mathbf{a}_2; W_2) + b_2))$
CNN Block 3 + Attn: Mid-level Feature	$\mathbf{a}_6 = \mathbf{A}_3(\text{LeakyReLU}(\mathbf{C}_3(\mathbf{a}_4; W_3) + b_3))$
Feature $\mathbf{f}_{128}$	$\mathbf{f}_{128} = \text{Flatten}(\text{GAP}(\mathbf{a}_6))$
CNN Block 4 + Attn: Deeper Feature	$\mathbf{a}_8 = \mathbf{A}_4(\text{LeakyReLU}(\mathbf{C}_4(\mathbf{a}_6; W_4) + b_4))$
Feature $\mathbf{f}_{256}$	$\mathbf{f}_{256} = \text{Flatten}(\text{GAP}(\mathbf{a}_8))$
CNN Block 5: Deep Semantic Feature	$\mathbf{a}_9 = \text{LeakyReLU}(\mathbf{C}_5(\mathbf{a}_8; W_5) + b_5)$
Projection Head: Feature Refinement	$\mathbf{f}_{deep} = \text{ReLU}(W_p \cdot \text{GAP}(\mathbf{a}_9) + b_p)$
3. Feature Fusion	
Fusion: Concatenate LBP and Deep Features	$\mathbf{H} = [\mathbf{L}\,;\, \mathbf{f}_{128}\,;\, \mathbf{f}_{256}\,;\, \mathbf{f}_{deep}]$
4. Classification	
SVM Classification: Linear / RBF Decision Boundary	$\mathbf{y} = \text{SVM}(\mathbf{H})$
Output: Real / Spoof Label	$\mathbf{y} \in \{0, 1\}$

images. There are 11 images per subject, one for each of the following facial expressions or configurations that are there in the dataset can be described as center-light, without glasses, happy, left-light, with no glasses, normal, right-light, sad, sleepy, surprised, and wink [27]. Some of the samples can be seen in fig.3.

3.1.3 PHOTO PRINT ATTACK DATASET

The dataset is taken from kaggle that has 3,000+ people used more than 7,000 picture print attacks. Models are trained using a 2D presentation attack dataset to identify printed photo faking in facial recognition systems. Over 3,000 Participants , participating in the project where there is variety in representation of images like gender

and racial diversity in balance and more than 7,000 photo print attacks were carried out on the participants [28]. Some of the samples can be seen in fig 3.



Fig. 3: (a)- NUAA DATASET SAMPLES, (b)- YALE FACE DATASET SAMPLES, (c)-PHOTO PRINT ATTACK DATASET SAMPLES

3.2 PREPPROCESSING

Before the data is processed , it goes through the preprocessing stage as the raw dataset accumulated contains in sufficient data, imbalanced data, might require transformation of data. Data Preprocessing is one of the major step for training the model where it ensures the dataset must go through to cleaning, scaling, and also performing various other techniques before it is finally processed and given to classifier for

classification/prediction [29]. In order to standardize the data each image underwent through certain step for preprocessing before giving it for feature extraction.

Unlike several traditional approaches that apply a common preprocessing pipeline to all feature extraction modules, the proposed framework adopts two complementary parallel preprocessing pipelines. Since handcrafted texture descriptors as well as CNN utilize fundamentally different visual characteristics, preprocessing is specifically optimised according to the requirements of each feature extraction branch. The handcrafted MS-LBP descriptor operates on grayscale intensity information, while the CNN learns hierarchical semantic representations directly from RGB facial images. Consequently, both grayscale and RGB representations are simultaneously preserved throughout the preprocessing stage.

- Acquisition of Face Image
- Conversion to Grayscale
- Resizing of Image
- Normalization of pixel
- Extraction of feature

3.2.1 Acquisition of RGB Face images

The proposed approach is intentionally designed for conventional RGB imaging devices without the need of specialized hardware such as infrared cameras, depth sensors, or structured-light imaging systems. Also due to RGB imaging devices represent the most widely deployed sensing modality in practical biometric authentication systems This design choice significantly enhances the practical applicability of the proposed framework since commodity RGB cameras are readily available in smartphones, laptops, surveillance systems, and embedded Internet of Things (IoT) devices. The use of standard RGB images also facilitates practical deployment by eliminating additional hardware costs while maintaining compatibility with existing face recognition infrastructures.

3.2.2 Conversion to Grayscale

Although CNN-based semantic learning benefits from color information, Local Binary Pattern fundamentally operates on local intensity relationships rather than chromatic information. Consequently, each RGB facial image is converted into a grayscale representation prior to Multi-Scale LBP extraction. LBP encodes binary relationships among neighboring pixel intensities. Consequently, chromatic information contributes minimally to the resulting texture descriptor while unnecessarily increasing computational complexity. Converting RGB images into grayscale preserves the luminance transitions responsible for facial micro-texture characterisation while simultaneously reducing computational overhead. The grayscale intensity value is computed as

$$\text{x,y} = 0.299R + 0.587G + 0.114B$$

Grayscale intensities preserve local structural transitions that are directly leveraged by the LBP operator to encode neighborhood relationships it also preserves discriminative

micro-texture patterns associated with paper wrinkles, edge discontinuities, as well as degradation of natural skin structures and printing artifacts.

3.2.3 Resizing Image

Images obtained from different datasets frequently possess varying spatial resolutions owing to differences in camera specifications and acquisition protocols. Directly processing images of inconsistent dimensions would generate feature vectors with varying spatial characteristics, thereby reducing classifier robustness.

To overcome this limitation, every facial image is resized to 64×64. The eq 1 expresses operation as

$$I_r = f(I) \tag{1}$$

3.2.4 Normalization

Following image resizing, RGB images supplied to the convolutional neural network undergo pixel normalization prior to deep feature extraction. Normalization transforms pixel intensities into a standardised numerical range, thereby improving optimization stability, accelerating convergence during feature learning, and preventing individual color channels from dominating gradient propagation.

3.3 Separate Preprocessing Pipelines

The proposed method intentionally maintains independent preprocessing streams for handcrafted texture extraction as well as deep semantic learning as both feature extraction paradigms exploit fundamentally different image characteristics. Multi-Scale Local Binary Pattern relies exclusively on grayscale intensity transitions to characterise local facial micro-textures, while the CNN benefits from the richer chromatic and semantic information preserved within RGB images. Employing separate preprocessing pipelines therefore enables each feature extraction module to operate on the most informative image representation while preserving their complementary characteristics. This design forms the foundation of the subsequent multi-level feature fusion strategy proposed in this work.

3.4 FEATURE EXTRACTION

Feature extraction comprises of creating, transforming, and selecting features to improve results by converting the raw data or information which is preprocessed into meaningful features representation given to classifier [30], [31]. This section of focuses on the techniques that our propsed model used to extract the features for detection of the monocular RGB 2D printed photo attacks. The techniques and models this study used are shown in the table 5 shows the various pipelines we implemented before our proposed work.

3.4.1 Local Binary Pattern Feature Extraction

Natural human skin exhibits highly irregular micro-texture patterns generated by fine wrinkles, pores, pigmentation variations, and subsurface skin structures. During presentation attacks, these intrinsic facial characteristics are partially degraded owing to image printing, paper reflections, display devices, image recapturing, or material deformation. Consequently, spoof media introduce characteristic texture artifacts including printing dots, paper fibers, edge distortions, as well as smooth intensity transitions that differ significantly from genuine facial skin. The LBP is a texture descripture which is non-parametric and handmade as we decide the value for the neighborhood which is modeled as a circle with variable radius R and sample points P. When these values are set at multiple levels than it is called Multi Scale Local Binary Pattern. The eq2 for LBP is shown below

$$\mathrm{LBP}_{P,R} = \sum_{p=0}^{P-1} s(g_p - g_c) \cdot 2^p \tag{2}$$

Where:
g_c : Gray-level intensity of the center pixel.
g_p : Gray-level intensity of the neighbor pixel.
P : Total number of neighbors in the neighborhood.
2^p : Binomial weight assigned to each neighbor.

$$f(x) = \begin{cases} x^2 & \text{if } x < 0 \\ x & \text{if } x \geq 0 \end{cases} \tag{3}$$

Since Local Binary Pattern explicitly models local intensity transitions, it becomes particularly effective for detecting these subtle spoof-related texture inconsistencies while remaining computationally lightweight.

3.4.2 Multi Scale Local Binary Pattern Feature Extraction

Among various handcrafted texture descriptors, Local Binary Pattern (LBP) has demonstrated remarkable effectiveness owing to its computational simplicity, robustness to monotonic illumination variations, and ability to characterize local intensity relationships. Nevertheless, conventional LBP computed at a single neighborhood scale often captures only limited texture information, thereby reducing its discriminative capability under complex presentation attack scenarios.

To overcome this limitation, the proposed framework employs MS-LBP, where local texture descriptors are extracted at multiple neighborhood radii and subsequently aggregated through spatial histogram representation. This multi-scale analysis enables simultaneous characterization of fine facial micro-textures and broader regional texture structures, thereby improving the discrimination between genuine and spoof facial images.

Unlike many existing hybrid face anti-spoofing approaches that utilize conventional single-scale LBP or combine multiple handcrafted descriptors, the proposed framework exploits a carefully optimized Multi-Scale LBP configuration specifically designed to

preserve complementary facial texture information while maintaining low computational complexity is defined as $MS\text{-}LBP = [\text{LBP}_{P_1,R_1}, \text{LBP}_{P_2,R_2}, \ldots, \text{LBP}_{P_n,R_n}]$.

- **Level 1 (Fine):** Radius $R = 1, P = 8$. It captures immediate neighbors.
- **Level 2 (Medium):** Radius $R = 2, P = 16$. It catches intermediate patterns.
- **Level 3 (Coarse):** Radius $R = 3, P = 24$. It looks at wide macro-textures.
- **Fusion:** Histograms from all levels are concatenated into one feature vector. As shown in 4

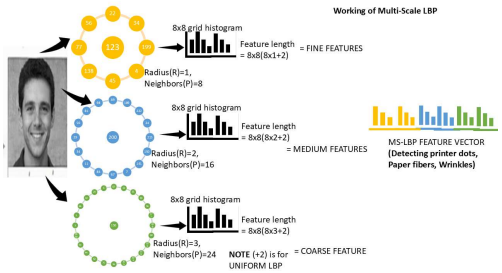


Fig. 4: Flowchart of working of MS- LBP

3.4.3 CNN Based Hierarchical Semantic Feature Learning

Following handcrafted texture extraction, semantic facial representations are learned using a custom convolutional neural network (CNN). While the Multi-Scale Local Binary Pattern module explicitly captures local facial micro-texture variations associated with presentation attack artifacts, it is unable to effectively model higher-level semantic relationships among facial components. Consequently, an additional deep learning branch is incorporated to automatically learn hierarchical facial representations capable of describing structural, contextual, and semantic characteristics that complement handcrafted texture descriptors.

Unlike conventional CNN classifiers that directly perform end-to-end prediction using a Softmax output layer, the proposed convolutional architecture is designed exclusively as a hierarchical feature extractor. The extracted deep representations are subsequently integrated with handcrafted Multi-Scale LBP descriptors through a multi-level feature fusion strategy before final classification. This design enables simultaneous exploitation of handcrafted and learned representations while reducing the dependency on large-scale end-to-end CNN optimization.

The overall CNN architecture consists of five convolutional stages, two self-attention modules, global average pooling, and a lightweight feature projection network. Instead of utilizing only the final deep embedding, semantic representations extracted at multiple abstraction levels are preserved to maximize complementary information during feature fusion.

Hierarchical Convolutional Feature Learning

Convolutional Neural Networks (CNNs) are widely used in computer vision because they can automatically learn meaningful features from raw images without relying on manually designed descriptors. Instead of extracting handcrafted features, CNNs learn feature representations directly from the input image by applying a series of convolutional operations. As the information passes through successive layers, the network gradually transforms simple visual patterns into more complex and discriminative representations.

In the proposed model, the CNN architecture consists of five convolutional blocks. Each block is responsible for extracting features at a different level of abstraction. The initial layers learn low-level features such as edges, corners, and texture patterns, while the intermediate layers capture facial components like the eyes, nose, and mouth. The deeper layers combine these features to learn high-level semantic representations that help distinguish genuine faces from spoofing attacks.

The output of each convolutional layer is computed using the following eq 4 :

$$F^{(l)} = \sigma(W^{(l)} * F^{(l-1)} + b^{(l)}) \tag{4}$$

where:

- $F^{(l)}$ is the output feature map of the l-th convolutional layer.
- $F^{(l-1)}$ is the input feature map from the previous layer.
- $W^{(l)}$ represents the learnable convolution kernel.
- $b^{(l)}$ denotes the bias associated with the convolution operation.
- $*$ denotes the convolution operator.
- $\sigma(\cdot)$ represents the nonlinear activation function, such as the Rectified Linear Unit (ReLU), which introduces nonlinearity into the network.

Convolutional blocks

The proposed architecture employs five convolutional stages to progressively transform facial appearance information from low-level texture primitives into highly discriminative semantic embeddings. The first two convolutional blocks primarily learn primitive visual structures including edges, corners, intensity gradients, and simple facial textures. The third convolutional stage begins capturing intermediate facial structures such as eyes, nose contours, mouth boundaries, and facial geometry. The fourth and fifth convolutional stages learn increasingly abstract semantic representations describing global facial appearance, contextual relationships among facial components, and presentation attack characteristics. Instead of relying exclusively on the final deep representation, intermediate semantic features generated after the third and fourth convolutional stages are preserved because they contain valuable structural information that complements handcrafted texture descriptors. This hierarchical representation learning strategy forms the basis of the proposed multi-level feature fusion mechanism.

Batch Normalization

Deep convolutional networks often experience unstable optimization due to continuously changing feature distributions during training. To alleviate this problem, Batch Normalization is incorporated after every convolutional layer. Without feature normalization, internal feature distributions continuously change during optimization, leading to unstable gradient propagation and slower convergence. Batch Normalization alleviates this internal covariate shift, enabling stable hierarchical feature learning while improving generalization.

Batch Normalization standardizes feature distributions before nonlinear activation, thereby improving optimization stability, accelerating convergence, and reducing sensitivity to parameter initialization.

LeakyReLU Activation

The proposed framework adopts LeakyReLU because subtle presentation attack characteristics frequently generate low-magnitude feature responses that may be completely suppressed by conventional ReLU activation. Preserving these weak responses enables richer semantic representation learning while improving gradient propagation throughout the network.

Max Pooling

After each convolutional stage, feature dimensionality is reduced using Max Pooling. Pooling reduces resolution while preserving dominant activations. Pooling progressively enlarges the effective receptive field while suppressing redundant spatial information. This enables the network to capture increasingly global facial characteristics without substantially increasing computational complexity.

Dropouts

To lessen overfitting, dropouts are incorporated after every convolutional block. Instead of relying excessively on individual neurons, dropout encourages distributed feature learning and improves robustness.

Self Attention

The proposed CNN incorporates two self-attention modules after the third and fourth convolutional stages.

Unlike conventional convolution, which primarily captures local neighborhood information, self-attention as shown in eq 5 models long-range dependencies among distant facial regions.

The attention map is computed as

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{5}$$

where Q, K, and V represent the query, key, and value matrices, respectively, and d_k denotes the scaling dimension of the keys. Anti spoofing attacks on biometric systems do not occur in one fixed area. Attack artifacts can appear anywhere—like

the eyes, nose, mouth, or paper edges. Self attention mechanism allows the network to dynamically decide which spatial regions of the face require the most concentration to detect spoofing attempts

Global Average Pooling

Instead of using large fully connected layers directly after convolution, the proposed framework employs Global Average Pooling (GAP). It is computed as in eq6

$$g_k = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} F_k(i, j) \tag{6}$$

where $F_k(i, j)$ denotes the activation at spatial location (i, j) in the k^{th} feature map, H and W represent the height and width of the feature map, respectively, and g_k is the pooled feature corresponding to the k^{th} channel.

Projection Network

The 512-dimensional global representation is further transformed through a lightweight projection network consisting of two fully connected layers (512→256→128). Rather than performing classification, the projection network learns a compact semantic embedding that preserves discriminative facial characteristics while reducing feature dimensionality prior to hierarchical feature fusion.

This compact representation minimizes redundancy and facilitates effective integration with handcrafted Multi-Scale LBP descriptors.

The proposed CNN therefore generates three complementary semantic representations corresponding to intermediate (128-dimensional), higher-level (256-dimensional), and projected deep (128-dimensional) embeddings. Unlike conventional CNN-based face anti-spoofing methods that utilize only the final network output, the proposed framework preserves semantic information across multiple abstraction levels. These hierarchical representations are subsequently integrated with handcrafted Multi-Scale LBP descriptors through the multi-level feature fusion strategy presented in the following subsection.

3.4.4 Multi-Level Feature Fusion

Following handcrafted texture extraction and hierarchical semantic feature learning, the proposed framework integrates the complementary feature representations through a multi-level feature fusion strategy. The primary objective of feature fusion is to simultaneously preserve local facial micro-texture information extracted by Multi-Scale MS-LBP and semantic facial representations learned at different abstraction levels by the proposed convolutional neural network.

Unlike conventional face anti-spoofing frameworks that utilize either handcrafted descriptors or deep learning representations independently, the proposed approach recognizes that both feature categories describe fundamentally different facial characteristics. Handcrafted texture descriptors explicitly capture local intensity transitions associated with presentation attack artifacts, whereas convolutional neural networks progressively learn semantic facial structures through hierarchical representation

learning. Individually, neither representation provides a complete description of facial authenticity. Therefore, combining these complementary representations enables more comprehensive characterization of genuine and spoof facial images.

Furthermore, unlike many existing hybrid approaches that concatenate handcrafted descriptors with only the final CNN embedding, the proposed framework preserves semantic information extracted from multiple intermediate convolutional stages. This hierarchical fusion strategy allows the classifier to simultaneously exploit local texture information, intermediate structural representations, and highly abstract semantic embeddings, thereby improving discriminative capability while maintaining computational efficiency.

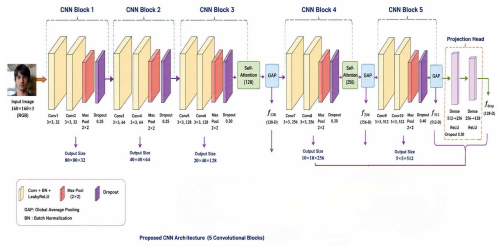


Fig. 5: Feature Extraction pipeline of CNN

Table 5: Feature extraction methods employed across different face liveness detection pipelines.

Pipeline	Feature Extraction Strategy
Decision Tree (Raw Pixels)	None (raw pixel intensities used directly)
Multi scale LBP with SVM classifier	Local Binary Pattern (LBP) texture descriptors with spatial and multi-scale histogram encoding
CNN	CNN-learned hierarchical spatial features (implicit feature learning)
Multi scale LBP fused with CNN and MLP classifier	CNN deep feature vectors extracted from penultimate layer
Multi scaleLBP fused with CNN and SVM (Proposed)	Feature-level fusion of multi-scale spatial LBP descriptors and multi-level CNN deep features (f128, f256, fdeep)

3.5 CLASSIFICATION

This part is all about how we decide if a face is real or fake by making sure we are being fair when we compare different ways of showing features, we're using several classification models in this study. Below are the listed classifiers this study implemented and the table 6 shows various classifiers and training strategies we worked on.

3.5.1 SVM Classifier

The proposed framework intentionally adopts Support Vector Machine instead of the conventional Softmax classifier commonly employed in end-to-end convolutional neural networks. This design choice is motivated by several important observations.

First, the convolutional neural network is not designed as a complete classification model but rather as a hierarchical semantic feature extractor. Consequently, the objective of the CNN is to generate discriminative facial embeddings rather than directly predicting class labels.

Second, the hybrid feature vector produced by the proposed framework contains complementary handcrafted and deep semantic information with substantially higher dimensionality than the original image representation. Support Vector Machine has consistently demonstrated excellent performance for high-dimensional feature spaces because its optimization process depends primarily on support vectors rather than the entire training dataset.

Finally, compared with end-to-end Softmax classification, SVM generally exhibits superior generalization when the available training dataset is relatively small, as is commonly observed in publicly available face presentation attack detection datasets.

Therefore, utilizing SVM as the final classifier enables the proposed framework to exploit the complementary characteristics of handcrafted texture descriptors and deep semantic representations while minimizing overfitting.

Hyperparameter-Optimized Support Vector Machine

Following the extraction of handcrafted texture descriptors and hierarchical semantic representations, the proposed framework performs facial authenticity classification using a Support Vector Machine (SVM). Rather than employing an end-to-end Softmax classifier within the convolutional neural network, the proposed architecture utilizes the CNN exclusively as a semantic feature extractor while delegating the final classification task to an optimized SVM. This design enables the classifier to operate directly on the high-dimensional hybrid feature representation generated through multi-level feature fusion, thereby improving discrimination between genuine and spoof facial images while maintaining robust generalization.

Support Vector Machine has demonstrated remarkable effectiveness for high-dimensional pattern recognition problems owing to its maximum-margin optimization principle. Unlike conventional classifiers that attempt to minimize empirical classification error alone, SVM simultaneously maximizes the separation margin between different classes. Consequently, the classifier exhibits improved robustness against overfitting, particularly when the number of training samples is relatively limited compared with the dimensionality of the feature space. Since the proposed framework produces a comprehensive hybrid representation by combining handcrafted texture descriptors and hierarchical CNN embeddings, SVM provides an effective mechanism for learning the optimal decision boundary within this discriminative feature space.

Table 6: Classification models and training strategies used across different pipelines.

Pipeline	Classifier and Training Strategy
Decision Tree (Raw Pixels)	Decision Tree classifier with constrained depth and class-weighted training
LBP , SVM	Support Vector Machine with RBF kernel; hyperparameters optimized using GridSearchCV
CNN (End-to-End)	CNN classification head trained end-to-end using binary cross-entropy loss
LBP, CNN with MLP	Fully connected neural network classifier trained on CNN feature vectors
LBP, CNN with SVM (Proposed)	Support Vector Machine with RBF kernel trained on fused LBP–CNN features; ; hyperparameters optimized using GridSearchCV

3.6 Working

As described in the fig 2the first step is to have the dataset of RGB images which is given two parallel independent streams of process where the first branch employs MS-LBP to capture handcrafted texture descriptors at multiple spatial scales. By computing LBP histograms over different radii and spatial grids, this branch effectively characterizes spoof-specific micro-texture cues such as printer dots, paper fibers, edge discontinuities, and the degradation of natural skin microstructures that are commonly introduced during 2D photo print attacks.

Simultaneously, the second branch processes the same input image using a custom convolutional neural network to learn hierarchical semantic facial representations. As the network progresses through successive convolutional blocks, it gradually transforms low-level visual primitives into increasingly discriminative facial representations. Intermediate convolutional layers preserve local structural information, including facial edges, eye contours, nose boundaries, and regional facial patterns, whereas deeper layers encode high-level semantic information representing the overall facial structure and identity-related characteristics.

These intermediate and deep CNN features are subsequently integrated with the handcrafted MS-LBP descriptors through a multi-level feature fusion strategy, allowing complementary texture and semantic information to interact before classification. This fusion produces a richer and more discriminative feature representation by combining local texture variations with hierarchical facial semantics.

The fused feature vectors obtained from different abstraction levels are concatenated to construct a unified high-dimensional representation that simultaneously preserves handcrafted texture information, intermediate structural cues, and deep semantic knowledge. Finally, the resulting feature vector is supplied to a

hyperparameter-optimized linear Support Vector Machine (SVM), which constructs a maximum-margin decision boundary for distinguishing bona fide presentations from spoof attacks. By decoupling feature representation learning from classification, the proposed framework improves generalization of 2-D photo print images while maintaining computational efficiency suitable for lightweight face anti-spoofing applications.

3.7 EVALUATION OF MODELS

To get a good glance at how well our detection systems, we are using following ways to measure the performance.

3.7.1 Accuracy

We use accuracy as shown in Eq 7 as our main measure to tell how many genuine and spoof samples we got right. This is basically just how many predictions we got right out of all the guesses we made during testing. Accuracy gives a quick idea of how good a model is and compare different setups. Different iteration accuracy of the proposed model given in the tables **??** and **??**

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}$$

3.7.2 Confusion Matrix

Confusion matrix is a heatmap that help us see how well a classification model did by breaking down the results into true positives, true negatives, false positives, and false negatives as shown in fig 8 and fig 9

3.7.3 Precision, Recall, F1- score

We calculate precision, recall, and F1-score as shown in Eqs 8, 9, and 10 respectively for each class. This helps us understand how well things are going, especially when some classes have way more data than others.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{8}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{9}$$

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{10}$$

3.7.4 Balanced Accuracy

Balanced Accuracy is a robust evaluation metric that measures the average classification performance across both classes can be seen in eq11. Unlike conventional accuracy, it reduces the effect of class imbalance by assigning equal importance to the positive

and negative classes. It is computed as the average of sensitivity (recall) shown in eq12, and specificity in eq13.

$$\text{Balanced Accuracy} = \frac{\text{Sensitivity} + \text{Specificity}}{2} \tag{11}$$

where

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{12}$$

and

$$\text{Specificity} = \frac{TN}{TN + FP}. \tag{13}$$

3.7.5 Matthews Correlation Coefficient

The Matthews Correlation Coefficient (MCC) is a comprehensive performance metric that simultaneously considers true positives, true negatives, false positives, and false negatives. It provides a balanced evaluation of binary classifiers, even when the dataset is imbalanced. The MCC value ranges from -1 to 1, where 1 indicates perfect classification, 0 represents random prediction, and -1 indicates complete disagreement between predictions and ground truth. The eq 14 can be described as

$$\text{MCC} = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{14}$$

3.7.6 Cohen's Kappa

Cohen's Kappa measures the level of agreement between the predicted and actual class labels after removing the agreement expected by chance. It provides a more reliable assessment of classifier performance than accuracy alone, particularly in classification problems with unequal class distributions. The formula is described in eq15, eq 16, eq17.

$$\kappa = \frac{P_o - P_e}{1 - P_e} \tag{15}$$

where

$$P_o = \frac{TP + TN}{TP + TN + FP + FN} \tag{16}$$

is the observed agreement, and

$$P_e = \frac{(TP + FP)(TP + FN) + (FN + TN)(FP + TN)}{(TP + TN + FP + FN)^2} \tag{17}$$

represents the expected agreement by chance.

3.7.7 ROC and AUC graph

We use ROC curves to check how good classifiers are at telling things apart, no matter where we set the cutoff or threshold point. It gives us a way to measure how well we can tell the real face images from the fake, and it doesn't depend on where we set the threshold or cutoff point as shown in fig 10

4 Results and Discussion

Experimental evaluation was conducted in which we trained our system using the NUAA Photograph Imposter Database. To see how well it worked in different situations, we also tested it with the Yale Recaptured Dataset and the Photo Print Attack Dataset. Our system correctly identified faces 97.26% of the time. This is a good result, especially since our system is quite simple and doesn't need huge pre-trained models or special cameras. It shows that we can detect fake faces effectively without complex setups. We also executed some pipelines systematically that led us to our proposed model. The proposed framework has been accepted through a progressively experimental process shown in tables 10 and 11. The fig 8, fig 9 and fig 10 shows the results of various implemented works and fig fig9 shows various accuracies of implemented works.

The results reflects the resistance of the system from 2-D photo printed attacks, warped facial attacks, wrapped printed photo attacks cutout prints and illumination conditions as well.

We executed a comparision with an existing model [3]. The following table 8 shows the difference. The fig 11 shows a comparison of accuracies. Furthermore, as shown in fig 7, in comparision to the existing model our proposed model are more suitable in terms of framework, complexity of parameters.

We learned that combining different types of features really helps. The CNN part focused on how the face looked, its structure, and colors on the printed paper, the CNN part focused more on facial apperance, illumination behavior shadow information, shape consistency, and edge characterstics. The LBP part picked up on tiny texture details, like print flaws or distortions from flat re-photographing wrinkles, fine lines and subtle reflectance variations. Many of these fine grained texture traits are either suppressed or substituted by printer generated patterns during image printing and capturing on paper as the texture of the flat 2-D paper is diffferent from skin. When we put these two types of information together, the system was much better at telling real from fake than if it just used one type. This approach is great for situations where you only have a regular RGB camera and no depth, heat, or infrared information sensors. Lastly SVM serves as a decision maker. It learns optimal boundary separation between genuine live face and a 2-D printed flat face on paper, warped photos, wrapped photos.

In addition to, the strong cross dataset performance reflected the improved generalization capabilty across different aquisition, camera conditions, illumination variation, different environments, lighting, and expression as well. We trained our model using the NUAA dataset, which was captured with a webcam. But the testing datasets involved different cameras and different conditions. For example, the Yale dataset and

Photo Print Attack dataset has face images taken with varying lighting and expressions. The spoof dataset also adds more variety because it includes 2-D flat print attacks. These are usually big problems for face anti-spoofing systems that only use regular color cameras. Even with all these big changes, our model still stayed very accurate. This suggests it learned what makes a fake image fake, rather than just recognizing specific camera patterns. The results of cross domain dataset generalization can be seen in fig 12, obtaining the accuracy of 93%.

Another benefit of this framework is that it is very practical to set up. Many current anti-spoofing systems require special sensors like infrared or depth cameras, but this approach only needs standard RGB images. This means you do not need any special hardware and can just use the cameras that are already in place. Because of this, the system works well for cheaper or simpler devices, such as smartphones, office attendance systems, and remote identity checks.

Beyond conventional performance measures, the proposed framework was further evaluated, as shown in fig 6 using Balanced Accuracy, Matthews Correlation Coefficient (MCC), and Cohen's Kappa to provide a more comprehensive assessment of classification reliability. The model achieved a Balanced Accuracy of 96.54%, confirming its ability to classify both bona fide as well as presentation attack samples with comparable effectiveness. Also, the obtained MCC of 0.9431 indicates a strong correlation between predicted and actual class labels while considering all the results of the confusion matrix, highlighting the robustness of the classifier. Similarly, the Cohen's Kappa value of 0.9421 demonstrates almost perfect agreement beyond chance, providing additional evidence that the proposed framework generalizes well along with making the model highly reliable classification decisions. Collectively, these metrics complement the high accuracy of 97.26% with reinforcing the effectiveness of the proposed multi-scale feature fusion framework for face presentation attack detection.

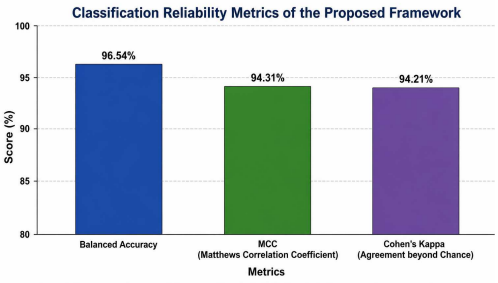


Fig. 6: Preprocessing steps of an image

Overall, our experiments show that the LBP-CNN-SVM system successfully reduces the problems caused by different lighting, environmental changes, varying camera sensors, and biases in the datasets. Its strong performance across different datasets means it can adapt better to new situations. This also highlights how useful hybrid feature learning approaches could be for practical single-camera face anti-spoofing applications.

The security benefits of the proposed framework extends beyond the classification accuracy and withstands the other evaluation metrics demonstrate that the proposed framework is not only accurate but also balanced, reliable, and capable of learning meaningful discriminative representations for attack detection and cross data generalization as shown in table 7contributing identity fraud prevention, to attack surface reduction that filters a flat 2-D printed paper photo or warped photo from a live user, to the practical deplyoment of the frame work unlike many existing solutions that require infrared, thermal or depth sensors where as the proposed framework operates on standard RGB camera images strengthening the biometric authentication security.

Table 7: Performance evaluation of the proposed face anti-spoofing framework using standard classification and Reliability metrics.

Metric	**Value**
Accuracy	97.34%
Precision	99.21%
Recall (Sensitivity)	93.52%
Specificity	99.57%
F1-score	96.28%
Balanced Accuracy	96.54%
Matthews Correlation Coefficient (MCC)	0.9431
Cohen's Kappa	0.9421

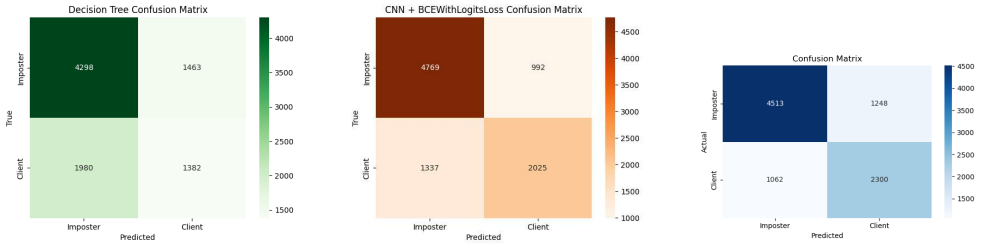


(a) (b)

Fig. 7: (a)- Confusion Matrix of existing Model(ECNN + LBP + LTP + SVM), (b)- Roc- Auc graph of existing Model(ECNN + LBP + LTP + SVM)

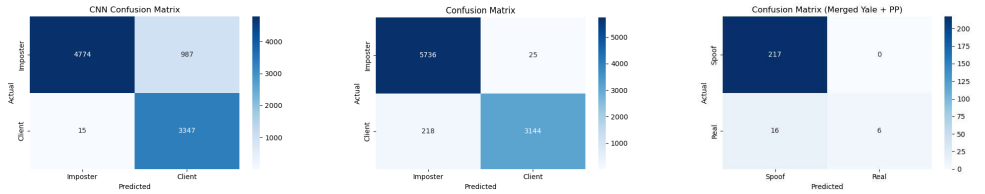
5 Conclusion and Future Scope

This study presented a lightweight and monocular hybrid framework for face presentation attack detection (PAD) using monocular RGB facial imagery. The proposed



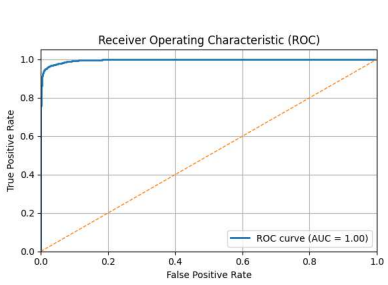
(a) (b) (c)

Fig. 8: (a)- Confusion Matrix of Decision Tree, (b)- Confusion Matrix of CNN, (c)- Confusion Matrix of LBP Feature Extraction and SVM classifier

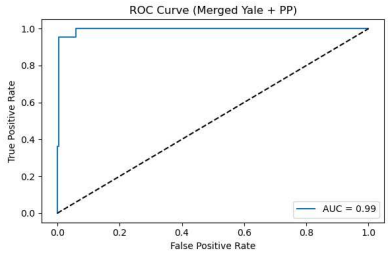


(a) (b) (c)

Fig. 9: (a)- Confusion Matrix of LBP Feature Extraction and CNN feature extraction and CNN fully connected, (b)- Confusion Matrix of proposed pipeline LBP Feature Extraction CNN feature extraction and SVM classifier, (c)- Confusion Matrix of proposed pipeline LBP Feature Extraction CNN feature extraction and SVM classifier trained cross dataset (yale and photoprint attack)



(a) ROC Graph for proposed model on NUAA Dataset



(b) ROC Graph for proposed model on merge dataset (yale and photo print attack)

Fig. 10: ROC graph of the proposed model on two different datasets

Table 8: Comparative analysis between Existing ECNN + LBP + LTP + SVM and the proposed LBP + CNN + SVM framework

Aspect	**ECNN + LBP + LTP + SVM**	**Proposed LBP + CNN + SVM**
CNN Architecture	Encoder CNN (ECNN) with 6 convolutional layers	CNN with 5 convolutional layers
Fully Connected Layers	Multiple FC embedding layers ($256 \rightarrow 128 \rightarrow 256$) increase parameter dependency	No additional fully connected projection layers
Architectural Complexity	Comparatively more complex	Simpler and lightweight architecture
Computational Cost	Higher due to deeper CNN and FC layers	Lower computational complexity
Handcrafted Features	Uses both LBP and LTP descriptors	Uses only LBP descriptor
Feature Redundancy	Possible redundancy due to LBP and LTP belonging to similar texture families	Reduced redundancy with balanced feature fusion
Fusion Strategy	Accumulation of multiple texture descriptors with ECNN features	Complementary fusion of texture and semantic features
Semantic Representation	Limited by shallow ECNN representation	Stronger semantic learning through CNN features
Texture Representation	Strong texture emphasis	Balanced texture and structural representation
Model Dependency	More dependent on handcrafted texture features	More balanced deep-learning and handcrafted integration
Generalization Evaluation	Primarily intra-dataset evaluation	Includes cross-dataset and multi-source evaluation
Cross-Dataset Testing	Not extensively explored	Evaluated using Yale + Photo Print Attack datasets
Robustness to Unseen Data	Generalization capability not fully validated	Improved robustness under unseen acquisition conditions
Real-World Applicability	Moderate	Better suited for practical deployment
Feature Representation	Texture-dominant representation	Complementary semantic-texture representation
Experimental Accuracy	$\sim$95% (reproduced implementation)	97.26%

approach has been designed to identify a practical gap between highly complex/sensor-dependent PAD systems and simple RGB systems, and designed a monocular semantic learning framework to address the balance among sensing simplicity, computational efficiency, complementary feature representation, and cross-dataset generalization. The results displayed the effectiveness of the proposed approach for the evaluated planar presentation attacks. The framework achieves an overall accuracy of 97.26%, with 99.21% Precision, 93.52% Recall, 99.57% Specificity, 96.28% F1-score, and 96.54% Balanced Accuracy. In addition, the obtained MCC of 0.9431 and Cohen's Kappa of 0.9421 indicate strong agreement between the predicted and actual classes. The approximately 93% accuracy obtained in cross-dataset evaluation further demonstrates promising generalization when the proposed framework is evaluated on an independent data distribution.

Future research can substantially advance the hybrid Multi-Scale LBP with CNN feature fusion and SVM classification framework suggested in this paper by investigating a number of interesting avenues of Vision Transformer based, Explainable Anti Spoofinf systems, deepfake-aware liveness detection, and multi-modal biometric

Table 9: Performance comparison of different anti-spoofing pipelines.

Pipeline	**Accuracy (%)**	**Precision**	**Recall**	**F1-score**	**Evaluation Type**
Decision Tree	62.26	0.49	0.41	0.45	Intra-dataset (NUAA)
CNN	74.47	0.67	0.60	0.63	Intra-dataset (NUAA)
Multi-scale LBP + SVM	74.68	0.648	0.684	0.666	Intra-dataset (NUAA)
Multi-scale LBP + CNN + FC	89.00	0.77	0.99	0.87	Intra-dataset (NUAA)
LBP + CNN → SVM (Proposed)	**97.30**	**0.99**	**0.93**	**0.96**	**Intra-dataset (NUAA), Cross-dataset (Yale + Photo Print Attack)**

security framework.

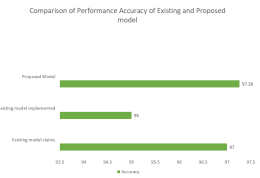


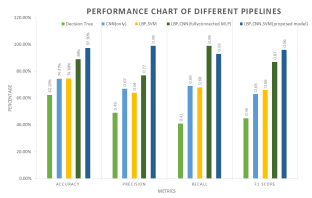
Fig. 11: Performance Comparison between existing and proposed hybrid model

Table 10: Proposed architectures's hyperpaprametertuning and their respective accuracy (Exp 1-20)

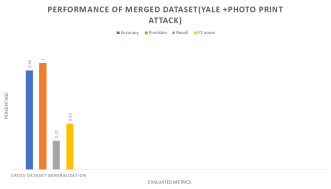
Exp	Configuration	Acc. (%)
Exp 1	Baseline CNN: 3 convolution blocks	84.12
Exp 2	Added 4th convolution layer with BatchNorm	87.64
Exp 3	Added 5th convolution layer	87.19
Exp 4	Activation changed from ReLU to LeakyReLU	88.72
Exp 5	LBP method: uniform → var; radius: 1 → 2	89.00
Exp 6	Increased LBP sampling points: 8 → 16	90.00
Exp 7	Enabled spatial-LBP (high-dimensional)	80.00
Exp 8	Grid = 4×4; n_points = 8	91.00
Exp 9	Polynomial SVM in GridSearchCV	79.00
Exp 10	CNN trained using Adam optimizer	93.00
Exp 11	CV = 6; LBP grid = 8×8	–
Exp 12	LBP n_points = 8	–
Exp 13	Grid = 8×8; n_points = 8	91.00
Exp 14	Radius = 2; n_points = 16	92.00
Exp 15	Radius = 3	**95.00**
Exp 16	Radius = 4	92.00
Exp 17	Grid = 2×2	93.00
Exp 18	Grid = 2×2; radius = 3	92.00
Exp 19	Grid = 2×2; radius = 2	92.77
Exp 20	n_points = 32; radius = 2	94.45

Table 11: Proposed architectures's hyperpaprametertuning and their respective accuracy(Exp 21-40)

Exp	Configuration	Acc. (%)
Exp 21	Grid = 4×4; radius = 3	92.95
Exp 22	Grid = 8×8; n_points = 8	93.47
Exp 23	Grid = 8×8; n_points = 16	93.98
Exp 24	Grid = 4×4; n_points = 16	91.00
Exp 25	Grid = 4×4; n_points = 8	89.49
Exp 26	Radius = 5	–
Exp 27	Spatial LBP with grid = 8×8	**97.17**
Exp 28	Spatial LBP with grid = 16×16	94.63
Exp 29	Added stacked kernel	95.00
Exp 30	Stacked kernel + grid 8×8	**97.26**
Exp 31	LBP method set to ROR	76.50
Exp 32	Dropout schedule refinement	97.27
Exp 33	Increased input resolution to 128×128	97.10
Exp 34	Conv kernel size: 3	97.00
Exp 35	Conv kernel size: 5	**97.34**
Exp 36	Added attention after Block 3	**97.34**
Exp 37	Added attention after Block 4	97.33
Exp 38	Dual attention (Block 3 + Block 4)	**97.34**
Exp 39	MS-LBP radii = [1, 3, 5]	97.07
Exp 40	MS-LBP radii = [1, 2, 3]	**97.26**



(a) Performances of different pipelines on NUAA Dataset



(b) Performances of proposed pipeline on merged (yale + photo print attack) Dataset

Fig. 12: Performances on different datasets

Acknowledgements. I would like to express my sincere gratitude to the Centre for Advanced Studies (CAS), Dr. A.P.J. Abdul Kalam Technical University (AKTU), Lucknow, for providing the necessary infrastructure, resources, and academic environment that made this research possible.

Declarations

Some journals require declarations to be submitted in a standardised format. Please check the Instructions for Authors of the journal to which you are submitting to see if you need to complete this section. If yes, your manuscript must contain the following sections under the heading 'Declarations':

- Funding - Authors declare that this work is not funded by any organisation or agency.
- Conflict of interest/Competing interests- The authors declare that they have no competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.
- Author contribution- Naimisha Awasthi: Conceptualization, Methodology, Software, Data Curation, Formal Analysis, Investigation, Validation, Visualization, Writing original draft preparation.
  Usha Sharma: Supervision, Methodology, Validation, Resources, Writing, Review & Editing.
  Yudhishthir Raut: Supervision, Resources, Validation, Project Administration, Writing, Review & Editing.
  Deepak Kumar Jain: Supervision, Resources, Validation, Project Administration, Writing, Review & Editing.
  All authors contributed to the study conception and design. Material preparation, data collection, implementation, experimentation, and analysis were primarily performed by Naimisha Awasthi. All authors reviewed the manuscript critically for important intellectual content, approved the final version of the manuscript, and agree to be accountable for all aspects of the work.

- Data Availability- The datasets generated and/or analysed during the current study are available in the KAGGLE repository. The link for NUA Dataset - https://www.kaggle.com/datasets/aleksandrpikul222/nuaaaa The link for Yale Face Dataset - https://www.kaggle.com/datasets/olgabelitskaya/yale-face-database/data The link for PHOTO PRINT ATTACK Datasset - https://www.kaggle.com/datasets/axondata/photo-print-attacks-dataset-1k-individuals

References

[1] Patel, S.S., Dash, A.A., Mishra, N., Yang, T., Rathore, R.S., Mo, D., Wang, L.: Enhancing biometric authentication through score-level fusion of gait and palm vein modalities. In: 2024 International Conference on Intelligent Algorithms for Computational Intelligence Systems (IACIS), pp. 1–7 (2024). https://doi.org/10.1109/IACIS61494.2024.10721667

[2] Yu, Z., Qin, Y., Li, X., Zhao, C., Lei, Z., Zhao, G.: Deep learning for face anti-spoofing: A survey. IEEE transactions on pattern analysis and machine intelligence **45**(5), 5609–5631 (2022)

[3] Hassani, A., Diedrich, J., Malik, H.: Monocular facial presentation–attack–detection: Classifying near-infrared reflectance patterns. Applied Sciences **13**(3) (2023)

[4] Radad, M., Enab, A.E., Elagooz, S.S., El-Fishawy, N.A., El-Rashidy, M.A.: Face anti-spoofing detection based on novel encoder convolutional neural network and texture's grayscale structural information. International Journal of Computational Intelligence Systems **18**(1), 175 (2025)

[5] Liu, S., Zhang, K.-Y., Yao, T., Bi, M., Ding, S., Li, J., Huang, F., Ma, L.: Adaptive normalized representation learning for generalizable face anti-spoofing. In: Proceedings of the 29th ACM International Conference on Multimedia, pp. 1469–1477 (2021)

[6] Mohan, K., Chandrasekhar, P., Ramanaiah, K.: Object-specific face authentication system for liveness detection using combined feature descriptors with fuzzy-based svm classifier. International journal of computer aided engineering and technology **12**(3), 287–300 (2020)

[7] Pandey, A., Tiwari, A.K.: Advancing face spoofing detection with lbp, pca, and svm: A robust ai security approach. ICTACT Journal on Image & Video Processing **15**(4) (2025)

[8] El-Rashidy, M.A., Enab, A.E., Elagooz, S.S., El-Fishawy, N.A., Radad, M.: A novel texture descriptor using machine learning for face anti-spoofing detection. International Journal of Machine Learning and Cybernetics, 1–22 (2025)

[9] Chakole, V.V., Dixit, S.R.: Mlbfa: A machine learning-based face anti-spoofing detection framework under replay attack. International journal of electrical and computer engineering systems **16**(4), 321–332 (2025)

[10] Shinde, S.R., Bongale, A.M., Dharrao, D., Thepade, S.D.: An enhanced light weight face liveness detection method using deep convolutional neural network. MethodsX **14**, 103229 (2025)

[11] Kuznetsov, O., Frontoni, E., Romeo, L., Rosati, R., Maranesi, A., Muscatello, A.: Deep learning models for robust facial liveness detection. arXiv preprint arXiv:2508.09094 (2025)

[12] Hiremath, S.S., Hatture, S., Hiremath, M.S., Shrimantrao, A., Hiremath, J.S., Sivanandan, S.K.: Real-time face detection and anti-spoofing with mtcnn and densenet201 for secure biometric authentication. In: 2025 International Conference on Computing Technologies (ICOCT), pp. 1–7 (2025). IEEE

[13] Rusia, M.K., Singh, D.K.: A color-texture-based deep neural network technique to detect face spoofing attacks. Cybernetics and Information Technologies **22**(3), 127–145 (2022)

[14] Singh, R.P., Dash, R., Mohapatra, R.K.: Lbp and cnn feature fusion for face anti-spoofing. Pattern Analysis and Applications **26**(2), 773–782 (2023)

[15] Jha, K., Srivastava, S., Jain, A.: A novel texture based approach for facial liveness detection and authentication using deep learning classifier. Int. J. Comput. Exp. Sci. Eng. **10**(3), 323–331 (2024)

[16] Gupta, N., Kaur, A.: An optimized feature extraction framework for hybrid learning-based face anti-spoofing. Lex Localis **23**(S4), 2236–2251 (2025)

[17] Liu, A., Tan, Z., Liang, Y., Wan, J.: Attack-agnostic deep face anti-spoofing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6336–6345 (2023)

[18] Lee, K.W., Lim, J.Y., Lim, K.M., Lee, C.P.: Face presentation attack detection via ensemble learning algorithm. In: 2023 IEEE 11th Conference on Systems, Process & Control (ICSPC), pp. 101–106 (2023). https://doi.org/10.1109/ICSPC59664.2023.10420311

[19] Vannurswamy, K., Shekar, B.H., Pilar, B., Karunakar Kotegar, A., Jiang, F.: Deep ensemble learning approach for face anti-spoofing detection based on pre-trained models. In: 2024 IEEE International Conference on Computer Vision and Machine Intelligence (CVMI), pp. 1–6 (2024). https://doi.org/10.1109/CVMI61877.2024.10782399

[20] Muhammad, U., Laaksonen, J., Romaissa Beddiar, D., Oussalah, M.: Domain

generalization via ensemble stacking for face presentation attack detection. International Journal of Computer Vision **132**(12), 5759–5782 (2024)

[21] Qi, H., Han, R., Duan, K., Shi, Y., Qi, X., Gao, C., Ren, L.: A high-performance adaptive fusion network for face anti-spoofing detection. Scientific Reports **15**(1), 37607 (2025)

[22] Singh, M., Anand, N., Sharma, G.: Exponentiating the precision of facial spoof detection: An intricately woven tapestry of hybrid machine learning paradigms and state-of-the-art textural feature extraction methodologies (2023)

[23] Prasad, M., Jain, S., Bhanodia, P., Priya, A.: Influence of standalone and ensemble classifiers in face spoofing detection using lbp and cnn models. European Journal of Electrical Engineering and Computer Science **8**(2), 17–30 (2024)

[24] Hamad, S.S., Ibrahim, N.J., Fadhil, M.N.: Texture features with resnet50v2 for face spoofing detection. In: AIP Conference Proceedings, vol. 3264, p. 030033 (2025). AIP Publishing LLC

[25] Chavan, D.U., Salunkhe, S., *et al.*: Ai-powered secure banking with face & liveness verification. International Journal on Advanced Computer Engineering and Communication Technology **14**(1), 563–570 (2025)

[26] Peixoto, B., Michelassi, C., Rocha, A.: Face liveness detection under bad illumination conditions. In: 2011 18th IEEE International Conference on Image Processing, pp. 3557–3560 (2011). IEEE

[27] Belhumeur, P.N., Hespanha, J.P., Kriegman, D.J.: Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. IEEE Transactions on pattern analysis and machine intelligence **19**(7), 711–720 (1997)

[28] Data, A.: Photo Print Attacks Dataset (1k Individuals). https://www.kaggle.com/datasets/axondata/photo-print-attacks-dataset-1k-individuals. Accessed: 2025-02-15 (2021)

[29] Ashraf, M.W.A., Singh, A.R., Pandian, A., Rathore, R.S., Bajaj, M., Zaitsev, I.: A hybrid approach using support vector machine rule-based system: detecting cyber threats in internet of things. Scientific Reports **14**(1), 27058 (2024)

[30] Awasthi, N., Gautam, P.R., Sharma, A.K.: Rfecv-dt: Recursive feature selection with cross validation using decision tree based android malware detection. In: 2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT), pp. 1–6 (2024). https://doi.org/10.1109/ICCCNT61001.2024.10725127

[31] Pandey, V.K., Sahu, D., Prakash, S., Rathore, R.S., Dixit, P., Hunko, I.: A lightweight framework to secure iot devices with limited resources in cloud

environments. Scientific Reports **15**(1), 26009 (2025)

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- DATAAVAILABILITY.docx