ISSN: 2093-5374 / E-ISSN: 2093-5382

Region-wise Explainability for Trustworthy Face Spoofing Detection using Gradient-weighted Class Activation Mapping and Facial Landmarks

S. Karthika[1*] and Dr.G. Padmavathi[2]

[1*]Assistant Professor, Department of Information Technology, Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore, Tamil Nadu, India.
karthika_it@avinuty.ac.in, https://orcid.org/0009-0002-6822-4281

[2]Professor, Department of Computer Science, Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore, Tamil Nadu, India. padmavathi_cs@avinuty.ac.in, https://orcid.org/0000-0002-5377-4451

Received: February 12, 2026; Revised: March 20, 2026; Accepted: May 07, 2026; Published: June 30, 2026

Abstract

Face spoofing poses a serious threat to biometric systems. Although deep learning models achieve high accuracy in detecting spoofed faces, their decision-making process often lacks transparency, and their decisions are often difficult to understand. This research presents a region-wise explainability approach to enhance transparency and interpretability in face spoofing detection. The method focuses on spoof-prone facial regions, such as the eyes, nose, and mouth, identified using landmark-based segmentation. A ResNet18 model integrated with a Convolutional Block Attention Module (CBAM) is trained separately on full-face and cropped facial regions. The model achieves high accuracy across all inputs, with the full-face model attaining 99.22% test accuracy, while the nose, eyes, and mouth regions also show strong performance with accuracies above 97%. In addition, all regions achieve a zero false acceptance rate, which is important for secure biometric systems. Subsequently, Gradient-weighted Class Activation Mapping, known as Grad-CAM, is applied across all face regions to visualize the regions that influence the model's decisions. The generated heatmaps show that the full face yields the highest prediction confidence of 0.956 among the other regions. Further, a confidence drop analysis is performed by masking the most important regions identified by Grad-CAM. This helps to understand how much each facial region contributes to the final decision and verifies whether the model is focusing on meaningful spoof-related features.

Keywords: Trustworthy AI, Region-wise Explainability, Face Recognition, Grad-CAM, Explainable AI.

1 Introduction

Face Recognition (FR) systems are now widely used for security and authentication purposes. However, these systems are still vulnerable to spoofing attacks, such as the use of printed images, replayed videos and 3D masks. Convolutional Neural Networks (CNNs) are proficient in detecting various spoofs, as they have the ability to learn complex feature representations. However, their internal decision-making process is not always transparent and understandable. This lack of interpretability creates concerns

Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications (JoWUA), volume: 17, number: 2 (June-2026), pp. 244-263. DOI: *10.58346/JOWUA.2026.I2.014*

*Corresponding author: Assistant Professor, Department of Information Technology, Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore, Tamil Nadu, India.

security-critical applications. Hence, there is a huge need in making these systems to be more transparent and trustworthy beyond improving detection accuracy.

To address this, a region-wise explainability framework is proposed in this research that highlights the contribution of facial regions in spoof detection using Grad-CAM. This grad-cam visualizes the heatmap relays of each face image so that it is easy to demonstrate which area the model is focus on. A core contribution of this work is to measure the confidence score for each prediction to assess how certain the model is about its decisions. In addition, prediction confidence drops and similarity changes are also measure upon region masking to understand their individual importance. This work bridges the gap between performance and interpretability in face anti-spoofing.

The proposed work is carried out in two phases that aim to improve the model performance and provide interpretable results. Phase I aims at developing a classification model using a ResNet18 architecture integrated with an attention-aware module called Convolutional Block Attention Module (CBAM), proposed by Woo et al. (2018). In this phase, the model is trained with full-face images and specific regions, such as eyes, nose, and mouth, cropped according to the detected facial landmarks.

In Phase II, a region-wise explainability approach using Gradient-weighted Class Activation Mapping (Grad-CAM) was proposed by Selvaraju et al. (2017). This is done to visualize how each region of a face contributes to the model's decision, rather than considering the face as a single sample. This phase also consists of concept called confidence drop analysis. Grad-CAM visualization is used to determine the most impactful parts of the face and masked out one at a time. Then, the corresponding change in prediction confidence is estimated to understand the contribution of the individual face region.

The following points summarize the core contributions of this proposed work:

- A face region-wise explainable framework has been proposed wherein facial regions are segregated based on the landmarks and processed independently to find out which regions are more vulnerable to attacks.
- A ResNet18 model with the attention-focused module, Convolutional Block Attention Module (CBAM) is used. This combination enables the model to pay attention to the most relevant and informative features in both spatial and channel-level.
- The analysis of confidence drop with similarity is done with the help of explainability approach Grad-CAM. Heatmaps are created, and the impact of important regions identified by the heatmaps is analysed through masking.

The rest of the paper is structured as follows. In Section 2, an overview of the related work in the field of facial spoofing detection and explainable artificial intelligence approaches is given. In Section 3, the proposed approach is introduced, which includes data preprocessing, model architecture, and the region-wise explanation paradigm. The results obtained from the experiments are presented in Section 4. Finally, Conclusions and Future Work are presented in Section 5.

2 Related Works

Several deep learning-based approaches have been developed to detect face spoofing in recent years to address some of the shortcomings associated with conventional approaches when dealing with spoofing. Some of the notable literature reviews on the subject of face anti-spoofing by Yu et al. (2022) and Xing et al. (2025) highlight the rapid advancements in deep learning-based techniques.

Chen et al. (2019) have suggested a two-stream CNN based attention mechanism to detect face spoofing in both RGB and illumination invariant Multi-Scale Retinex (MSR) domains. MobileNet and ResNet-18 are used for feature extraction in the two domains, and the attention-based fusion strategy has helped improve robustness to lighting conditions. Experiments conducted on CASIA-FASD, REPLAY-ATTACK, and OULU-NPU databases have shown that such an approach provides superior performance compared to models trained using single dataset.

Similarly, Wang et al. (2020) have used spatiotemporal gradient information, temporal depth learning, and contrastive depth loss functions to enhance robustness. In addition, a dataset called 'DMAD' is also compiled that consists of depth maps for real and spoof faces to evaluate the system's effectiveness.

The foundational work of a deep learning approach to detecting face spoofing has been provided by Najeebullah et al. (2025). In the work, a basic CNN framework is used for face anti-spoofing. Therefore, this work can be considered helpful to understand various possibilities that arise with deep learning in anti-spoofing applications. Najeebullah et al. (2025) performed a comparative analysis of popular architectural options such as MobileNetV2, ResNet50, and ViT in a face anti-spoofing setting. Based on the experiments conducted on a large-scale dataset, the authors concluded that MobileNetV2 provides the best performance efficiency trade-off.

Shinde et al. (2025a) proposed a deep CNN architecture for detecting face liveness through enhanced feature extraction. In other similar work, Shinde et al. (2025b), a lightweight deep CNN model is designed to perform face anti-spoofing in real-time. This work is considered as suitable for practical deployment as it achieves a balance between computational efficiency and detection performance.

Several studies have explored improving robustness through feature combinations. Jaswanth & Ramprasad, (2023) combined texture-based descriptors, such as Local Binary Patterns (LBP), with deep learning models to improve model robustness in challenging imaging conditions. Kong et al. (2022) enhanced ResNet50 with a channel attention mechanism to focus on important facial regions. Additionally, a feature fusion technique is employed by Zuama et al. (2025), where the combination of identity-aware and texture-aware features enhances detection performance. The study is based on the improvement of face spoofing detection using feature fusion techniques. The study suggests a face spoofing detection model based on the fusion of FaceNet and DenseNet201 architectures, which are effective in extracting identity-based and texture-based features, respectively. The suggested model is effective in achieving higher classification accuracy and is validated on various benchmark datasets, such as the NUAA and LCC-FASD datasets. The study suggests that data augmentation may not always be effective in improving the accuracy of the model and may even increase the error rates in some cases. Although the suggested model is effective in improving the accuracy and reliability of the face spoofing detection model, the study is based only on feature fusion techniques and does not provide any insights regarding the analysis of the contributions of the facial regions.

While the above methods focus on improving accuracy and generalization, the demand for explainability in face spoofing detection has led to the integration of visual explanation techniques. Some recent studies have explored explainability. Rajpal et al. (2023) investigated the use of Local Interpretable Model-Agnostic Explanations (LIME) to improve the interpretability of deep face recognition models such as LeNet-5, AlexNet, Inception-V3, and VGG16. Their experiments on the Yale, AT&T, and LFW datasets showed that LIME effectively highlights visually important facial regions.

However, there exists a problem related to the reliability of explainability approaches. For instance, Ghorbani et al. (2019) found that explainability approaches are sensitive and unstable to small perturbations on the input, even if the prediction is stable. Thus, the application of explainability can cause wrong interpretations in face anti-spoofing systems.

Recently, Singh et al. (2026) proposed a novel explainable AI framework. It consists of the combination of the learning process and the application of new explainability techniques that use Shapley values to determine important regions. Novel elements called 'XAIPooling' and 'XAIDropout' assist the algorithm in identifying the most relevant features during the learning process. As far as the architecture of the proposed framework is concerned, a multi-stream architecture is employed using both RGB and LBP images. This model demonstrates good generalization abilities on several datasets. However, the study does not analyze the impact of regions' importance on face spoofing detection.

Thus, the application of explainability in AI-based face anti-spoofing algorithms is a big breakthrough for developing trusted systems. The transparency of the model increases significantly, making it suitable for implementation. Nevertheless, the implementation of this approach implies certain problems, including performance consistency and transparency. However, overcoming those problems is essential for advancement. Moreover, the main goal of this research work is region-specific analysis to identify the significance of specific regions such as eyes, nose, and mouth in detecting spoofed faces.

3 Proposed Approach

The proposed work integrates CBAM and Grad-CAM techniques to enhance the region-wise explanation capability by enabling the model to focus on and explain the spoof-vulnerable facial areas, such as eyes, nose, and mouth.

Problem Definition

Face spoofing detection involves distinguishing genuine facial samples from spoofing attacks like printed photos or replayed videos. Given an input facial image $I \in \mathbb{R}^{(H \times W \times 3)}$, the objective is to learn a classification function: $f(I) \rightarrow y$, where $y \in \{0,1\}$, 0 denotes a spoof face, and 1 denotes a genuine face.

While deep convolutional neural networks attain high classification accuracy, their internal decision-making process lacks transparency. Therefore, this work formulates face spoofing detection not only as a classification task but also as an explainable learning problem.

Specifically, the proposed system is designed to:

i. Perform region-wise facial analysis by segmenting the input face into landmark-guided regions, including eyes, nose, mouth, and full face.
ii. Develop an attention-enhanced ResNet18 architecture integrated with Convolutional Block Attention Module to learn spoof-related features.
iii. Generate visual explanations using Gradient-weighted Class Activation Mapping to emphasize the decision-relevant the regions relevant to model decisions.
iv. Quantify region-wise reliability through prediction confidence scores.
v. Evaluate the importance of each region using confidence drop analysis via heatmap-guided masking.

System Model and Assumptions

The proposed framework works as a region-wise explainable face spoofing detection framework. For any input frame of a face, denoted by I, captured from the video of the Replay-Attack dataset, the framework first performs a series of structured pre-processing steps before region-wise classification and explainability.

Let the extracted frame be represented as in equation 1:

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

Where v denotes a video sample and $f(\cdot)$represents the frame extraction function.

First, facial landmarks are extracted using MediaPipe Face Mesh, which provides 468 key facial points. Based on these landmarks, four facial inputs are generated as shown in equation 2:

$$R = \{R_f, R_e, R_n, R_m\} \tag{2}$$

Where R_f: full face region, R_e: eyes region, R_n: nose region, R_m: mouth region.

Each region is resized to 224×224 and independently processed by an attention-enhanced convolutional neural network. For each region $R_i \in \mathcal{R}$, a dedicated ResNet18_CBAM model learns a mapping as in equation 3:

$$y_i = f_{\theta_i}(R_i) \tag{3}$$

Where f_{θ_i}denotes the ResNet18 integrated with CBAM, θ_i represents the trainable parameters, and $y_i \in \{0,1\}$ represents the predicted classes (0: spoof, 1: genuine).

The model produces softmax probabilities as represented in equation 4:

$$S_i = \max(\text{Softmax}(f_{\theta_i}(R_i))) \tag{4}$$

where $S_i$represents the prediction confidence for region i.

To provide interpretability, Grad-CAM generates a heatmap H_i based on the gradients of the target class with respect to the last convolutional layer. The importance of each region is further validated using heatmap-guided masking. Let M(.) denote a masking operation that suppresses high-activation pixels, and equation 5 shows:

$$R_i^{masked} = M(R_i, H_i) \tag{5}$$

The masked region $R_i^{masked}$is then passed again through the trained model to obtain a new confidence score S_i^{masked}.

The confidence drop is computed by using equation 6:

$$\Delta S_i = S_i^{original} - S_i^{masked} \tag{6}$$

A larger ΔS_i value indicates higher regional importance in spoof detection.

System Assumptions

The proposed system operates under the following assumptions:

1. Each input sample contains a single clearly visible face.
2. One representative frame per video is sufficient to capture spoof-related artifacts.
3. Facial landmarks extracted using MediaPipe are accurate for region segmentation.

4. Faces are approximately frontal and can be aligned using eye coordinates.
5. Spoof attacks are limited to photo and replay-based presentation attacks as defined in the Replay-Attack dataset.
6. Region-wise models are trained independently to evaluate the discriminative strength of each facial area.

These assumptions ensure controlled experimental evaluation and enable reproducible region-wise analysis.

Overall System Architecture

The system architecture of the proposed framework for detecting spoof faces is shown in figure 1 below. It is clear that the entire architecture of this framework involves multiple steps such as preprocessing, feature extraction using region classification, and explainability.

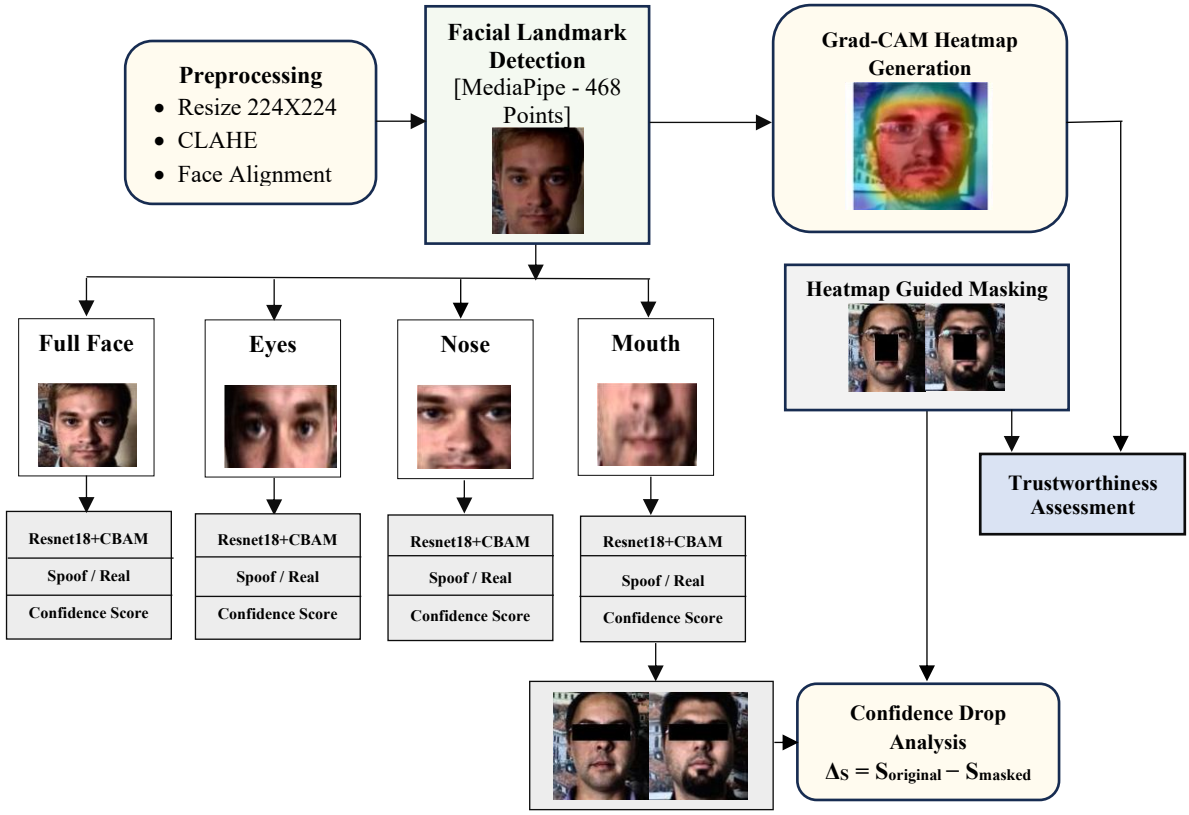

Figure 1: Overall system architecture

Dataset Preparation and Preprocessing

Replay-attack dataset containing real and spoof face videos. One image per video is extracted from each video to create an extensive and diverse dataset of real and spoof face images. In order to standardize input and improve facial features' interpretation, pre-processing techniques are implemented. All the input images are normalized by resizing them to 224x224 dimensions as required by the model's input dimensions. For stable training of the neural network model, normalization of pixel values is performed by limiting them to the [0,1] scale.

MediaPipe Face Mesh library is used to detection facial landmarks with 468 facial key points. These landmarks are important for several further steps, including face alignment and extraction of certain facial regions. Alignment of facial images is completed using eye coordinates. Contrast Limited Adaptive Histogram Equalization (CLAHE) technique is employed to increase the contrast of an image. The aim of this is to recognize artifacts that are hard to see under regular lighting conditions.

Some specific facial regions such as the eyes, nose, mouth and also the full face are extracted by region-wise cropping which is completed using facial landmark coordinates. A slight padding is applied on each region in order to ensure proper extraction of all the facial features. Finally, structured folder preparation is conducted in order to train, validate, and test face models in region-wise manner.

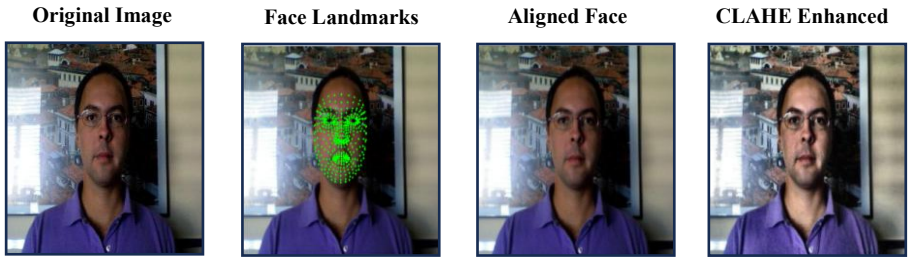

Figure 2(a): Landmark detection, face alignment and contrast enhancement

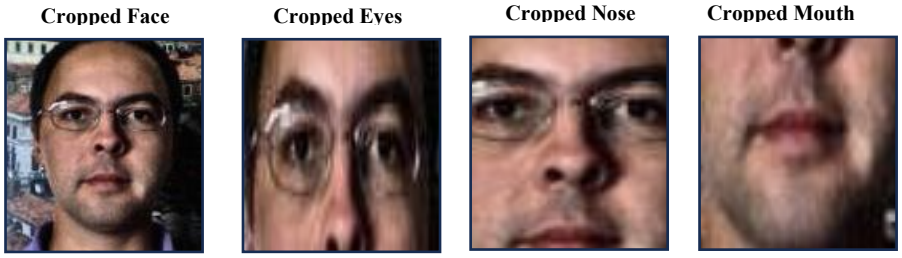

Figure 2(b): Region-wise cropped face images

The figure 2a illustrates the preprocessing pipeline, including facial landmark detection, face alignment based on eye positions, and contrast enhancement using CLAHE. Following this, region-wise cropping is performed using key landmarks, resulting in focused views of the eyes, nose, mouth, and full face, as shown in figure 2b. The Replay-Attack dataset used for experimental evaluation is publicly available and can be accessed from the Idiap Research Institute website: https://www.idiap.ch/en/scientific-research/data/replayattack.

Model Development

The proposed model used a ResNet18 as the base model and enhanced using the Convolutional Block Attention Module (CBAM).

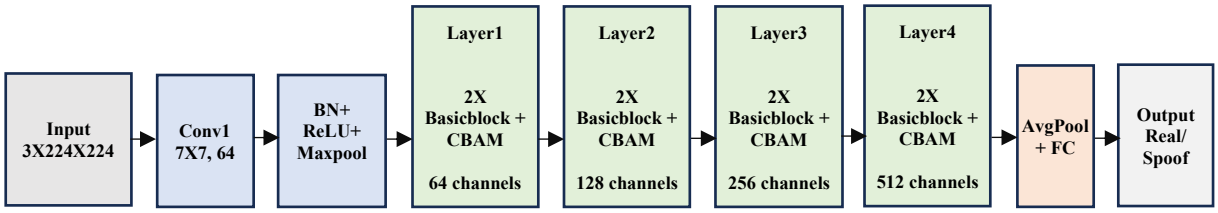

Figure 3: Architecture of ResNet18 with CBAM model

CBAM modules are inserted after certain residual blocks in the Resnet18-architecture. These modules enable the network to focus on most informative and spoof-prone regions in the feature map by introducing spatial and channel attention. Figure 3 illustrates the overall structure, where each CBAM block sequentially applies both channel and spatial attention. Channel attention learns which types of features such as texture, color and lighting are more important and spatial attention learns where in the image such as eyes, nose or mouth, the important features are located.

The ResNet18_CBAM model processes an RGB input image through a sequence of layers intended to extract meaningful features for face spoofing detection. The process starts with an initial convolutional block, which uses a 7×7 convolution to capture basic low-level patterns such as edges and fine-textures. This is followed by batch normalization, ReLU, and 3×3 max pooling layer to downsize the spatial dimension and preserve only the necessary characteristics. Then the image representation moves on to four residual layers from Layer1 up to Layer4, each containing two residual blocks. Two 3×3 convolutions are applied within each residual block, and each is followed by batch normalization and ReLU. Also, there are skip connections added after the convolutions, making each output equal to the sum of the output and the input.

The CBAM technique is applied after the second convolution in each residual block. Channel Attention within CBAM consists of average and max pooling on spatial axes, followed by a fully-connected network, called MLP, and a sigmoid activation, resulting in per-channel weights that amplify or mute those depending on their importance. Within Spatial Attention, the channel-wide average and maximum pools are taken. To enable the model focus on the most informative facial parts for the efficient detection task, their outputs are concatenated and a 7×7 convolution with sigmoid activation is used to produce spatial weights.

In this ResNet18_CBAM model, each next residual layer becomes deeper, but the input spatial size decreases. Layer 1 maintains the same spatial dimension and applies 64 filters. Layer 2 downsizes the spatial size and doubles the number of filters. It is repeated for Layers 3 and 4, both employing downsampling with 256 and 512 filters, respectively. The hierarchical architecture of the network allows capturing increasingly abstract higher-level features due to increased depth. After applying all residual blocks, a Global Average Pooling (GAP) layer is applied to squeeze the output of each feature map into one value summarizing the spatial information. GAP is fed to Fully Connected layer to make the final decision on what class does the input belongs either real or spoofed.

Training and Evaluation Setup

The resulting dataset after preprocessing the data according to Section 3.3.2 is split into train, validation, and test sets in a ratio of 70:15:15. The ResNet18_CBAM model is trained on each facial region such as full face, eyes, mouth, and nose. During the training process, the Adam optimization algorithm is used to efficiently update the parameters of the ResNet18_CBAM model including those influenced by channel and spatial attention mechanisms. The Adam optimizer helps minimize the binary cross-entropy loss by adjusting weights and biases during training to ensure the model effectively learns discriminative patterns between real and spoofed facial regions. The best-performing model checkpoint based on validation accuracy is saved for each region. The model is trained using a batch size of 16 and a learning rate of 0.0001.

The ResNet18 backbone is initialised with ImageNet pre-trained weights to facilitate transfer learning and speed up the convergence process. The newly added CBAM blocks in the model are initialised using the Kaiming initialisation method (He et al., 2015) to ensure stable gradient propagation

in the training process. The final fully connected layer is initialised using Xavier initialisation (Glorot & Bengio, 2010) and adapted for binary classification. All bias terms are initialised to zero, and the entire network is fine-tuned during training.

Python programming language is used to implement all experiments. The proposed deep learning models are developed with PyTorch and Torchvision. Image processing and transformations are performed with OpenCV and Torchvision libraries. Facial landmark detection is performed with MediaPipe Face Mesh. Grad-CAM visualization is performed with PyTorch-based gradient operations. Heatmap processing is performed with OpenCV. Experiments are performed in a Jupyter Notebook environment with an Anaconda distribution on a computer with Microsoft Windows 11. The computer hardware configuration is an Intel Core Ultra 7 155U processor with 12 cores and 16 GB RAM.

Explainability using Grad-CAM

Confidence Score Analysis helps to figure out that how confident the model is about its predictions. However, it does not reveal why the decision is made by the model; it means that the trustworthiness is still not obtained. Hence, this section aims to add explainability to the model's decision by highlighting the facial regions that influenced its predictions.

Gradient-weighted Class Activation Mapping (Grad-CAM) is used to produce heatmaps for each facial region. The heatmaps reveal important areas in the input image affecting the model's output and indicate which pixels influenced the model most in its prediction. As Grad-CAM uses the spatial information in convolutional layers to produce clear and localized explanations, it enhances interpretability by confirming that the model focuses on meaningful facial cues, improving trust and reliability.

Let $A^k \in R^{u \times v}$ denotes the activation map of the k^{th} channel in the last convolutional layer, where u and v are spatial dimensions. The Grad-CAM method computes the importance weights α^k_c for a target class c by averaging the gradients of the score y^c with respect to the activation map A^k as in equation 7.

$$\alpha_k^c = \frac{1}{u \times v} \sum_{i=1}^{u} \sum_{j=1}^{v} \frac{\partial_{y^c}}{\partial A_{ij}^k} \quad (7)$$

These weights represent the importance of feature map k for the class c. The Grad-CAM heatmap, $L^c_{\text{Grad-CAM}}$ is then computed using equation 8 as the weighted sum of the activation maps, followed by a ReLU operation to focus on positive influences:

$$L^c_{Grad\_CAM} = \text{ReLU}\left(\sum_k \alpha_k^c A^k\right) \quad (8)$$

The resulting heatmap $L^c_{Grad\text{-}CAM}$ highlights the spatial regions most relevant for predicting class c, which is then upsampled and overlaid on the input image for visualization. Figure 4 illustrates the heatmaps generated using Grad-CAM for each region such as full face, eyes, nose and mouth.

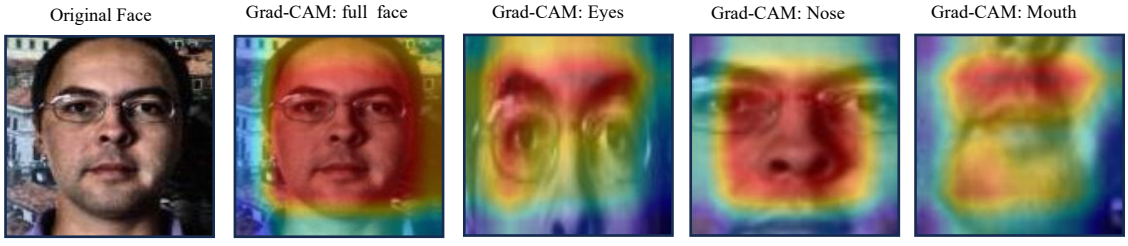

Figure 4: Heatmaps using Grad-CAM

An algorithmic representation of the proposed framework has been added as Algorithm 1 to clearly describe the workflow. The method begins with the extraction of frames from the input videos. Next, some basic preprocessing steps include resizing, normalization, face alignment, and contrast enhancement using CLAHE, are applied to the extracted frames to improve image quality. Subsequently, the facial landmarks are identified to detect the key regions include full-face, eyes, nose and mouth. The ResNet18 model with CBAM is used to process each region, which produces both the predicted class label and the corresponding confidence score.

To enhance the interpretability of model's decisions, heatmaps are generated using Grad-CAM technique and the most influential regions are highlighted. These regions are then masked, and the modified input samples are again processed using the model to observe the chances in the confidence. The difference between the original and updated confidence scores, namely confidence drop, shows that how each region is important for the prediction. This approach helps to understand the individual contribution of different face regions.

Algorithm 1: Region-wise Explainable Face Spoofing Detection

Input: Video dataset V

Output: Prediction $\hat{y}_i$, confidence Si, heatmap H_i, confidence drop ΔS_i

1: Extract frame I from each video in V

2: Preprocess I (resize, normalize, align, apply CLAHE)

3: Detect facial landmarks and extract regions: Full_face, Eyes, Nose, Mouth

4: for each region Ri:

5: $\hat{y}_i, S_i = \text{ResNet18\_CBAM}(R_i)$

6: Generate Grad-CAM heatmap H_i

7: Mask high-activation regions → R_i_masked

8: S_i_masked = model (R_i_masked)

9: $\Delta S_i = S_i - S_i$_masked

10: end for

Return $\hat{y}i, S_i, H_i, \Delta S_i$

Heatmap-guided Masking and Confidence Drop Analysis

To investigate the effect that the regions which have driven the decision-making process of the model, the heatmaps produced by the use of the Grad-CAM algorithm are analyzed.

Figure 5: Sample masked images for regions (a) eyes (b) mouth (c) nose

The next stage in the assessment process involves investigating the effect that the masking of those regions has on the overall performance of the model. This stage is important since it will help to establish whether the regions selected do have a substantial effect on the decision that the model makes. Figure 5 illustrates sample images with the most important regions masked.

Heatmap-guided masking involves using the regions identified by the heatmaps generated by the use of Grad-CAM and then masking those regions in the original image by assigning the pixel values to zero. This is done in order to determine how the model's decision is affected.

Performance Evaluation Metrics

TP and TN refer to the true classifications for spoof and non-spoof samples respectively, whereas FP and FN are the incorrect classifications. The performance measures used in this paper are shown below, from Equations (9) - (16).

Accuracy is an index that refers to the effectiveness of the model in discriminating real and spoof face samples and is defined in equation 9.

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \tag{9}$$

Error Rate refers to the ratio of misclassified samples, is calculated using equation 10.

$$\text{Error Rate} = \frac{\text{FP+FN}}{\text{TP+TN+FP+FN}} \tag{10}$$

Precision refers to the ratio of correctly detected spoof samples out of all the samples identified as spoofs and is calculated using equation 11.

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \tag{11}$$

Recall measures the ability of the model to correctly identify spoof samples and is defined in equation 12.

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

F1-score is the harmonic mean of precision and recall, providing a balanced evaluation of classification performance, and is defined in equation 13.

$$\text{F1}_{\text{score}} = 2 * \frac{\text{Precision}*\text{Recall}}{\text{Precision+Recall}} \tag{13}$$

False Acceptance Rate (FAR) represents the proportion of real samples incorrectly classified as spoof, and is calculated using equation 14.

$$\text{FAR} = \frac{\text{FP}}{\text{FP+TN}} \tag{14}$$

False Rejection Rate (FRR) refers to the ratio of the number of spoof samples that are misclassified as genuine, and is calculated using equation 15.

$$\text{FRR} = \frac{\text{FN}}{\text{FN+TP}} \tag{15}$$

An overall error measure is computed by Half Total Error Rate (HTER), the mean of FAR and FRR as shown in equation 16.

$$\text{HTER} = \frac{\text{FAR+FRR}}{2} \tag{16}$$

Overall, the efficiency of the proposed face spoof detection method is completely assessed by all the metrics.

4 Results and Discussion

This section presents the results obtained through implementing the described approach and further discusses their analysis along with implications for improving face spoofing detection. Also, in addition to the evaluation of accuracy, other factors like confidence scores and techniques for increasing explainability of models are considered.

Training and Test Performance

Once the training process is finished, the performance of the model is assessed using the testing set, with the results shown below in table 1.

The very high accuracy of 99.22% proves that the model was able to identify between genuine and spoofed images with high efficiency due to the use of complete facial characteristics. But the mouth region shows a slightly higher validation loss of 0.0441, evidencing a minor overfitting even though it has the lowest training loss of 0.0053. The eyes and nose regions provide test accuracy over 97% proves that these regions contain appropriate information for detecting face spoofs.

Table 1: Summary of model training and evaluation performance

Region	Training Loss	Validation Loss	Train Accuracy	Validation Accuracy	Test Loss	Test Accuracy
Full face	0.0158	0.0152	0.9940	0.9944	0.0036	0.9922
Eyes	0.0211	0.0385	0.9952	0.9889	0.0562	0.9778
Nose	0.0200	0.0241	0.9928	0.9889	0.0607	0.9889
Mouth	0.0053	0.0441	0.9976	0.9833	0.0298	0.9833

This region-wise evaluation also works as an ablation study as it highlights that how each facial region contributes when compared to the full-face. The findings clearly state that individual regions provide very informative cues as well combining all these facial features together enhancing the overall performance. The model's ability of generalizing is shown through the close match between validation and test accuracies across the regions, which is illustrated in figure 6. It is evident that CBAM-based region-wise analysis effectively captures important spoofing features.

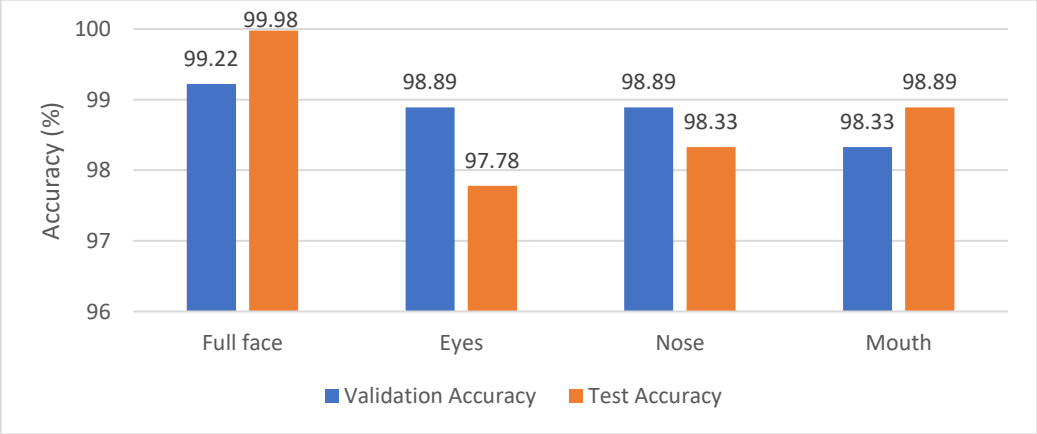

Figure 6: Validation vs test accuracy across facial region

The nose region shows the least HTER at 0.0323, indicating a higher level of reliability in comparison with other facial areas. It is therefore evident that particular areas of the face may offer consistent spoofing clues rather than the whole face.

Table 2: Comprehensive evaluation metrics across facial regions

Region	Test Accuracy	Error rate	Precision	Recall	F1-Score	FAR	FRR	HTER
Full face	0.9922	0.0078	0.99	1.00	0.99	0.0000	0.0613	0.0306
Eyes	0.9778	0.0222	0.97	0.99	0.98	0.0010	0.1290	0.0645
Nose	0.9889	0.0111	0.99	0.99	0.99	0.0010	0.0645	0.0323
Mouth	0.9833	0.0167	0.98	1.00	0.99	0.0000	0.0968	0.0484

As presented in table 2, the evaluation incorporates key performance metrics, including error rate, precision, recall, F1-score, False Acceptance Rate (FAR), False Rejection Rate (FRR), and HTER. Notably, the FAR remains negligible across all facial regions, attaining a value of 0.0000 in certain regions and only 0.0010 in others, thereby indicating an almost non-existent misclassification of spoof samples as genuine.

Discussion on Robustness, Limitations and Generalizability

The strength of the model is demonstrated through its ability to maintain consistency in performance during training, validation, and testing. Additionally, the model demonstrates stability even under partial occlusion conditions, as evidenced by the experiments carried out through the use of regional masks. Further improvement in feature learning is achieved through the inclusion of CBAM.

However, certain limitations are introduced as the model is dependence on texture-based features. Particularly, some challenging conditions such as poor illumination and low image quality may make the model to outperform. Moreover, its resilience against unknown or more sophisticated spoofing attacks such as deepfakes remains uncertain, while full-face representations continue to provide superior discriminative power compared to localized regions. The generalization capability of the model is constrained by its evaluation on a single dataset and the absence of repeated experimental validation. Future work will therefore focus on cross-dataset evaluation and the integration of additional modalities and temporal dynamics to further improve robustness and adaptability.

Confidence Score Analysis

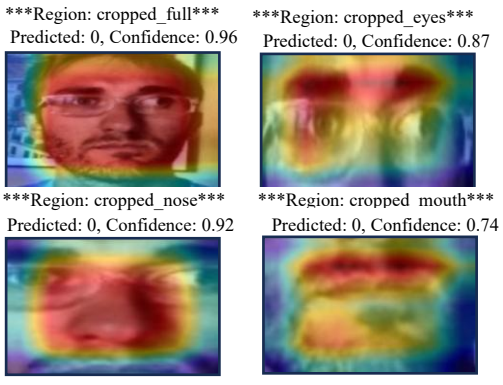

Figure 7: Confidence scores for real and spoof samples across different facial regions

Deep learning models prove to be effective in terms of their efficiency and excellent performance when dealing with face spoof images recognition tasks; however, it is crucial to verify the reliability of the

algorithms used. For this reason, it is necessary to calculate and analyze confidence scores, which can demonstrate how certain the model is of its classification such as genuine or spoof faces. The confidence score refers to the calculated probabilities, produced in the course of the model's testing procedure.

The figure 7 depicts the confidence scores for real and spoof samples across various facial regions, facilitating interpretation of region-wise model responses.

Table 3: Confidence scores

Region	True Label	Predicted	Confidence
Full face	0	0	0.956
Eyes	0	0	0.924
Nose	0	0	0.737
Mouth	0	0	0.873

The table 3 presents the mean confidence scores for real and spoof samples across different facial regions, indicating the discriminative capability of each region.

Grad-CAM Explainability Results

The Grad-CAM approach is employed to indicate which parts of the faces play an essential role in influencing the decision-making process within the ResNet18_CBAM network. Heatmaps obtained through Grad-CAM provide more information and increase interpretability, since it becomes evident which parts of the image contribute significantly to the model's prediction. Examples of Grad-CAM visualization for the faces that have been correctly predicted as real and fake can be found in figure 8. It is noticeable that when dealing with full-face inputs, the activation is rather strong in all essential parts, including the eyes, nose, and mouth.

In region-wise inputs, the activation can be noticed in the eyes and mouths in spoof images due to the fact that spoof attacks are characterized by artifacts in terms of boundaries and texture. The examples of heatmaps obtained from incorrectly classified images are provided in figure 9. They show that the activation is either weak or spread over different irrelevant regions such as the background of the image or certain facial parts. This fact explains why there was a prediction error the algorithm missed essential spoofing clues. Therefore, strong and local attention on the face regions related to spoofing attacks correlates with correct classification, while dispersion or misalignment of attention leads to prediction errors.

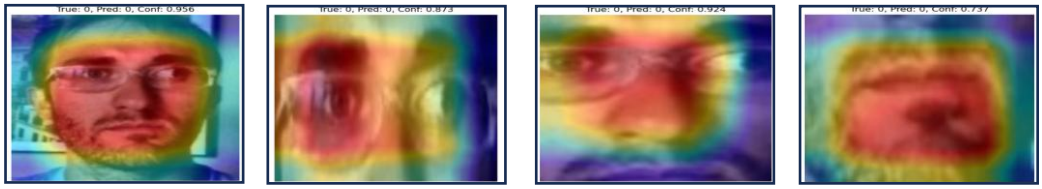

Figure 8: Grad-CAM heatmaps for correctly classified images

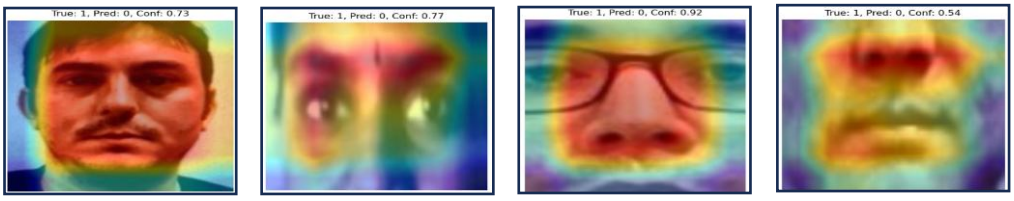

Figure 9: Grad-CAM heatmaps for misclassified images

Heatmap-Guided Masking and Confidence Drop Analysis

The masked images are reintroduced into the model to evaluate the impact of removing salient regions on prediction confidence and overall accuracy. The full-face region is intentionally excluded, as complete masking would invalidate spoof detection and conflict with the objective of region-wise analysis. A significant degradation in performance following masking indicates the importance of the corresponding regions and confirms the model's reliance on them. The confidence drop is computed using equation 17:

$$\text{Confidence Drop} = \text{Original Score} - \text{Masked Score} \quad (17)$$

A larger confidence drop signifies a greater contribution of the respective region to the model's decision-making process. Following the computation of confidence drops (e.g., the mouth exhibiting a drop of 0.05), the Structural Similarity Index (SSIM) is employed to quantify the similarity between the original and masked Grad-CAM heatmaps. The similarity scores reported in table 4 provide insights into the extent of feature representation changes induced by masking.

This region-wise analysis, illustrated in figure 10, offers a deeper understanding of redundancy and complementarity among facial regions. High similarity scores suggest compensatory behavior from other regions, whereas lower scores indicate unique and non-redundant contributions. The corresponding drop in confidence further reinforces the significance of each region.

Table 4: Region-wise confidence drop and similarity score analysis

Region	Confidence Drop	Similarity Score	Interpretation
Eyes	0.03	0.76	Moderately important as contributes useful features for detection
Nose	0.02	0.71	Least impactful as plays a supporting role in model prediction
Mouth	0.05	0.80	Most critical region and high drop show strong influence on decision-making

The findings reveal that the mouth region exerts the most substantial influence on model predictions, as evidenced by the highest confidence drop and distinct heatmap patterns. The eyes also contribute meaningfully, while the nose demonstrates comparatively limited impact. The similarity analysis further supports that the mouth encodes more distinctive spoof-related features.

Key findings consist of:

- The full-face model gives the best results of 99.22% accuracy, indicating that global feature representations work well
- Region-wise models are also highly effective, demonstrating the contributions of local features
- Grad-CAM and confidence drop analysis consistently identified the mouth as the most critical region

Overall, the integration of CBAM and Grad-CAM enhances the model's interpretability, which makes the model to suitable for trustworthy decision-making in security-critical applications. These findings also support the advancement of competent, region-focused face spoof detection models.

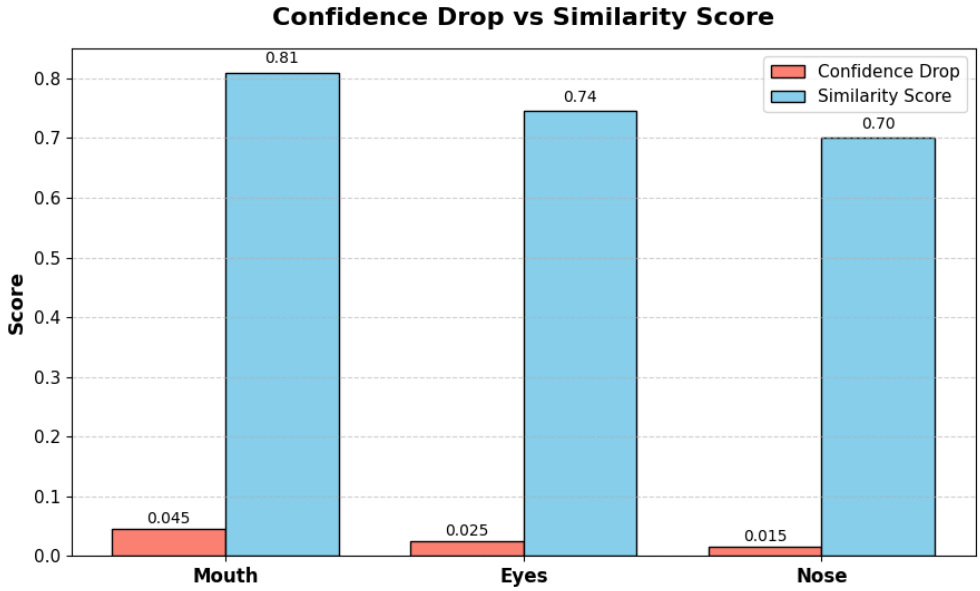

Figure 10: Region-wise confidence drop and similarity score plot for explainability evaluation

Ablation Study

Ablation study was performed to examine how much different facial regions contribute to the overall performance of the proposed system. During the ablation study, the model was evaluated using only inputs corresponding to particular facial regions such as full face, eyes, nose, and mouth. It is noteworthy that in the experiment, all architecture parameters and training settings remained the same.

The experiment showed that the test accuracy of the model with the full-face configuration is equal to 99.22%. Thus, the use of the global facial context is extremely important to distinguish between the genuine and spoofed samples. Meanwhile, the nose and mouth configurations yield relatively high test accuracy rates that equal 98.89% and 98.33%, respectively. The findings can be explained by the fact that these regions contain enough discriminative clues to detect spoofing artifacts, including texture variations and presentation artifacts.

On the other hand, the test accuracy rate of the eye region is somewhat lower and equals 97.78%. Thus, one may conclude that this facial region does not have a significant discriminative ability. In conclusion, the ablation study demonstrates that although local regions can capture subtle spoofing artifacts, the addition of the global facial context results in higher performance.

Comparison with Existing Work

The performances of the proposed framework using ResNet18 in combination with CBAM and Grad-CAM have been compared to state-of-the-art face spoofing detection techniques in table 5. It can be seen that the proposed system provides detection accuracy of 99.22%, and thus, outperforms some of the recently developed approaches.

Najeebullah et al., (2025) used ResNet18, MobileNetV2, and Vision Transformer and those pretrained models performed outstandingly when tested on the Replay-Attack dataset. Likewise, Abdullakutty et al. (2022) tested various architectures including VGG-16, ResNet-50, Inception V3, and

DenseNet-121. According to their findings, DenseNet-121 had a test accuracy rate of 97.32% with HTER rate of 4.37%.

However, the proposed approach attained higher accuracy of 99.22%. Therefore, the performance of this approach was better than those used by these researchers. The high accuracy achieved in the proposed model is mainly due to the use of CBAM, which enhances attention mechanism. Moreover, the application of Grad-CAM improves the performance of the model in feature extraction.

Table 5: Performance and explainability comparison of face anti-spoofing models

Model	Accuracy (%)	Explainability Technique	Highlights
ResNet18 (Najeebullah et al., 2025)	86.46	None	Standard deep CNN
MobileNetV2 (Najeebullah et al., 2025)	93.02	None	Lightweight and efficient architecture
Vision Transformer (Najeebullah et al., 2025)	92.09	None	Captures global feature representations
VGG-16 (Abdullakutty et al., 2022)	86.84	None	Transfer learning baseline model
ResNet-50 (Abdullakutty et al., 2022)	93.72	None	Deep residual learning-based model
Inception V3 (Abdullakutty et al., 2022)	79.50	None	Multi-scale feature extraction architecture
DenseNet-121 (Abdullakutty et al., 2022)	97.32	None	Best performance among pretrained CNNs
ResNet18 (Thiruchelvam et al., 2023)	98.20	Basic Grad-CAM	Provides visual explanation with heatmaps
SE-ResNet50 (Kong et al., 2022)	99.98	Implicit (channel attention)	Focuses on important facial regions using attention
MobileNetV3 + CBAM + CDC (Sun et al., 2024)	97.68	CBAM only	Lightweight model with attention mechanism
Proposed: ResNet18 + CBAM + Grad-CAM	99.22	Grad-CAM + CBAM	Combines high accuracy with region-wise explainability

The proposed model not only maintains competitive performance but also incorporates explainability through Grad-CAM alongside channel-spatial attention via CBAM. This facilitates both efficient spoof detection and visualization of the decision process, making the method more appropriate for use in security-sensitive environments. In summary, the comparative analysis proves that the proposed method provides a good trade-off between efficiency and interpretability, overcoming shortcomings found in other methods.

Novelty and Real-World Applicability

The principal novelty of this study lies in the unified integration of region-wise analysis, attention mechanisms, and model interpretability within a face anti-spoofing framework. Unlike conventional approaches that predominantly rely on holistic facial representations, the proposed method systematically examines distinct facial regions, namely the eyes, nose, and mouth, to capture localized discriminative cues. This fine-grain analysis not only facilitates the identification of the most informative regions for spoof detection but also provides deeper insight into the model's decision-making process.

A further significant contribution is the integration of the Convolutional Block Attention Module into the ResNet18 architecture. This CBAM module enables the model to selectively highlight salient spoof-related patterns by enhancing feature representation through adaptive channel-wise and spatial attention, thereby improving its discriminative ability.

In addition, the framework explicitly addresses model interpretability through the integration of Grad-CAM-based visual explanations and confidence score analysis. These components collectively enhance transparency by revealing the regions influencing predictions while simultaneously quantifying prediction certainty. In contrast to many existing works that primarily emphasize accuracy, the proposed approach achieves a balanced trade-off between performance, interpretability, and robustness.

From an application perspective, the proposed model demonstrates promising potential for deployment in real-world security systems, such as mobile authentication, banking security, and access control. This is justified due to the high accuracy level demonstrated by the method and resistance to partial occlusion, proven using the region masking approach. Moreover, the selection of ResNet18 as the backbone network indicates that it may be used in constrained computing environments, although more research is needed.

5 Conclusion

This research introduced a novel region-wise explainability framework for trustworthy face spoofing detection using facial landmarks, Gradient Weighted Class Activation Mapping (Grad-CAM), and the attention-enhanced ResNet18 model with CBAM. By focusing on the important regions for face spoofing, such as the eyes, nose, and mouth, this research shows the unequal contribution of different regions to the model's decision-making process.

In this research, the proposed model attained high accuracy in the test phase, which is 99.22%, for full-face images. This shows the importance of global facial features in face spoofing detection. Moreover, the region-wise analysis shows the high accuracy of the model for the nose region, which is 98.89%, and the mouth region, which is 98.33%. For the eyes region, the model attained 97.78%, which shows the importance of the entire face in the decision-making process.

Additionally, the confidence drop analysis using the heatmap-based masking method shows the importance of different regions. Among the different regions, the mouth region shows the maximum confidence drop, which is 0.05. These results show the importance of different regions in the model's decision-making process. It is evident from the above results that the model is using spatially meaningful features, which play crucial role in its decision-making process.

In conclusion, the proposed model using the region-wise analysis, attention, and explainability techniques is important for trustworthy face spoofing detection. The proposed model is compared with the existing models, which show the proposed model's high performance compared to the existing models. Moreover, the proposed model is important for its high interpretability compared to the existing models. Future work can be done on this research by incorporating different modalities such as depth images to enhance the model's performance in the presence of sophisticated attacks like 3D masks. Moreover, the proposed model can be extended to work with videos to improve the performance and interpretability.

References

[1] Abdullakutty, F., Elyan, E., Johnston, P., & Ali-Gombe, A. (2022). Deep transfer learning on the aggregated dataset for face presentation attack detection. *Cognitive computation*, *14*(6), 2223-2233. https://doi.org/10.1007/s12559-022-10037-z

[2] Chen, H., Hu, G., Lei, Z., Chen, Y., Robertson, N. M., & Li, S. Z. (2019). Attention-based two-stream convolutional networks for face spoofing detection. IEEE Transactions on Information Forensics and Security, 15, 578-593. https://doi.org/10.1109/TIFS.2019.2922241

[3] Ghorbani, A., Abid, A., & Zou, J. (2019). Interpretation of neural networks is fragile. In *Proceedings of the AAAI Conference on Artificial Intelligence, 33*(1), 3681–3688. https://doi.org/10.1609/aaai.v33i01.33013681

[4] Glorot, X., & Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 249-256.

[5] He, K., Zhang, X., Ren, S., & Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision* (pp. 1026-1034). https://doi.org/10.1109/ICCV.2015.123

[6] Jaswanth, P., & Ramprasad, M. V. S. (2023). Deep learning based intelligent system for robust face spoofing detection using texture feature measurement. *Measurement: Sensors*, *29*, 100868. https://doi.org/10.1016/j.measen.2023.100868

[7] Kong, Y., Li, X., Hao, G., & Liu, C. (2022). Face anti-spoofing method based on residual network with channel attention mechanism. *Electronics*, *11*(19), 3056. https://doi.org/10.3390/electronics11193056

[8] Najeebullah, S., Salman, M., & Swati, Z. N. K. (2025). Face spoofing detection technology using deep learning. arXiv. https://doi.org/10.48550/arXiv.2503.19223

[9] Rajpal, A., Sehra, K., Bagri, R., & Sikka, P. (2023). Xai-fr: explainable ai-based face recognition using deep neural networks. Wireless Personal Communications, 129(1), 663-680. https://doi.org/10.1007/s11277-022-10127-z

[10] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision* (pp. 618-626). https://doi.org/10.1109/ICCV.2017.74

[11] Shinde, S. R., Bongale, A. M., Dharrao, D., & Thepade, S. D. (2025). An enhanced light weight face liveness detection method using deep convolutional neural network. *MethodsX*, *14*, 103229. https://doi.org/10.1016/j.mex.2025.103229.

[12] Shinde, S. R., Bongale, A. M., Dharrao, D., Jadhav, D., & Yadav, N. (2025). Enhancing Face Liveness Detection: Novel Deep CNN Architectures for Anti-Spoofing. *Engineering, Technology & Applied Science Research*, *15*(5), 27206-27212. https://doi.org/10.48084/etasr.12431

[13] Singh, R. P., Dash, R., & Mohapatra, R. K. (2026). Unveiling explainability in face anti-spoofing: Hybrid feature extraction with XAI-guided feature aggregation. *Pattern Recognition*, *169*, 111905. https://doi.org/10.1016/j.patcog.2025.111905

[14] Sun, Z., Yan, H., Guo, M., & Hao, Z. (2024). Lightweight face anti-spoofing for improved MobileNetV3. *Journal of Image Processing Theory and Applications, 7*(1), 97–149. https://doi.org/10.23977/jipta.2024.070117

[15] Thiruchelvam, P., Sathiyarasah, S., Paranthaman, T., & Thaneeshan, R. (2023, March). Design Face Spoof Detection using Deep Learning. In *2023 International Conference on Energy, Power, Environment, Control, and Computing (ICEPECC)* (pp. 1-5). IEEE. https://doi.org/10.1109/ICEPECC57281.2023.10209524

[16] Wang, Z., Yu, Z., Zhao, C., Zhu, X., Qin, Y., Zhou, Q., ... & Lei, Z. (2020). Deep spatial gradient and temporal depth learning for face anti-spoofing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 5042-5051). https://doi.org/10.1109/CVPR42600.2020.00509

[17] Woo, S., Park, J., Lee, J. Y., & Kweon, I. S. (2018). CBAM: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision (ECCV)* (pp. 3–19). Springer. https://doi.org/10.1007/978-3-030-01234-2_1

[18] Xing, H., Tan, S. Y., Qamar, F., & Jiao, Y. (2025). Face anti-spoofing based on deep learning: A comprehensive survey. *Applied Sciences*, *15*(12), 6891. https://doi.org/10.3390/app15126891

[19] Yu, Z., Qin, Y., Li, X., Zhao, C., Lei, Z., & Zhao, G. (2022). Deep learning for face anti-spoofing: A survey. *IEEE transactions on pattern analysis and machine intelligence*, *45*(5), 5609-5631. https://doi.org/10.1109/TPAMI.2022.3215850

[20] Zuama, L. R., Susanto, A., Santosa, S., Gan, H. S., & Ojugo, A. A. (2025). High-Performance Face Spoofing Detection using Feature Fusion of FaceNet and Tuned DenseNet201. *Journal of Future Artificial Intelligence and Technologies*, *1*(4), 385-400. https://doi.org/10.62411/faith.3048-3719-62

Authors Biography

S. Karthika is an Assistant Professor in the Department of Information Technology at Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore, India, with 11 years of teaching experience. She is currently pursuing her Ph.D. in Computer Science at the same institution. She holds an M.C.A. and M.Phil. in Computer Science and is qualified in the UGC-NET examination. Her research interests include network security, face biometric authentication, deep learning, and explainable artificial intelligence. She has published six Scopus-indexed conference papers, authored two books, and holds two copyrights.

Dr.G. Padmavathi is a retired Professor of Computer Science from Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore, India. She has over three decades of teaching and research experience. She has published more than 200 research papers in reputed journals and conferences and has successfully guided several Ph.D. scholars. She has also executed multiple funded research projects supported by agencies such as UGC, AICTE, DRDO, and DST. Dr. Padmavathi is a life member of professional bodies including CSI, ISTE, and ISCA, and serves as a reviewer for several international journals and IEEE conferences. Her research contributions focus on developing secure and intelligent systems in computer science and cyber security.