EFFICIENT FACE ANTI-SPOOFING IDENTIFIER NETWORK (FASIN) WITH DEPTH AND NEAR INFRARED DEEP LEARNING METHODS

Mudunuru Suneel[1] and Tummala Ranga Babu[2]
[1]Department of Electronics and Communication Engineering, University College of Engineering, Acharya Nagarjuna University (ANU), Guntur, Andhra Pradesh, India
[2]Department of Electronics and Communication Engineering, RVR and JC College of Engineering, Guntur, Andhra Pradesh, India
E-Mail: suneel007@gmail.com

ABSTRACT

Face anti-spoofing (FAS) is a crucial task in the field of face recognition practices, which aims to detect and prevent attempts to spoof/attack a facial recognition system using fake or manipulated images. In this work, we aimed to develop a novel Face Anti-Spoofing Identifier Network (FASIN) with Depth and Near Infrared (NIR) embeddings, trained on multi modalities using multi stream Convolutional Neural Networks (CNNs). The proposed FASIN model is capable of processing RGB, Depth, and NIR images to extract discriminative features and effectively distinguish between genuine and fake faces. The depth maps and the near infrared (NIR) images are acquired from RGB images by constructing the depth map construction network (DMCN) and near infrared construction network (NIRCN) respectively. The FASIN model comprises three sub-networks: one for processing RGB images, a second for processing depth images, and the other sub-network for processing NIR images. All the sub-networks consist of multiple CNN layers, which extract features at different scales and levels of abstraction. The inherent noise and other variables may reduce the efficacy of CNN, the wavelet spatial attention mechanism has been proposed to support the RGB CNN stream and it is named wavelet attention CNN (WA-CNN). The extracted features are then concatenated using a multi modal feature fusion module to obtain a robust feature representation that is used to classify real and fake faces. An ensemble learning mechanism has been attached to the model to learn the concatenated features effectively. Experimental results obtained on four benchmark datasets (namely, CelebA-Spoof, CASIA-SURF, WMCA, and MSU-MFSD) demonstrate the efficacy of the proposed FASIN model collated with the state-of-the-art methods. The proposed FASIN model achieves high accuracy and low average classification error rates (ACER), indicating its potential for real-world applications in face anti spoofing identification systems.

Keywords: face anti-spoofing (FAS), multi modal feature fusion (MMFF), wavelet attention, depth map construction network (DMCN), near infrared construction network (NIRCN), ensemble learning.

Manuscript Received 1 August 2024; Revised 27 October 2024; Published 22 December 2024

1. INTRODUCTION

Facial recognition (FR) is a vital biometric application that is exploited in a variety of fields such as security, surveillance, and validation. Due to its non-intrusive temperament, face biometrics became one of the extremely convenient modalities for biometric validation. Even though FR systems are approaching human presentation in identifying people in several challenging datasets [1], most of them are still helpless to detect presentation attacks (PA), also acknowledged as spoofing attacks [2], in which an attacker creates face forgeries such as digital displays, printed photographs, masks and then launches spoofing attacks by bestowing the forgeries to face recognition systems camera sensors. Face Presentation Attack Detection (Face PAD), also known as Face Anti-Spoofing (FAS), has been developed to sense fake faces and ensure the precision and security of face recognition systems. [3].

Face anti-spoofing (FAS) systems are critical for the reliable deployment of facial recognition systems [4] to identify and avoid attacks such as print attacks [5], 3D attacks [6], and replay attacks [7]. Face anti-spoofing techniques have been increasingly significant in upgrading face recognition systems in recent years, due to the extensive use of facial recognition in areas such as phone unlocking, access control, financial payments, and monitoring [8]. It is the automatic detection of presentation attacks by discriminating between an authentic face and a presentation attack instrument (PAI) that attempts to imitate genuine biometric attributes. Presentation attacks are outlined as the direct presentation of human traits or artifacts to the input sensor of a biometric system to disrupt its regular procedure. Face PAD methods try to detect presentation attacks by measuring and evaluating anatomical traits or automatic and voluntary reactions.

The majority of available research focuses on detecting replay and print attacks by visible spectral data. Color, texture [9, 10], motion [11], and physiological signals [12] are frequently used for PAD in visible-spectrum imagery. Conventional and CNN-based approaches have demonstrated efficiency in distinguishing between the living and spoofing face by formalising face anti-spoofing as a two-class (binary) classification between spoofing and living of genuine images. Nevertheless, these methodologies make it difficult to investigate the characteristics of spoofing patterns, such as the loss of skin features, colour falsehood, moiré pattern, and spoofing artifacts [13-15].

Earlier, researchers used liveness indicators such as head motion and eye blinking to identify print attacks [16]. Yet, when faced with unspecified attacks, such as pictures with the eye area removed and video replay, these strategies fail. Eventually, the study moved to a broader texture analysis, addressing print and replay attacks. To make a binary choice, researchers primarily use handmade features, such as LBP [17, 18], HoG [19], SIFT [20], and SURF [21], in combination with conventional classifiers, such as SVM and LDA. However, when the testing environments, such as lighting and background are changed, they frequently experience a significant performance decline, which might be interpreted as an overfitting issue. Furthermore, they have problems in dealing with 3D mask attacks [22].

Researchers have sought to tackle the overfitting problem in a variety of methods. Some researchers, for example, extract spoofing features in HSV+YCbCR space, whilst others investigate features in the temporal domain [23, 24]. Later work updated the data with image patches and fused the scores from the patches to a single decision [25].

The topic of detecting unspecified face spoof attempts, known as zero-shot face anti-spoofing (ZSFA) [26], is a new and uncertain challenge for the research community. Researchers are highly engaged in the generalisation of anti-spoofing develops, or how effectively they detect spoof/presentation attacks that were never detected during training. Image-based face anti-spoofing is concerned with face anti-spoofing algorithms that only use RGB photos as input, with no additional features such as depth or heat maps.

Another critical part of ZSFA is the selection of the most relevant features for detecting unknown spoof attacks. Conventional feature extraction methods may not be sufficient to capture the characteristics of unknown spoof attacks, necessitating more complex and extensive feature extraction methods. Deep learning is one promising way to address the ZSFA problem. Deep learning algorithms have demonstrated great effectiveness in a diversity of computer vision tasks, comprising face recognition and anti-spoofing. Deep learning algorithms, in particular, can train high-level features that are more robust and discriminative than hand-crafted features.

To address the ZSFA issue, several deep learning based approaches have been developed. One option is to utilise generative adversarial networks (GANs) to produce synthetic spoof photos and train anti-spoofing models on both actual and synthetic spoof images [26]. Another strategy is to practice transfer learning, in which pre-trained models are fine-tuned on a small collection of known spoof attacks before being evaluated on unknown spoof assaults.

While deep learning based approaches have shown positive outcomes in terms of ZSFA, there are still various problems to overcome. One significant challenge is the scarcity of large-scale datasets that cover a wide range of spoof attacks. Another problem is generalising anti-spoofing models across domains such as various lighting conditions and camera settings. Moreover, deep learning based approaches require a considerable quantity of training data and processing resources, which may be insufficient in some real-world settings.

While great progress has been achieved in identifying known spoof attacks, the ZSFA problem remains an unresolved challenge that requires additional investigation. Deep learning-based algorithms have shown significant promise in addressing the ZSFA problem, but more work is needed to increase their resilience and generalisation capabilities. Finally, the creation of effective anti-spoofing technologies will be critical in maintaining trust and confidence in facial recognition systems and preventing identity fraud and other malicious acts.

In conclusion, this paper presents the following key contributions:

a) This work suggests exploring the advantages of various modalities to support the task of identifying face spoofing attacks. Depth map construction network (DMCN) and near infrared construction network (NIRCN) were implemented for depth and NIR retrieval from RGB data.
b) Multi stream CNN architecture has been proposed to be used to train on different modalities of data. As the attention mechanism boosts the important feature to be in the race, spatial wavelet attention is included in the RGB CNN stream (RGB WA-CNN).
c) Multi modal feature fusion (MMFF) has been proposed for improving the accuracy in identifying spoofing attacks.
d) It presents an ensemble learning strategy to learn discriminative fused features extracted from multiple CNN streams.
e) Comprehensive tests were carried out, and the outcomes show that the proposed scheme has advanced the state-of-the-art in terms of performance on a number of different public benchmarks.

The remainder of the paper is prepared as follows: section 2 presents a detailed literature review of the existing works in face anti-spoofing (FAS). Data processing and multi-modality data construction are detailed in section 3. Section 4 presents the proposed methodology, its workflow, and the training and testing strategies. Results and discussion were carried out in section 5. Finally, section 6 depicts the conclusions and the future working directions.

2. LITERATURE REVIEW

Face recognition technology has become a fundamental part of our everyday lives, with countless applications. However, privacy and security issues have been highlighted due to the fact that this technology is susceptible to spoofing attacks by malicious actors. In recent years, various researchers have presented anti-spoofing algorithms that use depth and near-infrared (NIR) images to boost the accuracy and generalisation capability of the model.

In their work, Atoum *et al*. [25] developed a depth-based anti-spoofing system that combines a combination of deep neural networks and 3D facial surface information to detect spoofing attempts. The method generates depth maps from RGB images and utilises them as input to a CNN-based network to obtain a binary classification result. The scientists also presented a patch-based technique that captures local depth characteristics from facial regions and feeds them into a separate CNN for classification. Experimental results demonstrated that the suggested solution outperformed standard 2D-based anti-spoofing methods. In another paper, Li *et al*. [27] presented a multi-channel CNN for anti-spoofing detection using depth and colour data. The method produces depth maps and colour pictures from the input facial image and feeds them into separate CNNs for feature extraction. The fused features are then sent into a final classification layer. The authors also presented a data augmentation technique that employs random rotation and translation to enhance the model's generalisation capabilities. Using multiple benchmark datasets, experimental results demonstrated that the suggested technique attained state-of-the-art performance. Liu *et al.* [28] introduced a system that integrates depth and NIR information for anti-spoofing detection. The method collects depth maps and NIR pictures from the input face and feeds them to different CNNs to extract features. The combined features are then sent into a final classification layer. The authors also suggested a feature fusion method that integrates depth and NIR information at multiple levels of the network. Experimental results show that the suggested method outperformed traditional 2D-based anti-spoofing methods and methods that employ only depth or NIR information on multiple benchmark datasets. Boulkenafet *et al.* [29] proposed a system for anti-spoofing detection that integrates RGB, depth, and NIR information. The method extracts RGB images, depth maps, and NIR images from the input face and feeds them into separate CNNs for feature extraction. The features are then concatenated and supplied into a final classification layer. The authors also suggested a feature selection approach that selects the best discriminative features from each modality to minimise the dimensionality of the input. Experimental results showed that the suggested method achieved state-of-the-art performance on several standard datasets, outperforming traditional 2D-based anti-spoofing methods and methods that use only one or two modalities. Jourabloo *et al.* [30] resolved the face anti-spoofing by inversely decomposing a spoof/ fake face into the live/ genuine face and the spoof noise pattern, then used the spoof noise for classification. Kim *et al.* [31] offer a new dataset to differentiate between mask materials and facial skin by using radiance measures. Zhang *et al*. [32] proposed a method that combines depth and colour information for anti-spoofing detection. The method collects depth maps and colour images from the input face and feeds them into independent CNNs for feature extraction. The features are then concatenated and supplied into a final classification layer. The authors also proposed a spatial attention mechanism that dynamically weights the importance of different facial regions based on their discriminative power. Testing results demonstrated that the suggested method achieved state-of-the-art performance on multiple benchmark datasets, beating classic 2D-based anti-spoofing algorithms and methods that employ only depth or colour information.

In addition to the above-mentioned works, there have been several other studies that have investigated the use of depth and NIR information for anti-spoofing detection. For example, Wang *et al*. [33] offered a method that uses depth information to detect 3D mask assaults, while Nguyen *et al*. [34] proposed a way that uses NIR information to detect print attacks. These studies demonstrate the potential of depth and NIR information in improving the accuracy and generalisation capability of anti-spoofing methods. Yu *et al*. [35] presented a novel frame-level FAS approach built on Central Difference Convolution (CDC) that can capture intrinsic detailed patterns by combining gradient information and intensity. Stehouwer *et al*. [36] developed a GAN-based method for generating and identifying noise patterns from known and unknown medium combinations. Liu *et al*. [37] proposed a spoof trace disentanglement network (STDN) as a hierarchical collection of patterns at several sizes to extricate spoof traces from input faces. Xu *et al*. [38] proposed a deep neural network architecture that combines CNN and long short-term memory (LSTM) units. Pinto *et al*. [39] used the shape-from-shading (SfS) technique to generate reflectance, depth, and albedo maps as input to the SfSNet to examine the material difference between authentic and spoofing faces. Wen et al. [40] claim that texture features comprise the information about personal identity, which is surplus for anti-spoofing and could guide poor generalization performance. Therefore, few research findings propose handcrafted features built on image quality and distortion analysis for the anti-spoofing task [40-42]. Furthermore, several approaches in [43], [11], and [44] extracted dynamic texture features from numerous video frames to assess motion data in the temporal domain rather than the spatial domain.

Depth and NIR information have emerged as viable modalities for increasing the accuracy and generalisation capability of face anti-spoofing algorithms. These modalities add complementary information to RGB images and can help to overcome the limits of classic 2D-based anti-spoofing approaches. Many researchers have presented deep learning-based algorithms that exploit depth and NIR information for anti-spoofing detection, and have achieved state-of-the-art performance on various standard datasets. However, there is still a need for further study to explore the efficiency of these modalities in real-world circumstances and to develop approaches that can handle more sophisticated spoofing attacks.

3. DATA PROCESSING

3.1 Datasets

A. CelebA-Spoof [45]: The CelebA Spoof dataset was developed for anti-spoofing faces. It is made up of approximately 10,000 colour images of 10-second recordings of people executing one of three actions: live, print attack, or replay attack. The dataset was generated by researchers at the University of Albany, State University of New York, and it is an extension of the original CelebA dataset, which is a large-scale face attribute collection containing over 200,000 celebrity photos. The dataset contains a variety of attributes that can be used in anti-spoofing studies. Each video is tagged with the sort of action taken (live, print, or replay attack) as well as to identify the individual in the video. The dataset contains images of various quality and resolution, as well as a variety of positions, expressions, and lighting situations. The dataset also includes a set of masks for identifying the location of the face in each image, which is important for training machine learning models.

B. CASIA-SURF [46]: The CASIA-SURF is a large-scale benchmark dataset for evaluating image based face recognition systems. It was created by the Chinese Academy of Sciences' Institute of Automation (CASIA) in collaboration with the Surveillant Camera-based Face Recognition Group (SURF) at the University of Maryland. The dataset contains 10,575 images of 486 individuals, with an average of 21.75 images per subject. The images were captured under varying illumination conditions, facial expressions, and poses, making it challenging for face recognition algorithms to correctly identify individuals. The images were captured using a Canon EOS 10D digital camera with a resolution of 1,536 x 1,024 pixels. The subjects were asked to stand in front of a plain background with a neutral facial expression. The dataset also includes manually annotated landmarks for each image, which can be used to align the faces for pre-processing. The CASIA-SURF dataset has become a widely used benchmark dataset in the field of face recognition and has been used to evaluate the performance of many state-of-the-art face recognition algorithms.

C. WMCA [47]: The wide multi-channel presentation attack (WMCA) dataset is a set of spoofed (fake) biometric samples created to test biometric recognition systems. The dataset includes printed pictures, replay attacks, and 3D masks, among other presentation attacks (also known as spoofing attacks). The dataset was developed by the University of Notre Dame in collaboration with the Warsaw Institute of Technology and the University of Cagliari. The dataset contains around 60,000 samples separated into two subsets: training and testing. The training set has over 50,000 samples, and the testing set has over 10,000 samples. Each sample is labeled with information on the presentation attack employed, the sensor used to take the biometric sample, and other pertinent information.

Table-1. Showing the dataset summary.

Dataset Name	Modalities	Attack Types	Details of Samples			
			Type	Number of subjects	Genuine face	Spoof face
CelebA-Spoof [45]	RGB	Print, 3D Mask, Paper cut, Photo attack	Images	10177	202599	422938
CASIA-SURF [46]	RGB	3D Mask attack	Videos	48	288	846
WMCA [47]	RGB+Depth+NIR	Printed pictures, replay attacks, 3D masks	Images	72	347	1594
MSU-MFSD [48]	RGB	Printed pictures, replay attacks	Videos	35	-	-

D. MSU-MFSD [48]: The MSU-MFSD dataset serves as a standard for assessing the performance of face anti-spoofing algorithms. It was developed by Michigan State University's Biometrics and Pattern Recognition Lab (MSU). The collection contains spoofing attacks such as printed pictures, replay assaults, and video attacks. The spoofing presentations are collected using various devices, under various lighting circumstances, and with various forms of spoofing attacks. The dataset also contains tough conditions, such as wearing spectacles or having facial hair, which might make face identification more difficult. Table-1 gives the summary of the all benchmarking datasets used in this work.

3.2 Data Augmentation

Data augmentation can be a useful technique for improving the performance of anti-spoofing models that aim to detect fake biometric traits such as fake faces, voices, or fingerprints. In this context, data augmentation involves creating additional variations of the genuine and spoofed samples in the training dataset by applying various transformations to the original data. In the case of face spoofing, data augmentation techniques could include adding random noise to the images, rotating, flipping, or cropping the images, changing the brightness or contrast, and applying blur or distortion filters [49]. One important consideration when applying data augmentation to spoofing datasets is to ensure that the augmented samples still represent realistic spoofing attacks. For instance, if the spoofing attack involves using a printed photo of a genuine user, the augmented samples should reflect variations of that scenario, such as different lighting conditions or printing quality. Another important consideration is to balance the distribution of genuine and spoofed samples in the augmented dataset. If the original dataset is imbalanced, the augmented dataset could lead to further imbalances, which could affect the model's ability to detect spoofing attacks accurately. Figure-1 shows the sample images of augmented data.

Figure-1. Sample images for data augmentation.

3.3 Depth Map Construction

Depth information can be useful in image classification tasks using deep learning algorithms for several reasons. Depth information can help improve the accuracy of image classification algorithms by providing additional spatial information about the objects in an image. For example, knowing the depth of an object can help distinguish between objects that appear similar in 2D but have different depths. Depth information can provide valuable information about the structure of a scene, such as the distance between objects and their relative positions in 3D space. This can help improve the overall understanding of a scene and make it easier to classify objects within it. When objects in an image are partially occluded by other objects, depth information can handle occlusions more effectively and it provides valuable additional information that can help improve the accuracy and efficiency of image classification using deep learning algorithms.

As the depth maps provide additional information, we proposed to use depth information in recognizing the genuine and spoofed faces. The RGB images of both genuine and spoofed face categories were transformed into depth maps using the method proposed in [50].

This section presents a Convolutional neural network that employs transfer learning to generate a depth map with high resolution from a single RGB image.

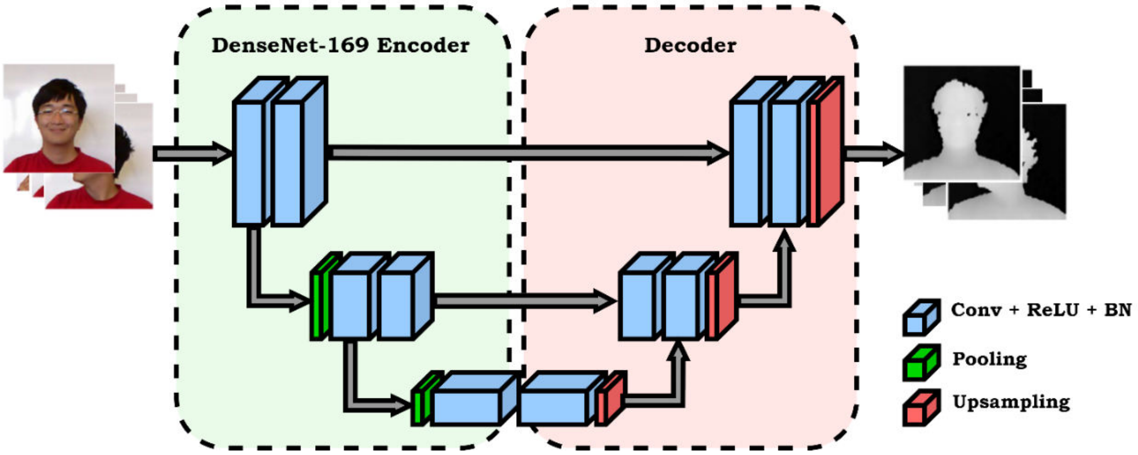

Figure-2. Process of depth map generation from RGB image.

In Figure-2, we can see a high-level diagram of the encoder and decoder network used for estimating the depth. In the encoder, DenseNet-169 [51] architecture, pre-trained on ImageNet [52], is utilised to turn the input RGB face image into a feature vector. After passing this vector through a series of up-sampling layers, we end up with a depth map with half the resolution of the original. These up-sampling tiers and their corresponding skip connections form the decoder. Even though state-of-the-art methods have recommended adding sophisticated layers such as Batch Normalization, the decoder does not use them. Starting with the same number of output channels as the truncated encoder, a 1x1 convolutional layer is used for the decoder. Up-sampling blocks are then added together one by one. Each unit incorporates a bilinear up-sampling of 22 samples, two convolutional layers of 33 samples each, and output filters that are each tuned to receive half as many input samples. Then, the output of the previous layer and the pooling layer from the encoder, both of which have the same spatial dimension, are fed into the first convolutional layer of the two. Each block is then followed by a leaky ReLU activation function with parameter = 0.2, except for the last up-sampling block. The photos' original colours in the [0, 1] range are used as inputs rather than any kind of normalisation of the input data.

In-depth regression, the standard loss function takes into account the discrepancy of the ground-truth depth map. y and prediction of the network $\hat{y}$. This paper proposes a biased sum of three loss functions as the definition of the loss. L between y and $\hat{y}$.

$$Loss = L_{depth} + L_{gradient} + L_{SSIM}$$

The loss L_{depth} is the point-wise loss which is calculated on the depth standards:

$$L_{depth}\left(y_{in}, y_{pred}\right) = \left(\frac{1}{m}\right)\sum_{i=1}^{m}\left|y_{in} - y_{pred}\right|$$

The loss $L_{gradient}$ is the loss derived on the gradient of a depth image g_{im}

$$L_{gradient}\left(y_{in}, y_{pred}\right) = \left(\frac{1}{m}\right)\sum_{i=1}^{m}\left|g_{imX}\right| + \left|g_{imY}\right|$$

where g_{imX} and g_{imY} compute the changes in the x and y components of depth image gradients.
The L_{SSIM} loss is written as:

$$L_{SSIM}\left(y_{in}, y_{pred}\right) = \frac{1 - SSIM\left(y_{in}, y_{pred}\right)}{2}$$

Figure-3 shows the sample RGB and its estimated depth maps.

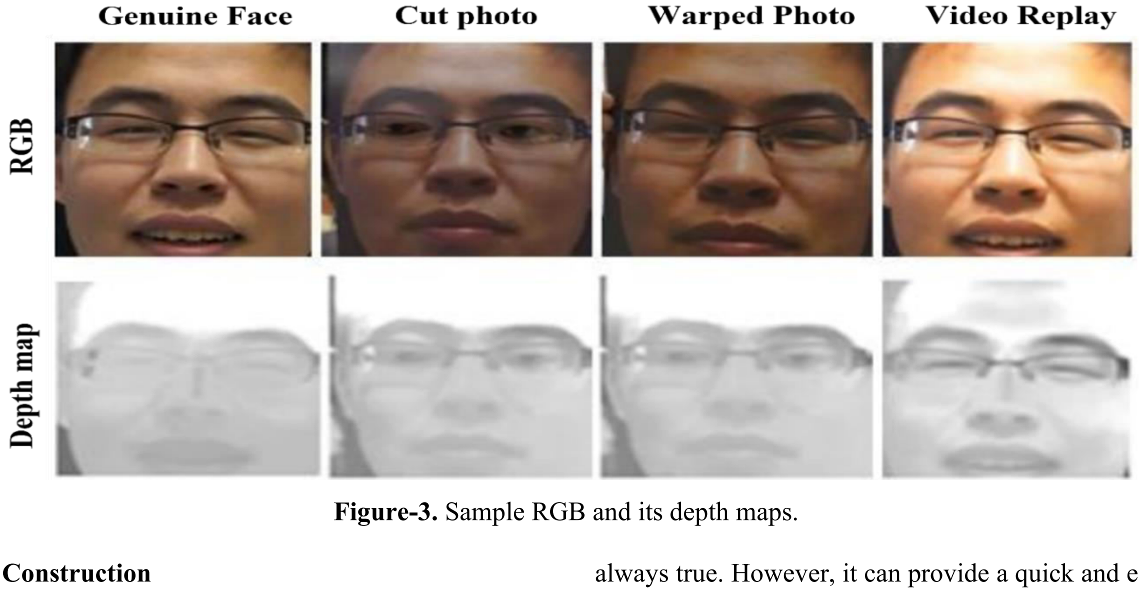

Figure-3. Sample RGB and its depth maps.

3.4 NIR Construction

Near-infrared (NIR) images can be constructed from RGB images by using various methods, including:

Channel substitution: In this method, the red channel of the RGB image is replaced with the NIR channel. This method assumes that the NIR information is present in the red channel of the RGB image, which is not always true. However, it can provide a quick and easy way to generate NIR-like images.

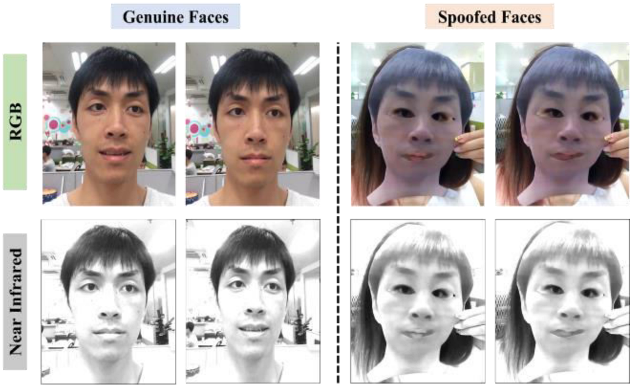

Figure-4. Sample images showing the RGB images and NIR converted images.

Multispectral imaging: Multispectral imaging involves capturing images of a scene at multiple wavelengths, including NIR. This method requires specialized equipment, such as a multispectral camera or a modified RGB camera with a NIR filter. The captured images can then be combined to generate an NIR image.

Inverse modelling: Inverse modelling involves using a mathematical model to estimate the NIR values from the RGB values. This method requires knowledge of the spectral characteristics of the scene and the camera used to capture the RGB image. The model can then be used to estimate the NIR values for each pixel in the image.

Deep learning: Deep learning techniques, such as convolutional neural networks (CNNs), can be trained to generate NIR images from RGB images. This method requires a large dataset of paired RGB and NIR images for training the network. Once trained, the network can generate high-quality NIR images from RGB images.

In this work we used CNN to train on paired RGB and NIR images further to estimate the NIR images for the given RGB images. Figure-4 shows the sample NIR images converted from RGB.

4. PROPOSED METHODOLOGY

In this section, we discuss the overview of the proposed face spoofing detection system. Later, we present a detailed architecture of the proposed system.

4.1 System Overview

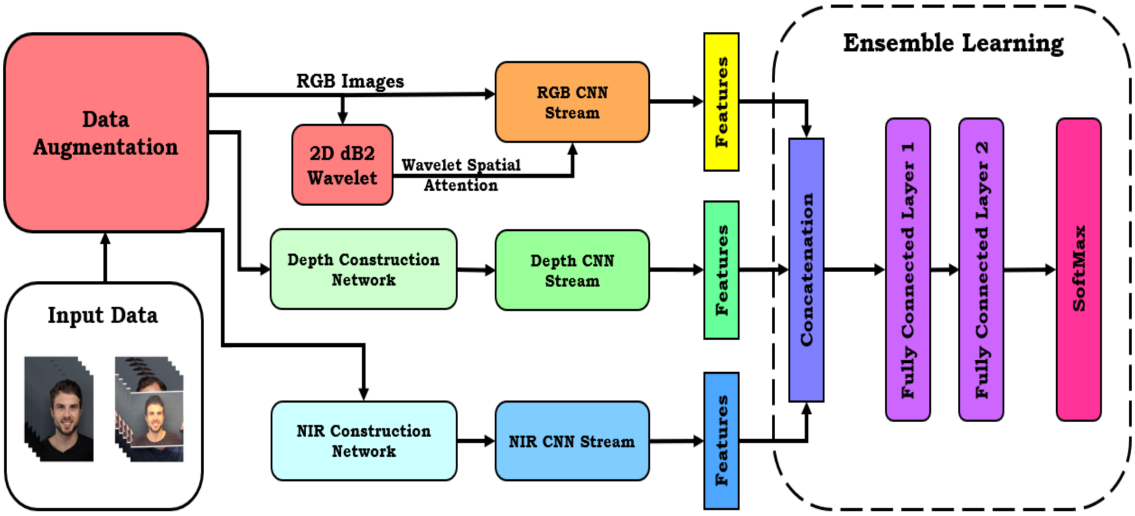

Figure-5. Block diagram representing the proposed system overview.

A Face Anti-Spoofing Identifier Network (FASIN) has been proposed in this work to detect face spoofing activity. The system is proposed to use multiple CNN streams to train on multi modal data. Figure-5 shows a detailed overview of the proposed FASIN working mechanism.

Data augmentation can be particularly useful when working with small training datasets, where the lack of data can lead to overfitting and poor model performance. Hence, we applied data augmentation to increase the size of the dataset.

The RGB CNN stream will take the raw images as input and extract the features in the convolutional layers. An RGB CNN stream will receive spatial attention to boost the performance. Wavelet attention mechanism has gained popularity in recent years for its effectiveness in improving the performance of convolutional neural networks (CNNs). In CNNs, attention mechanisms are used to selectively focus on certain parts of the input data, allowing the network to better capture relevant features and reduce noise. Wavelet attention is particularly useful in CNNs because it enables the network to efficiently analyze both low-frequency and high-frequency features of the input data. This is important because different features may have different spatial frequencies, and traditional CNNs are better at capturing high-frequency features than low-frequency features. By incorporating wavelet attention into a CNN, the network can more effectively capture features at all spatial frequencies, leading to improved performance. This RGB CNN stream with a wavelet attention mechanism will produce the RGB feature space.

Depth data helps to accurately recognize objects in an image. Depth data provides crucial information about the spatial relationships and geometry of objects in a scene. Depth data can be used to enhance the quality of images, such as by providing more accurate focus and depth-of-field effects. Depth maps were constructed for the inputted RGB images using the depth construction network (DCN) and trained on depth CNN stream to produce depth feature space.

NIR images often have better contrast and resolution than visible light images, which can lead to improved accuracy in machine learning tasks. The increased contrast can make it easier to distinguish between objects, while the higher resolution can provide more detail for analysis. Hence, we proposed to use near infrared (NIR) images to improve the accuracy of the spoofing detection mechanism. The RGB images were converted to NIR images using a NIR construction network. An NIR CNN stream takes the NIT images as inputs and produces a NIR feature space.

All the features extracted from three different CNN streams were concatenated and further trained on a fully connected dense layers. An ensemble learning mechanism was adopted to learn the details from three different modalities. Finally, the SoftMax layer produces the probability scores to identify the spoofed and genuine faces. The next section will discuss the proposed architecture.

4.2 Proposed System Architecture

The system is proposed to use three streams of convolutional neural networks. Every individual stream is fed with RGB, Depth maps, and NIR images respectively. Figure-6 shows the proposed architecture with three CNN streams, feature concatenation, and an ensemble learning strategy on the concatenated features.

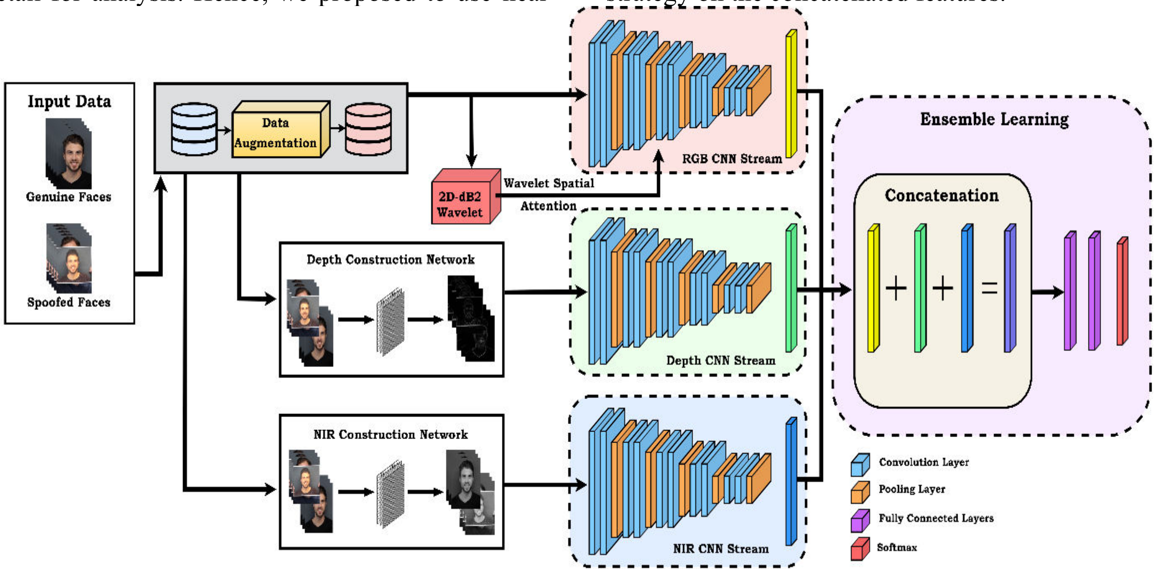

Figure-6. Proposed architecture with three CNN streams and ensemble learning block.

A. Convolutional neural network

The neurons that make up a convolutional neural network are organised in a hierarchical structure [53]. The formula for a single neuron is a function taking in I_{ij} and returning an output y. The notation for this function is:

$$y^{k} = f\left(I_{ij}^{(k-1)} * w^{(k-1)} + b^{(k-1)}\right)$$

where the scalar b represents the neuron's bias and the vector w represents the weights. The term "activation function" refers to the function $f(\square)$. For this purpose, we employ the Rectified Linear Unit activation function (ReLU). Training involves iteratively trying out different values for a set of parameters (such as weights and biases) until the neural network achieves an optimal fit between inputs and outputs. Typically, a loss function is defined at the outset of training with the backpropagation algorithm, which is the most popular method. For a single training example, the loss function L can be defined as an ℓ_2 Norm of the error.

$$L\left(I_{ij}^{n}, y^{n}\right) = \frac{1}{2}\left\| y^{\hat{n}} - y^{n} \right\|^{2}$$

Where, $h_{(w,b)} * I_{ij}$ is the neuron output $y^{\hat{n}}$. For N training examples, the same loss can be given as

$$L\left(I^{n}_{ij}, y^{n}\right)=\frac{1}{n}\sum_{n=1}^{N}\left\| y^{\hat{n}}-y^{n}\right\|^{2}$$

Backpropagation uses gradient descent to look for the best solution by reducing the loss function to its minimum value.

As can be seen in Figure 6, the deep architecture built on top of CNN specifically for face spoofing detection consists of 15 hidden layers (10 convolution layers and 5 pooling layers). Ten convolution layers acquire the knowledge of convolution kernels with sizes of 5×5. The average pooling is implemented with a factor of 2, and there are five pooling layers. All three streams namely RGB with wavelet attention CNN stream, depth CNN stream, and NIR CNN streams use the above-mentioned CNN mechanism for learning the features. The flattened layers in three streams provide the three feature vectors. fv_{RGB}, fv_{depth} and fv_{NIR}. These three feature vectors from three different CNN streams were concatenated to form a single feature vector fv_{c}.

$$fv_{c} = fv_{RGB} + fv_{depth} + fv_{NIR}$$

Further this concatenated feature vector fv_{c} is learned by the ensemble learning mechanism. The sigmoid function is incorporated into the deep architecture because face anti spoofing detection is fundamentally a two-class classification problem.

B. CNN with spatial wavelet attention

The human visual system allows us to efficiently capture and focus on essential areas through a sequence of glimpses while we observe a scene or item. Many works take this as inspiration and incorporate the attention mechanism to help models focus on the most important details as they learn. An abstract framework for describing the global context of the data can be derived from the structure of attention blocks. Features may be irreparably damaged by a down-sampling process in CNN. The discrete wavelet transform (DWT) is an optimal solution since it not only does the down-sampling process with good quality but also retrieves the finer details of the feature map. Hence, we propose adding the wavelet attention (WA) block to the RGB CNN stream, based on the global context modelling framework taken from [54].

The WA block applies DWT on the input image to generate a low-frequency component I_{LL} and three high-frequency components I_{LH}, I_{HL}, and I_{HH}. The low-frequency component prevents down-sampling from corrupting the primary information structure of the feature map. These high-frequency components retain a great deal of noise while preserving the specific information present in the data. To prevent interference from noise, we have decided to eliminate I_{HH}. Mathematically, the WA block is given as

$$\zeta = Aggregation\left(I_{LL}, Attention_{gen}\left(I_{LL}, Soft\max\left(Aggregation\left(I_{LH}, I_{HL}\right)\right)\right)\right)$$

The wavelet attention block can be simply added to CNN by substituting the max-pooling or average-pooling layer, however adding it to other locations requires modifying network structures.

C. Ensemble learning

When numerous learning models are combined into one, an "ensemble" model is created that outperforms any of the individual models. Because a group of learners may often outperform an individual, ensemble learning is frequently employed in data analytics issues. The ensemble learning method of confidence averaging averages the probabilities of classification from many base learners to determine which category should be assigned the highest degree of certainty [55]. When used in deep learning models, SoftMax layers can generate a posterior probability list that details the degree to which each class was correctly identified. The final classification result is based on the class label that has the highest average confidence value, which is determined by summing the base learners' probabilities of classification for each class. The SoftMax function is utilised to derive the confidence value for each category and is mathematically given as:

$$\text{Softmax}\left(fv_{c}\right)_{i} = \frac{e^{\left(fv_{c}\right)_{i}}}{\sum_{j=1}^{K} e^{\left(fv_{c}\right)_{j}}}$$

Where, fv_{c} is the concatenated input feature vector and the number of classes is denoted by K. $e^{\left(fv_{c}\right)_{i}}$ and $e^{\left(fv_{c}\right)_{j}}$ are the exponential components of both input and output respectively.

The predicted or anticipated class label that can be derived through the use of the confidence averaging method is denoted by:

$$y_{predicted} = \text{argmax}\,\frac{\sum_{j=1}^{n} p_{j}}{n}$$

Where n is the selected learners and p_{j} is the class confidence value. Confidence averaging helps the ensemble learning model to detect uncertain classification results and fix the wrongly classified samples, whereas the traditional voting technique simply evaluates the class labels. All ensemble models' computational complexity is proportional to the complexity of their base learners.

Another DL model ensemble approach is concatenation [56]. With the use of concatenate procedures, a concatenated CNN takes the highest-order features produced by the basic CNN models' top dense layer and combines them into a concatenated layer that comprises all the features/ gradients. After the concatenated layer, a drop-out layer is used to eliminate superfluous features, and a SoftMax layer is added to build a new CNN model. Concatenation's strength lies in its ability to build a holistic new model by fusing elements of the highest degree.

5. RESULTS AND DISCUSSIONS

Since face authorization systems could be compromised by presentation attacks (PAs), testing their resistance to such attacks is essential. To conduct cross-data testing, we employ CelebA-Spoof [45], CASIA-SURF [46], WMCA [47], and MSU-MFSD [48]. This section discusses the various performance protocols used to judge the performance of the proposed system for face anti spoofing identifier networks. Later this section presents the results obtained during various experiments designed. Further, a detailed discussion of the results was carried out.

5.1 Evaluation Protocols and Metrics

In this section, we present the quantitative and qualitative analysis of the proposed method. We used three metrics namely attack presentation classification error rate (APCER), bona fide presentation classification error rate (BPCER), and average classification error rate (ACER) to quantitatively assess the effectiveness of the proposed approach. The attack presentation classification error rate is defined as follows:

$$APCER = \frac{1}{N_{fake}} \sum_{k=1}^{N_{fake}} \left(1 - R_k\right)$$

$$BPCER = \frac{\sum_{k=1}^{N_{bonafide}} R_k}{N_{bonafide}}$$

$$ACER = \frac{APCER + BPCER}{2}$$

Where N_{fake} is the number of presentation attacks of fake faces, $N_{bonafide}$ is the number of genuine or bona fide face images and R_k is given as:

$$R_k = \begin{cases} 1 & \text{if the input is classified as presentation attack} \\ 0 & \text{if the input is classified as genuine of bona fide} \end{cases}$$

5.2 Training/ Testing Data Manifestation and Experimental Setup

We evaluated the performance of the proposed FASIN in different strategies. Firstly, the RGB WA-CNN sub-network has been implemented on all the datasets. Secondly, we implemented the combination of WA-CNN and depth stream CNN with both the modalities (RGB + Depth). Next, our proposed method with three sub-networks consisting of all three streams with three modalities (RGB + Depth + NIR) is implemented. Further, cross data validation is carried out to examine the robustness of the FASIN. Finally, FASIN has been compared to other existing state-of-the-art methods to conclude the effectiveness of the findings in this work. The entire system is implemented in the Python platform using Keras with TensorFlow backend. We have used high performance computer with two 6 GB Tesla K20MsGPUs from NVIDIA to train and test the model on various benchmark datasets. 80% of the data samples were used for training and the remaining 20% samples were used for validation and testing.

5.3 Performance Evaluation of FASIN using RGB WA-CNN Stream

As a first attempt, we initiated the training on the anti-spoofing datasets with RGB WA-CNN stream. The training was carried out for 200 epochs and the achieved results are presented in Table-2. Moderately acceptable performance was identified with the single stream RGB WA-CNN.

Table-2. Performance of the FASIN using only RGB stream CNN with wavelet attention (RGB WA-CNN).

Data Set	Modality	Performance Metrics					
		Without Attention			With wavelet attention		
		APCER (%)	BPCER (%)	ACER (%)	APCER (%)	BPCER (%)	ACER (%)
CelebA-Spoof [45]	RGB	18.53	16.79	17.66	13.44	12.03	12.74
CASIA-SURF [46]	RGB	19.56	17.80	18.68	14.86	11.54	13.20
WMCA [47]	RGB	18.26	16.22	17.24	12.96	10.69	11.83
MSU-MFSD [48]	RGB	18.60	16.65	17.63	13.02	10.58	11.80
Average Performance		18.74	16.87	17.80	13.57	11.21	12.39

Table-2 presents the values for various performance metrics on four datasets considered for the experimentation. This experiment uses only pre-processed augmented RGB image data as input. An average APCER, BPCER, and ACER of 18.74 %, 16.87%, and 17.80% is achieved with the RGB CNN stream without an attention mechanism. We also trained the RGB CNN stream with wavelet attention (RGB WA-CNN) and an average ACER of 12.39% has been achieved. From the ACER values, it is noted that the inclusion of an attention mechanism has greatly improved the performance of FASIN. However, these values were further improved by adding one more stream (D-CNN) with depth map modality in the next experiment.

5.4 Performance Evaluation of the FASIN using RGB WA-CNN and Depth CNN Stream (D-CNN)

As a second experiment, we implemented FASIN with two streams of CNN (RGB WA-CNN + D-CNN) with two modalities, RGB and depth maps respectively. When compared to the first experiment, the ACER values were greatly accelerated during the second experiment. The performance of two stream CNN with RGB and depth embeddings trained via ensemble learning strategy was evaluated over three performance protocols APCER, BPCER, and ACER.

Table-3 depicts the values achieved during this particular experiment.

Table-3. Performance of the FASIN using RGB WA-CNN and depth CNN stream.

Data Set	Modality	Performance Metrics		
		APCER (%)	BPCER (%)	ACER (%)
CelebA-Spoof [45]	RGB + Depth	8.19	5.02	6.61
CASIA-SURF [46]	RGB + Depth	11.36	7.01	9.19
WMCA [47]	RGB + Depth	7.44	3.16	5.30
MSU-MFSD [48]	RGB + Depth	6.35	2.23	4.29
Average Performance		8.34	4.36	6.35

From Table-3, an average APCER score of 8.34%, 4.36 % of BPCER, and 6.35% of ACER were observed. In conclusion, RGB WA-CNN + D-CNN has improved the performance of our proposed FASIN. Again, to further improve the performance, one more stream with a different modality has been introduced in the third experiment.

5.5 Performance Evaluation of the FASIN using the Proposed Three Stream CNN Architecture

During the third experiment, we added a NIR-CNN stream with near infrared modality to the architecture in the second experiment. We trained the three stream FASIN architecture (RGB WA-CNN + D-CNN + NIR-CNN) with three different modalities, RGB images, depth maps, and NIR images respectively. As the NIR images highlight specific parts of the face, it became a great added advantage for the proposed FASIN. From Table 4, it is observed that the performance of FASIN has improved to the next level with a minimum average ACER of 4.88%.

Table-4. Showing the various performance of the proposed method on different online available datasets.

Data Set	Modality	Performance Metrics		
		APCER (%)	BPCER (%)	ACER (%)
CelebA-Spoof [45]	RGB+Depth+NIR	6.18	4.01	5.09
CASIA-SURF [46]	RGB+Depth+NIR	9.35	6.21	7.78
WMCA [47]	RGB+Depth+NIR	5.42	2.14	3.78
MSU-MFSD [48]	RGB+Depth+NIR	4.36	1.36	2.86
Average Performance		6.33	3.43	4.88

Further, we planned to examine the proposed FASIN on cross data to know the ability of the proposed in real time implementation when unknown data has been inputted. We implemented the cross data validation as a fourth experiment in the next section.

5.6 Cross Data Validation

It's possible to find data distributions that are very similar within the same dataset, but which differ greatly between datasets. As a result, the latter may present greater difficulty. Nonetheless, our method is superior to alternatives in the vast majority of cases and achieves top placement across the board, showing that our techniques are highly adaptable and can fend off even previously unseen attacks. We can see the cross data performance of our FASIN in Table-5.

Table-5. Showing the performance of the proposed method for cross data validation.

Training Data	Testing Data	Performance Metrics		
		APCER (%)	BPCER (%)	ACER (%)
CelebA-Spoof	MSU-MFSD	13.33	10.14	11.74
CASIA-SURF	WMCA	16.50	12.13	14.32
CelebA-Spoof + WMCA	CASIA-SURF	12.58	8.28	10.43
MSU-MFSD + WMCA	CASIA-SURF	11.49	7.35	9.42
Average Performance		13.46	9.46	11.47

We firstly trained with a single dataset and tested with other dataset. Further, the training has been done on two different datasets and the test was carried out with other dataset. An average performance of different training and testing combinations was noted as ACER of 11.47%. Figure-7 shows the accuracy of identifying spoofed faces in different experimental strategies presented in this work on various benchmark datasets.

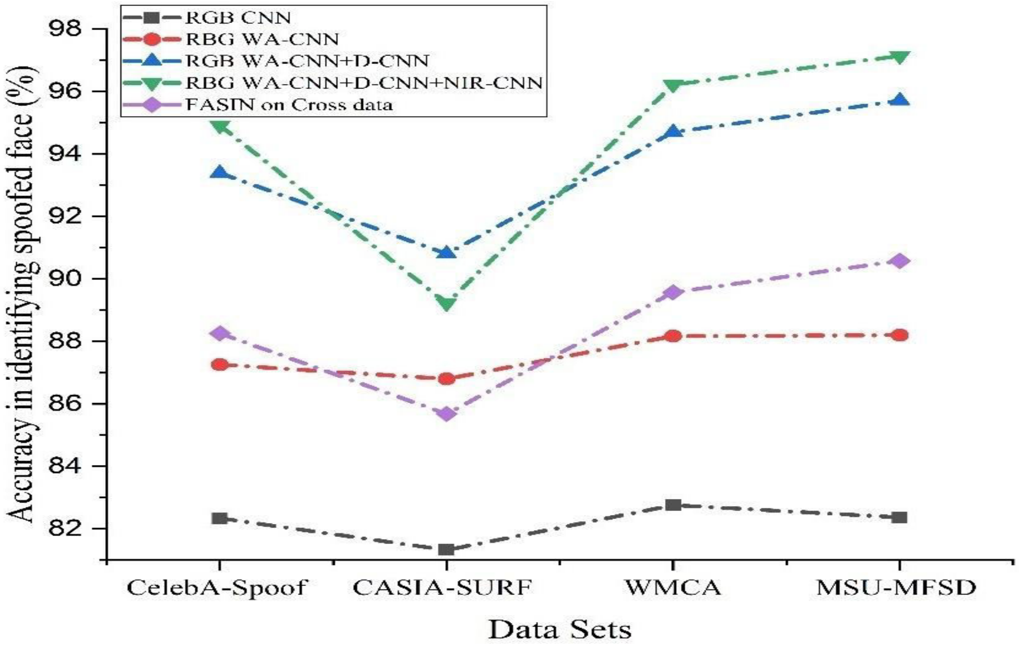

Figure-7. Accuracy plot of FASIN in various experimental strategies on various benchmark datasets.

Table-6. Comparing the Performance of the proposed with other state-of-the-art methods.

Method	Dataset	Modality	ACER (%)
LBP+SVM [57]	SiW-M	RGB	26.90
Patch+BCN [58]	SiW-M	RGB	11.81
CDCN [59]	CASIA-MFSD	RGB	5.32
CDCN [59]	WMCA	RGB	21.12
CMFL [60]	WMCA	RGB+Depth	7.62
MC-ResNetDLAS [61]	WMCA	RGB+Depth	33.45
TransFAS-NHF [62]	WMCA	RGB+Depth	7.01
CNN Meta-Learning [63]	CASIA	RGB	22.70
MsLBP [64]	MSU	RGB	37.20
FASNet [65]	CASIA-RFS	RGB	24.25
FAS-SGTD [66]	CASIA-MFSD	RGB+Depth	15.52
Proposed Multi stream CNN with ensemble learning (FASIN)	CelebA-Spoof [1]	RGB+Depth+NIR	5.09
	CASIA-SURF [2]		7.78
	WMCA [3]		3.78
	MSU-MFSD [4]		2.86

5.7 Compared with Other State-of-the-Art Methods

In this section, we compare the proposed FASIN with various existing methods implemented on different modalities. Table-6 presents the comparison of FASIN with other state-of-the-art methods for identifying presentation attacks. From Table-6, it is observed that our proposed FASIN with three streams (RGB WA-CNN + D-CNN + NIR-CNN) performs better when compared to other methods. Figure-8 depicts the training and validation loss plots of three streams FASIN and RGB WA-CNN stream.

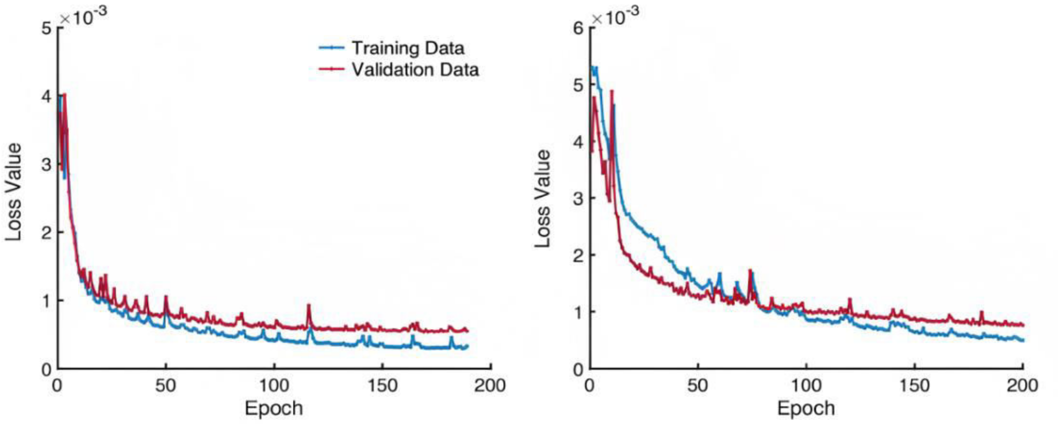

Figure-8. Training and validation error plots for proposed three stream CNN and RGB WA-CNN respectively.

The training lasted for 180 epochs for three streams FASIN and 200 epochs for RGB WA-CNN to achieve minimum loss.

6. CONCLUSIONS

In conclusion, as presented in this paper, the FASIN system is a highly effective method for combating face spoofing attacks in biometric security systems. FASIN can detect and differentiate both real and fake faces with high accuracy by RGB, Depth, and Near-

Infrared embedding's. FASIN study shows that it is a dependable and practical solution with good accuracy rates even when subjected to advanced spoofing tactics. Implementing the FASIN system with three different CNN streams namely, RGB WA-CNN, D-CNN, and NIR-CNN has greatly uplifted the process of identifying presentation attacks (PAs) by differentiating fake and genuine faces effectively. Various experiments were designed strategically to study the performance at various levels. During the first experiment, only a single RGB CNN stream with wavelet attention was implemented. In, the second experiment depth modality has been added and trained with D-CNN as an additional stream. During the third experiment, an NIR stream was added which is trained on NIR modality. AN ensemble learning strategy was adopted to greatly boost the learning capabilities of the proposed FASIN. An average ACER on all datasets was observed as 4.88%. Furthermore, the system's utilisation of innovative technologies such as deep learning and infrared imaging distinguishes it as a cutting-edge answer to the growing need for secure authentication across industries. Overall, the FASIN system makes an important contribution to biometric security, and its implementation has the potential to alter how businesses handle data and physical security. The system's effectiveness, dependability, and practicality make it a valuable addition to the present range of available security solutions, and its further development is anticipated to have a substantial impact on the industry.

Conflict of Interest: The authors declare that we have no conflict of interest.

REFERENCES

[1] E. Learned-Miller, G. B. Huang, A. RoyChowdhury, H. Li and G. Hua. 2016. Labeled faces in the wild: A survey. in Advances in face detection and facial image analysis. Springer. pp. 189-248.

[2] S. Marcel, M. Nixon and S. Li. 2014. Handbook of biometric anti spoofing-trusted biometrics under spoofing attacks. Advances in Computer Vision and Pattern Recognition. Springer.

[3] Information technology Biometric presentation attack detection Part 1: Framework. International Organization for Standardization, Standard, Jan. 2016.

[4] J. Guo, X. Zhu, C. Zhao, D. Cao, Z. Lei and S. Z. Li. 2020. Learning Meta face recognition in unseen domains. In Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 6163-6172.

[5] Z. Zhang, J. Yan, S. Liu, Z. Lei, D. Yi and S. Z. Li. 2012. A face anti spoofing database with diverse attacks. In Proc. 5th IAPR Int. Conf. Biometrics (ICB). pp. 26-31.

[6] N. Erdogmus and S. Marcel. 2013. Spoofing in 2d face recognition with 3d masks and anti-spoofing with Kinect. In Proc. IEEE 6th Int. Conf. Biometrics, Theory, Appl. Syst. (BTAS). pp. 1-8.

[7] I. Chingovska, A. Anjos and S. Marcel. 2012. On the effectiveness of local binary patterns in face anti-spoofing. In Proc. BIOSIG. pp. 1-7.

[8] R. Shao, X. Lan, J. Li and P. C. Yuen. 2019. Multi-adversarial discriminative deep domain generalization for face presentation attack detection. in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 10023-10031.

[9] Z. Boulkenafet, J. Komulainen and A. Hadid. 2015. Face anti-spoofing based on color texture analysis. in Image Processing (ICIP), 2015 IEEE International Conference on. IEEE. pp. 2636-2640.

[10] J. Ma¨att ¨ a, A. Hadid and M. Pietik ¨ ainen. 2011. Face spoofing detection ¨ from single images using micro-texture analysis. In Biometrics (IJCB), 2011 international joint conference on. IEEE. pp. 1-7.

[11] A. Anjos and S. Marcel. 2011. Counter-measures to photo attacks in face recognition: a public database and a baseline. in Biometrics (IJCB), 2011 international joint conference on. IEEE. pp. 1-7.

[12] R. Ramachandra and C. Busch. 2017. Presentation attack detection methods for face recognition systems: a comprehensive survey. ACM Computing Surveys (CSUR). 50(1): 8.

[13] Oeslle Lucena, Amadeu Junior, Vitor Moia, Roberto Souza, Eduardo Valle, and Roberto Lotufo. Transfer learning using convolutional neural networks for face anti-spoofing. In International Conference Image Analysis and Recognition, pages 27–34, 2017.

[14] Chaitanya Nagpal and Shiv Ram Dubey. A performance evaluation of convolutional neural networks for face anti spoofing. arXiv preprint arXiv:1805.04176, 2018.

[15] Xiao Song, Xu Zhao, Liangji Fang and Tianwei Lin. 2019. Discriminative representation combinations for accurate face spoofing detection. Pattern Recognition. 85: 220-231.

[16] K. Kollreider, H. Fronthaler, M. I. Faraj and J. Bigun. 2007. Realtime face detection and motion analysis with application in liveness assessment. In TIFS. 2.

[17] T. de Freitas Pereira, A. Anjos, J. M. De Martino, and S. Marcel. LBP-TOP based countermeasure against face spoofing attacks. In ACCV, 2012. 2

[18] J. Ma¨att a, A. Hadid and M. Pietik ainen. 2011. Face spoofing detection from single images using micro-texture analysis. In IJCB. 1, 2.

[19] J. Komulainen, A. Hadid and M. Pietikainen. 2013. Context based face anti-spoofing. In BTAS. 2.

[20] K. Patel, H. Han and A. K. Jain. 2016. Secure face unlock: Spoof detection on smartphones. In TIFS. 1, 2.

[21] Z. Boulkenafet, J. Komulainen and A. Hadid. 2017. Face anti spoofing using speeded-up robust features and fisher vector encoding. IEEE Signal Processing Letters. 2

[22] S. Liu, B. Yang, P. C. Yuen, and Guoying Zhao. 2016. A 3D mask face anti-spoofing database with real world variations. In CVPRW. 1, 2, 3

[23] W. Bao, H. Li, N. Li and W. Jiang. 2009. A liveness detection method for face recognition based on the optical flow field. In IEEE International Conference on Image Analysis and Signal Processing (IASP). 2.

[24] L. Feng, L. Po, Y. Li, X. Xu, F. Yuan, T. C. Cheung and K. Cheung. 2016. Integration of image quality and motion cues for face anti-spoofing: A neural network approach. Journal of Visual Communication and Image Representation. 1.

[25] Y. Atoum, Y. Liu, A. Jourabloo and X. Liu. 2017. Face anti-spoofing using the patch and depth-based CNNs. In IJCB. 1, 2.

[26] Nguyen Son Minh, *et al.* 2022. Self-Attention Generative Distribution Adversarial Network for Few-and Zero-Shot Face Anti-Spoofing. 2022 IEEE International Joint Conference on Biometrics (IJCB). IEEE.

[27] H. Li, P. He, S. Wang, A. Rocha, X. Jiang, and A. C. Kot. 2018. Learning generalized deep feature representation for face anti-spoofing. IEEE Transactions on Information Forensics and Security. 13(10): 2639-2652.

[28] Y. Liu, A. Jourabloo and X. Liu. 2018. Learning deep models for face anti spoofing: Binary or auxiliary supervision. in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. pp. 389-398.

[29] Z. Boulkenafet, J. Komulainen and A. Hadid. 2017. Face anti spoofing using speeded-up robust features and fisher vector encoding. IEEE Signal Processing Letters. 2.

[30] A. Jourabloo, Y. Liu and X. Liu. 2018. Face de-spoofing: Anti spoofing via noise modeling. In ECCV.

[31] Y. Kim, J. Na, S. Yoon and J. Yi. 2009. Masked fake face detection using radiance measurements. J. Opt. Soc. Amer. A, Opt. Image Sci. 26(4): 760.

[32] S. Zhang *et al*. 2020. CASIA-SURF: A large-scale multi-modal benchmark for face anti-spoofing. IEEE Trans. Biometrics, Behav., Identity Sci. 2(2): 182-193.

[33] Z. Wang *et al*. 2020. Deep spatial gradient and temporal depth learning for face anti-spoofing," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 5042-5051.

[34] Nguyen K., Fookes C., Ross A., Sridharan S. 2017. Iris Recognition with Off-the-Shelf CNN Features: A Deep Learning Perspective. IEEE Access. 6, 18848-18855.

[35] Z. Yu *et al*. 2020. Searching central difference convolutional networks for face anti-spoofing," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 5295-5305.

[36] J. Stehouwer, A. Jourabloo, Y. Liu and X. Liu. 2020. Noise modeling, synthesis, and classification for generic object anti-spoofing. In Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 7294-7303.

[37] Y. Liu, J. Stehouwer and X. Liu. 2020. On disentangling spoof trace for generic face anti-spoofing. In Proc. ECCV. pp. 406-422.

[38] Xu Zhenqi, Shan Li and Weihong Deng. 2015. Learning temporal features using LSTM-CNN

architecture for face anti-spoofing. 2015 3rd IAPR Asian conference on pattern recognition (ACPR). IEEE.

[39] Allan Pinto *et al*. 2020. Leveraging shape, reflectance, and albedo from shading for face presentation attack detection. In: IEEE Transactions on Information Forensics and Security. 15, pp. 3347-3358.

[40] D. Wen, H. Han and A. K. Jain. 2015. Face spoof detection with image distortion analysis. IEEE Trans. Inf. Forensics Security. 10(4): 746-761

[41] J. Galbally and S. Marcel. 2014. Face anti-spoofing based on general image quality assessment. In Proc. Int. Conf. Pattern Recognit., Stockholm, Sweden. pp. 1173-1178.

[42] H. Li, S. Wang and A. C. Kot. 2016. Face spoofing detection with image quality regression. In Proc. 6th Int. Conf. Image Process. Theory, ToolsAppl. (IPTA). pp. 1-6.

[43] T. D. F. Pereira *et al*. 20143. Face liveness detection using dynamic texture. EURASIP J. Image Video Process. 2014(1): 1-15.

[44] J. Komulainen, A. Hadid, M. Pietikäinen, A. Anjos, and S. Marcel. 2013. Complementary countermeasures for detecting scenic face spoofing attacks. In Proc. Int. Conf. Biometrics, Madrid, Spain. pp. 1-7.

[45] Zhang Yuanhan, *et al*. 2020. Celeba-spoof: Large-scale face anti-spoofing dataset with rich annotations. Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16. Springer International Publishing.

[46] Zhang Shifeng, *et al*. 2020. Casia-surf: A large-scale multi-modal benchmark for face anti-spoofing. IEEE Transactions on Biometrics, Behavior, and Identity Science. 2.2: 182-193.

[47] George Anjith, *et al*. 2019. Biometric face presentation attack detection with a multi-channel convolutional neural network. IEEE Transactions on Information Forensics and Security. 15: 42-55.

[48] Wen Di, Hu Han and Anil K. Jain. 2015. Face spoof detection with image distortion analysis. IEEE Transactions on Information Forensics and Security. 10.4: 746-761.

[49] Mikołajczyk, Agnieszka and Michał Grochowski. 2018. Data augmentation for improving deep learning in image classification problems. 2018 international interdisciplinary PhD workshop (IIPhDW). IEEE.

[50] Alhashim Ibraheem and Peter Wonka. 2018. High quality monocular depth estimation via transfer learning. arXiv preprint arXiv: 1812.11941.

[51] Huang Gao, *et al*. 2017. Densely connected convolutional networks. Proceedings of the IEEE conference on computer vision and pattern recognition.

[52] Deng Jia, *et al*. 2009. Imagenet: A large-scale hierarchical image database. IEEE conference on computer vision and pattern recognition. Ieee.

[53] Y. Hao, Q. Li, H. Mo, H. Zhang, and H. Li. 2018. AMI-net: Convolution neural networks with affine moment invariants. IEEE Signal Processing Letters. 25(7): 1064-1068.

[54] Zhao Xiangyu, Peng Huang, and Xiangbo Shu. 2022. Wavelet-Attention CNN for image classification. Multimedia Systems. 28.3: 915-924.

[55] J. Large, J. Lines, and A. Bagnall. 2019. A probabilistic classifier ensemble weighting scheme based on cross-validated accuracy estimates. DataMin. Knowl. Discov. 33(6): 1674-1709.

[56] M. Rahimzadeh and A. Attar. 2020. A modified deep convolutional neural network for detecting COVID-19 and pneumonia from chest X-ray images based on the concatenation of Xception and ResNet50V2. Informatics Med. Unlocked. 19 100360.

[57] Boulkenafet Z., Komulainen J., Li L., Feng X., Hadid A. 2017. Oulu-npu: A mobile face presentation attack database with real-world variations. In: FGR. 612-618.

[58] Yu Zitong, *et al*. 2020. Face anti-spoofing with human material perception. Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part VII 16. Springer International Publishing.

[59] Z. Yu *et al*. 2020. Searching central difference convolutional networks for face anti-spoofing. in Proc. CVPR. pp. 5295-5305.

[60] A. George and S. Marcel. 2021. Cross modal focal loss for RGBD face anti spoofing. In Proc. CVPR. pp. 7882-7891.

[61] A. Parkin and O. Grinchuk. 2019. Recognizing multi-modal face spoofing with face recognition networks. In Proc. CVPR Workshops. pp. 1-8.

[62] Wang Zhuo, *et al*. 2022. Face anti-spoofing using transformers with relation-aware mechanism. IEEE Transactions on Biometrics, Behavior and Identity Science 4.3: 439-450.

[63] C. Finn, P. Abbeel, and S. Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In Proc. Int. Conf. Mach. Learn. pp. 1126-1135.

[64] J. Määttä, A. Hadid and M. Pietikainen. 2011. Face spoofing detection from single images using micro-texture analysis. In Proc. IEEE Int. Joint Conf. Biometrics. pp. 1-7.

[65] O. Lucena, A. Junior, V. Moia, R. Souza, E. Valle and R. Lotufo. 2017. Transfer learning using convolutional neural networks for face anti-spoofing. In ICIAR.

[66] Zezheng Wang *et al*. 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.