Received: 8 October 2023 | *Revised:* 20 December 2023 | *Accepted:* 21 December 2023
DOI: 10.37917/ijeee.22.1.5 *Vol. 22 | Issue 1 | June 2026*

Iraqi Journal for Electrical and Electronic Engineering
Original Article

Open Access
IJEEE University of Basrah College of Engineering

A Multi-Modal Convolutional Neural Network for Face Anti-Spoofing Detection

Hala S. Mahmood*[1], Salah Al-Darraji[2]
[1]Dept. of Computer Science, College of Education for Pure Sciences, University of Basrah, Basrah, Iraq
[2]Dept. of Computer Science, College of Computer Science & Information Technology, University of Basrah, Basrah, Iraq

Correspondance
*Hala S. Mahmood
Department of Computer Science, College of Education for Pure Sciences,
University of Basrah, Basrah, Iraq
Email: hala.shaker@uobasrah.edu.iq

Abstract
Recently, face recognition technology has become more prevalent in various applications, including mobile devices, access control, and financial transactions. Therefore, it is crucial to address potential vulnerabilities that attackers might exploit. In this study, a method for face presentation attack detection (PAD) is introduced. The method utilizes the diversity of modalities provided by some cameras and sensors to detect face spoofing using convolutional neural networks (CNN) within the context of deep learning. To assess the effectiveness of the proposed approach in real-world scenarios, the wide multi-channel presentation attack (WMCA) dataset is used. The presented method exploits the multi-modal data, including RGB, depth, IR, and thermal channels, to enhance system performance and explore different techniques for combining the results from each modality. Furthermore, this study explores diverse techniques for fusing results from each channel in two fusion scenarios, pre-fusion and post-fusion. In the pre-fusion scenario, data from the four channels is combined, resulting in an ACER value of 0.19%. In the post-fusion scenario, the results of each modality are fused using different fusion techniques, such as majority voting, weighted voting, average pooling, and a stacking classifier. The stacking classifier yields the most favorable outcome with an ACER ratio of 0.03%. This performance is notably superior when compared to state-of-the-art methodologies.

Keywords
Anti-spoofing, CNN, deep learning, face recognition, multi-modal, presentation attack detection (PAD), WMCA dataset

I. Introduction

In recent years, face recognition technology has witnessed remarkable advancements, achieved state-of-the-art performance and even surpassed human-level capabilities [1]. A significant contributing factor to this success is the availability of massive annotated face datasets, often collected from the Internet [2, 3].

However, while face recognition systems have made significant strides, they remain vulnerable to a critical security threats known such as face spoofing or anti-spoofing attacks and adversarial attack [4, 5], the presence of presentation attacks (PAs) These threats poses significant challenges to the implementation and security of biometric systems, particularly facial recognition systems these attacks and can compromise the integrity and reliability of biometric authentication, making it crucial to address these difficulties in order to enhance the overall security of such systems [4, 6].

Face anti-spoofing is the process of identifying and differentiating between genuine facial images and deceptive presentations, such as printed photographs, masks, or replay attacks, to enhance the security and reliability of face recognition systems. To address this challenge, researchers have worked on developing anti-spoofing methods that aim on detecting presentation attacks and verifying the liveness of the cap-

tured face. However, one of the key obstacles faced by the research community in advancing face anti-spoofing methods is the limited availability of diverse and comprehensive datasets for training and evaluation and creating reliable face anti-spoofing datasets requires meticulous efforts in manually collecting samples, leading to restrictions in the number of unique individuals and samples.

Furthermore, most existing face anti-spoofing databases solely comprise RGB photos, limiting the information available for analysis [7–9]. Previous attempts to create multi-modal datasets have also been constrained by the scarcity of subjects, potentially resulting in over-fitting issues during training [10, 11].

To address these challenges and push the boundaries of the face anti-spoofing field, an approach is introduced in this study, harnessing multi-modal data integration that includes RGB, depth, infrared (IR), and thermal modalities. The objective is to optimize the performance of a neural network-based face anti-spoofing model.

The recently released WMCA [12] dataset for face anti-spoofing represents a significant advancement in liveness identification, as it is notable for its large dataset size and the inclusion of modalities such as RGB, Depth, IR, and Thermal, as in Fig. 1.

The method endeavors to extract rich and diverse information from each modality to enhance the system's ability to discern between real and fake faces.

In this paper, a method is introduced to utilize CNN for addressing the face anti-spoofing problem. A combination of RGB, IR, depth, and thermal modalities will be employed to enhance the performance of a neural network model. Each modality will be analyzed individually to get separate results for each modality. These results will be compared with the outcomes of references [12, 13], revealing that the current findings exhibit superior performance. Subsequently, the results of all modalities will be combined using different techniques.

First, majority voting [14] will be employed, wherein the most frequent value from each modality is chosen as the fusion value. Then, weighted voting will be employed [15], where each modality is assigned, weights based on its expected importance, and then the fusion value is calculated using the specified weights. In addition, average / pooling will be employed [16], where the values from each modality are arithmetically averaged and used as the pooling value.

Finally, stacking classifier will be employed [17], where a new model will be trained utilizing the results from each modality as features to predict the final outcome. By using these different techniques to combine modality results, capitalizing on the diversity of available information enhances the performance of the resulting model in search.

The main contributions can be summarized as follows:

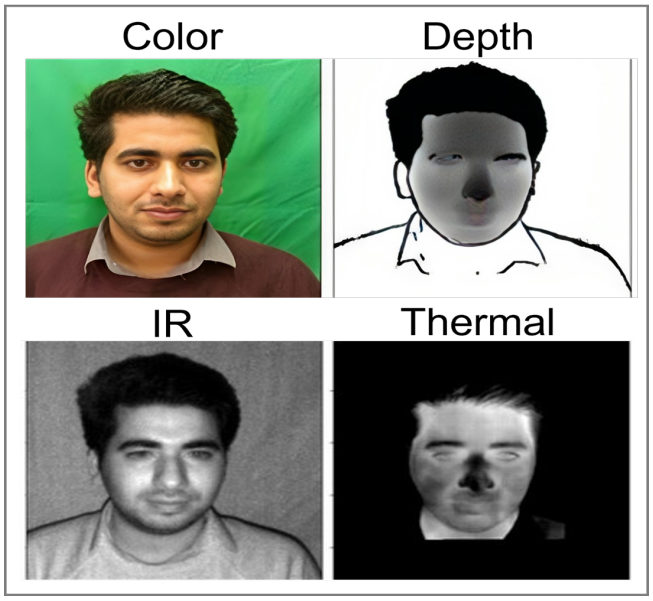

Fig. 1. The WMCA dataset sample of four different modalities [12]

- The paper introduces a multi-modal approach for face anti-spoofing, utilizing four different modalities: RGB, Depth, IR, and Thermal. by integrating information from multiple modalities, which enhances the ability to discriminate between genuine and spoofed faces.

- The paper explores different techniques for merging the results obtained from individual modalities. A new model is trained to utilize the results from each modality as features and predict the final anti-spoofing outcome. The stacking approach allows the model to learn from the combined information and make more informed decisions.

- The quality of the model was evaluated by employing cross-validation, which helps mitigate issues related to dataset partitioning and provides a more robust assessment.

This paper is organized as follows: Section II. provides a discussion of the most relevant literature review. In Section III. , the proposed method is introduced, covering the network architecture. Section IV. presents the experimental research, including comprehensive information about the dataset used, the applied protocol, and the specific working details, along with the methods used to evaluate the proposed PAD algorithm trained on the WMCA database. Section V. results and discussion the findings of this paper. Finally, Section VI. concludes the paper.

II. Related Work

For the development of non-cooperative liveness detection, there are a variety of datasets containing false and genuine photos that might be employed.

Replay-Attack [18], CASIA-FASD [19], and SiW [20] are well-known datasets for Presentation Attack Detection (PAD). These datasets are widely used in the field. It's important to note that these datasets solely consist of RGB data. Even the recently released SynthASpoof [21] a synthetic-based face PAD dataset, only contains RGB data.

Due to the ubiquitous use of face recognition in mobile devices, there are also several RGB datasets that were recorded by replaying face videos with smartphones, such as MSU-MFSD [9], Replay-Mobile [8], and OULU-NPU [7]. However, these datasets only include the RGB modality, which restricts the applicability and efficacy of related algorithms. Presentation attacks (PAs) in new forms, such as 3D attacks [22] and silicone masks [23], which are more complex and realistic than conventional 2D attacks, have evolved as a result of the ongoing advancement of attack methodologies. These improvements in attack techniques have shown the shortcomings of visible cameras in spotting such convincing face masks. Fortunately, face presentation attack detection (PAD) approaches now have more options than ever because of the development of new sensor technologies, including depth cameras, multi-spectral cameras, and infrared light cameras.

Kose et al. [24] introduced a dataset specifically designed to study the impacts of 2D and 3D face mask attacks. Modern CNNs have been effectively used in face anti-spoofing research, leveraging their strong representation power as demonstrated in [25–29]. Kim et al. [30] proposed a method that utilizes reflectance properties to differentiate between facial skins and mask materials. In [26, 29], pretrained CaffeNet or VGG-face models were employed as feature extractors to discern live and spoof faces.

The 3DMAD dataset [11], the first publicly accessible 3D mask dataset, was generated using the Microsoft Kinect sensor. This dataset comprises RGB and Depth modalities and consists of data from 17 subjects. Msspoof [10] is another multi-modal face PAD dataset that includes visible (VIS) and near-infrared (NIR) images. This dataset consists of real access as well as printed spoofing attacks and encompasses data from 21 subjects. CSMAD [31] is another multi-modal face PAD dataset that includes (VIS, Depth, NIR, and thermal) images. This dataset consists of data from 14 subjects. However, one of the limitations of existing datasets in face PAD is their restricted number of subjects and samples, which may lead to a potential over-fitting risk when evaluating face PAD algorithms on these datasets.

To enhance liveness detection, the newly published face anti-spoofing datasets such as WMCA [12] pushes the boundaries of the liveness identification challenge both in terms of the dataset size and the number of presented modalities (RGB, Depth, IR, and Thermal).

III. The Proposed Method

This section outlines the various stages of the proposed Presentation Attack Detection (PAD) framework.

A. Preprocessing

In the context of presentation attack detection, preprocessing plays a crucial role in improving the accuracy and robustness of the detection algorithms and enhancing the quality of the input image [12]. By applying preprocessing techniques such as face detection and landmark detection, the input images can be standardized and aligned, reducing variations caused by pose, illumination, and occlusions. A level of image preprocessing included face recognition and facial landmark detection. This step was utilized to improve the face's visual representation and image quality. Using the preprocessed images as training data, a connected sequential neural network was then created.

B. Network Architecture

In this research paper, a multi-modal CNN design is introduced for the purpose of detecting face spoofing. The proposed CNN architecture utilizes inputs such as RGB, depth, IR, and thermal images, all with dimensions of 128×128 pixels, sourced from the WMCA dataset [12]. The CNN consists of layers with filter sizes of 32, 64, and 128 followed by max pooling. Activation functions like ReLU and tanh are incorporated to introduce non-linearity in the model [32]. For the layers, the 'sigmoid' activation function is employed [33]. Train the model using the 'binary-crossentropy' loss function with the 'adamax' optimizer [34]. for the model, the evaluation metric is set to 'accuracy'.

The method of the proposed work consists of two scenarios:

1) Pre-Fusion

In the pre-fusion scenario, A series of experiments are performed to evaluate the effectiveness of the proposed system, Fig. 2. The system is tested using various combinations of multi-modal data, and all modalities are swapped with each other to ensure realistic and comprehensive results. This analysis aims to understand how the system performs with different input sets. The experiments demonstrate promising results when multiple modalities are combined together (four-modal: CDIT C=color, D=depth, I=IR, and T=thermal, as in Fig 1). The fusion approach enhances safety, robustness, and output reliability of the system.

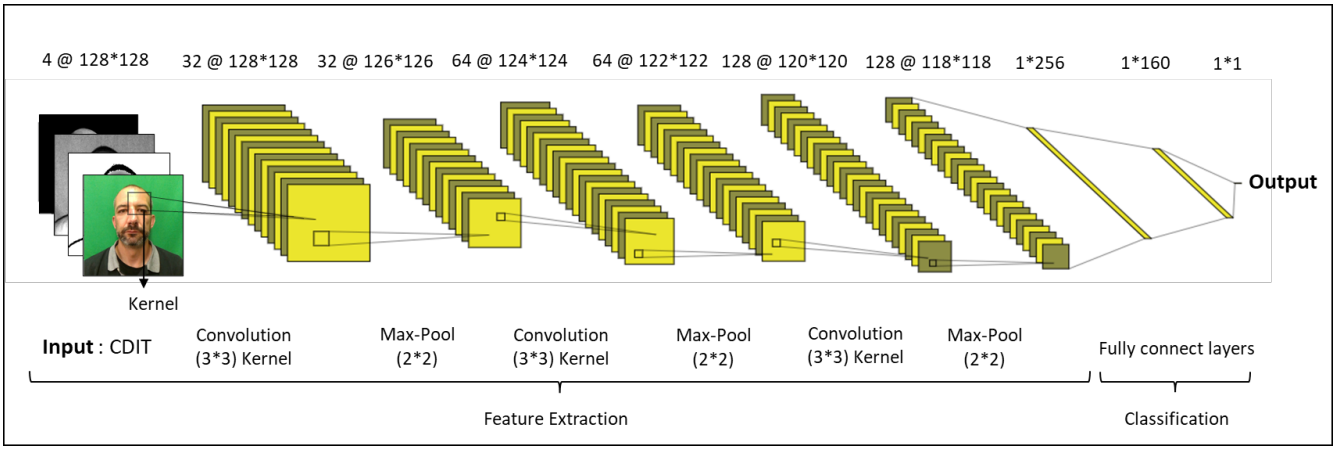

Fig. 2. Architecture for pre-fusion scenario

2) Post-Fusion

In the post-fusion scenario, individual modalities are processed independently. The anti-spoofing algorithm is applied to each modality, generating individual outputs, Fig. 3. These outputs are then combined using different fusion methods:

- Majority Voting [14]: Each classifier in the ensemble makes a prediction, and the final prediction is determined by selecting the class that receives the majority of votes.
- Weighted Voting [15]: Instead of equal weights for each classifier's prediction, a weighted scheme assigns different weights based on performance or confidence. Weights can be determined through cross-validation or model performance on a validation set.
- Averaging/Pooling [16]: The predicted probabilities or scores from multiple classifiers are averaged to obtain the final prediction. For multi-class classification, the predicted probabilities for each class from each classifier are averaged.
- Stacking/Stacked Generalization [17]: A meta-classifier is trained to take predictions from individual classifiers as input features and learn to make the final prediction. The individual classifier predictions serve as additional features for the meta-classifier.

The performance of the fusion methods is assessed to select the most effective approach for the final system output.

The proposed CNN architecture, along with the pre-fusion and post-fusion scenarios, demonstrates promising results for face anti-spoofing. The combination of multi-modal data fusion and diverse fusion methods enhances the system's security, robustness, and anti-spoofing accuracy. The findings contribute to advancing face anti-spoofing methods and emphasize the significance of boosting multiple modalities and fusion techniques for reliable facial recognition systems.

IV. Experimental Results

In this section, detailed explanations are provided regarding the utilized database and the experimental results achieved through the proposed approach.

The following sub-sections detail the different stages of the proposed framework.

A. Dataset

The Wide Multi-Channel Presentation Attack (WMCA) database was employed to assess the methodology [12]. WMCA has 1679 videos, 347 of which are bonafide presentations and 1332 of which are attacks as shown in Fig. 4.

The WMCA database comprises data recorded in four different modalities: RGB color data, depth maps, infrared data, and thermal data. The first three modalities were captured using an Intel RealSense SR300 camera, while thermal data was obtained using a Seek Thermal Compact PRO camera. The images are all 128×128 pixels in dimension and are geometrically aligned. In Fig. 2, samples of four image modalities are provided. Each image in the collection is shown in four modalities: color image intensity (C), depth (D), infrared image (I), and thermal data (T). In a previous study [12], the authors employed approximately 50 frames per video to perform data augmentation, resulting in a total of 83950 photos.

B. Protocol

The suggested model was trained using the WMCA database, which included 1679 photos. The images were split into a training set and a test set using a ratio of 70:30, respectively. The training set receives 70% of the data, whereas the test

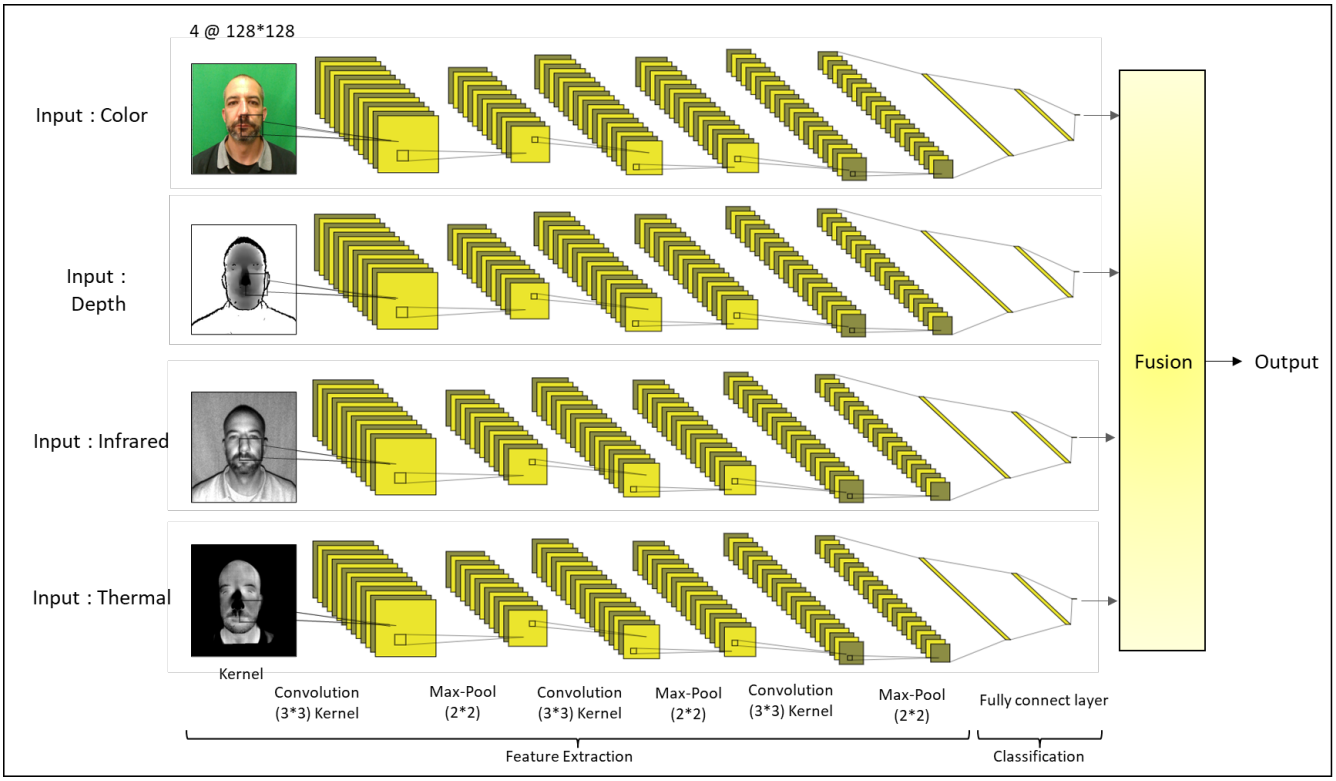

Fig. 3. Architecture for post-fusion scenario

set receives 30%. This ratio was established to guarantee that there would be an adequate amount of data for training and validation of the face discrimination model's overall performance.

From each video, only 50 uniformly sampled frames in the temporal domain were chosen. A video's individual frames are regarded as separate samples that provide one score per frame. A biometric sample comprises geographyically and temporally aligned frames from each of the four modalities [12].

Cross-validation [35] was employed to evaluate the performance of the proposed multi-modal fusion method. Specifically, the evaluation process involved the utilization of the k-fold cross-validation technique [36].

Cross-validation is a commonly used approach in machine learning to estimate how well a model will perform on unseen data and assess its generalizability. In k-fold cross-validation, the dataset is divided into k subsets, or folds. During each iteration, the model is trained on k-1 folds and then evaluated on the remaining fold, which serves as the validation set. This process is repeated k times, ensuring that each fold is used as the validation set exactly once.

C. *Implementation Details*

For the research, the authors of the WMCA dataset [12] added 83950 more images to the collection by including around 50 frames per video. The database is split into training and testing sets of 70:30. Throughout the model-building process, the activation functions tanh and ReLU were utilized [32]. Additionally, the sigmoid activation function was incorporated in the neural network's output layer to yield categorical outputs ranging from 0 to 1 [33] For training purposes, the batch size and number of epochs are assumed to be 100 and 30, respectively. subsequently trained the proposed network with a batch size of 100 using the AdaMax optimizer [34].

The employed loss function was binary crossentropy. The implementation made use of the library Tensorflow [37].

D. *Metrics*

The effectiveness of the proposed PAD algorithm was assessed using metrics specified by the ISO/IEC 30107-3 standards [38]:

- Attack Presentation Classification Error Rate (APCER): It is described as the rate of misclassified spoof images (spoof expected to be live).

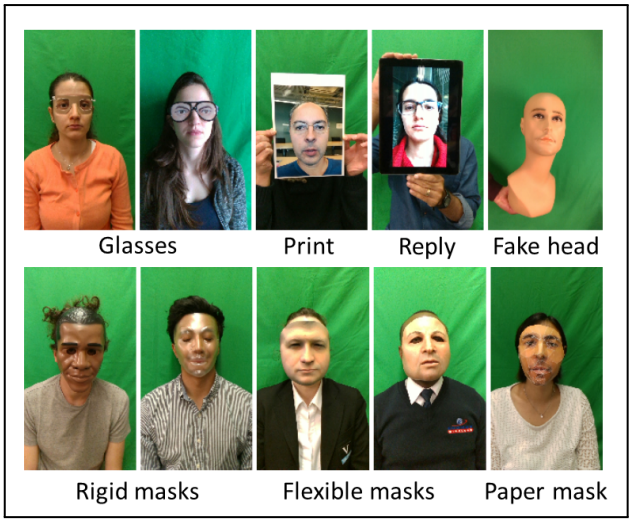

Fig. 4. Presentation attacks with different PAs [12]

$$APCER(\%) = \frac{\#\ of\ accepted\ attacks}{\#\ of\ attacks} \quad (1)$$

- Bonafide Presentation Classification Error Rate (BPCER): It is described as the rate of misclassified live images (live predicted as spoof).

$$BPCER(\%) = \frac{\#\ of\ rejected\ real\ attempts}{\#\ of\ real\ attempts} \quad (2)$$

- Average Classification Error Rate (ACER): It is the average of APCER and BPCER and is used to summarize the results as a single number. Both validation and test sets are reported with the ACER.

$$ACER(\%) = \frac{APCER(\%) + BPCER(\%)}{2} \quad (3)$$

The threshold for a BPCER value of 1% is determined in the validation set. Additionally, ROC curves for both the baseline and multi-modal methods are presented, along with the error metrics.

V. Results and Discussion

The primary objective of this paper is to tackle the issue of face impersonation attacks. To address this, a new multi-modal approach is proposed to counteract face impersonation.

The system employs four distinct modalities of information RGB, Depth, IR, and Thermal. The combination of these modalities allows the neural network model to benefit from a wide range of data sources, enhancing the overall performance of the anti-spoofing system compared to using individual modalities independently.

To achieve effective data fusion, two fusion scenarios were explored: pre-fusion and post-fusion.

In the pre-fusion scenario, A comprehensive set of experiments was conducted to assess the effectiveness of the system. Table I presents the results of these experiments, comparing different permutations of the four modalities (CDIT) within the system. This analysis aims to understand how the system performs when the inputs are different sets of data. The results show that when the inputs are multiple modalities together (i.e., the four-modal CDIT), it leads to promising results in terms of safety, robustness, and reliability of the outputs. The value of ACER is 0.19% for the CDIT compared to [12] where the value of ACER is 2.91%.

Moreover, the study reveals an important observation: when one modality is used separately, for example, modality C, the value of ACER is 24.01, which is a very large value, i.e., the system becomes vulnerable to the simplest attacks such as print or replay attacks. In such cases, adversaries can mimic the input data from that specific modality and deceive the system. This finding emphasizes the critical role of multi-modal data fusion to bolster the system's security and defense against spoofing attempts.

In the post-fusion scenario, the modalities are processed individually, and their respective outputs are fusion to obtain the final system output. This process involves taking each modalities data and applying the anti-spoofing algorithm independently. Once the algorithm processes all the modalities, their outputs are combined using various fusion methods. In this study, four types of fusions were used: majority voting, weighted voting, average pooling, and stacking classifiers.

The performance of the post-fusion scenario was evaluated based on the ACER. The ACER ratio for the stacking classifier is 0.03%, which is the best result, as indicated in Table II.

Interestingly, the ACER ratio achieved with post-fusion was even lower than the ACER ratio obtained from using CDIT fusion in the pre-fusion scenario. This comparison suggests that the post-fusion stage outperforms the pre-fusion scenario in terms of anti-spoofing accuracy.

Consequently, the conclusion derived from the analysis indicates that the post-fusion technique emerges as the most effective approach among the fusion methods tested. By fusing the outputs of each modality after individual processing, the system can leverage the diverse information captured by the different modalities and achieve improved face anti-spoofing performance. The combination of the four types of fusion methods further contributes to the overall success of the post-fusion stage.

TABLE I.
TABLE OF RESULTS AND COMPARISONS BETWEEN MODALITIES IN TEST SET

Modality	Test set								
	MC-CNN [12]			Basic feature set + SVM A total of 1679 images were used [13]			Proposed		
	APCER (%)	**BPCER (%)**	**ACER (%)**	**APCER (%)**	**BPCER (%)**	**ACER (%)**	**APCER (%)**	**BPCER (%)**	**ACER (%)**
CDIT	0.6	0	0.3	4.07	1.74	2.91	0	0.39	**0.19**
DIT	—	—	—	—	—	—	0	2.55	**1.27**
CDT	—	—	—	7.01	6.96	6.99	1.64	4.94	**3.29**
CIT	—	—	—	5.20	1.74	3.47	1.07	3.0	**2.04**
CDI	2.07	0	1.04	9.28	0	4.64	0.97	0.87	**0.92**
CD	—	—	—	12.90	6.09	9.49	2.79	11.84	**7.32**
CI	—	—	—	6.11	8.70	7.40	7.85	0	**3.92**
CT	—	—	—	10.86	1.74	6.30	3.72	4.65	**4.18**
DI	—	—	—	—	—	—	0.91	0.52	**0.71**
DT	—	—	—	—	—	—	0	6.35	**3.18**
IT	—	—	—	—	—	—	1.50	9.19	**5.35**
C	65.65	0	32.82	—	—	—	34.96	13.06	**24.01**
D	11.77	0.31	6.04	—	—	—	5.49	4.03	**4.76**
I	5.03	0	2.51	—	—	—	0	0.77	**0.39**
T	3.14	0.56	**1.85**	—	—	—	1.33	7.19	4.26

The study involved a comparison of the results of two algorithms: RDWT-Haralick-SVM, as proposed by reference [39], which relies on the SVM classifier, and the MC-CNN algorithm, which employs an artificial neural network for feature extraction and classify them. This methodology is based on the Light CNN deep learning model initially introduced in [40]. Both of these algorithms demonstrated outstanding performance in the context of face spoofing detection, as evidenced by their results in reference [12]. In that study, the CDIT metric yielded an ACER of 3.44% when utilizing the RDWT-Haralick-SVM algorithm and an ACER of 0.3% when employing MC-CNN. Furthermore, as stated in [13], the outcomes were similarly remarkable, as an ACER of 2.91% was obtained while using a collection of 1679 images, and an ACER of 1.18% was achieved when employing a dataset consisting of 83,950 images. Nevertheless, it's noteworthy that the proposed method surpassed these results in terms of performance as shown in the Table III.

In order to assess the performance of the proposed multi-modal fusion method, a k-fold cross-validation technique with k = 5 was employed. This involved dividing the dataset into five subsets. During each iteration, the model was trained on four subsets, while one subset was set aside for validation. This process was repeated five times, ensuring that every subset served as the validation set exactly once.

The evaluation metric used to gauge the model's performance was accuracy. After completing the cross-validation process, the model's accuracy was determined to be 99%. This high accuracy demonstrates the robustness and efficacy of the proposed fusion method.

TABLE II.
TABLE OF RESULTS POST-FUSION

Fusion Type	**APCER (%)**	**BPCER (%)**	**ACER (%)**
Majority voting	0.26	1.79	1.02
Weighted voting	0.03	0.23	0.13
Average / pooling	0.03	0.23	0.13
Stacking classifier	0.07	0	**0.03**

The fact that the proposed fusion method achieved such high accuracy on unseen data indicates its potential practical utility for face anti-spoofing applications. By successfully making accurate predictions on new and unseen samples, the model proves its ability to effectively detect and prevent face spoofing attacks, showcasing its reliability and real-world applicability.

The proposed multi-modal fusion method was evaluated using the Area Under the Curve (AUC) metric and the Receiver Operating Characteristic (ROC) curve. The AUC value was found to be 1 is shown in Fig. 5 indicates that the

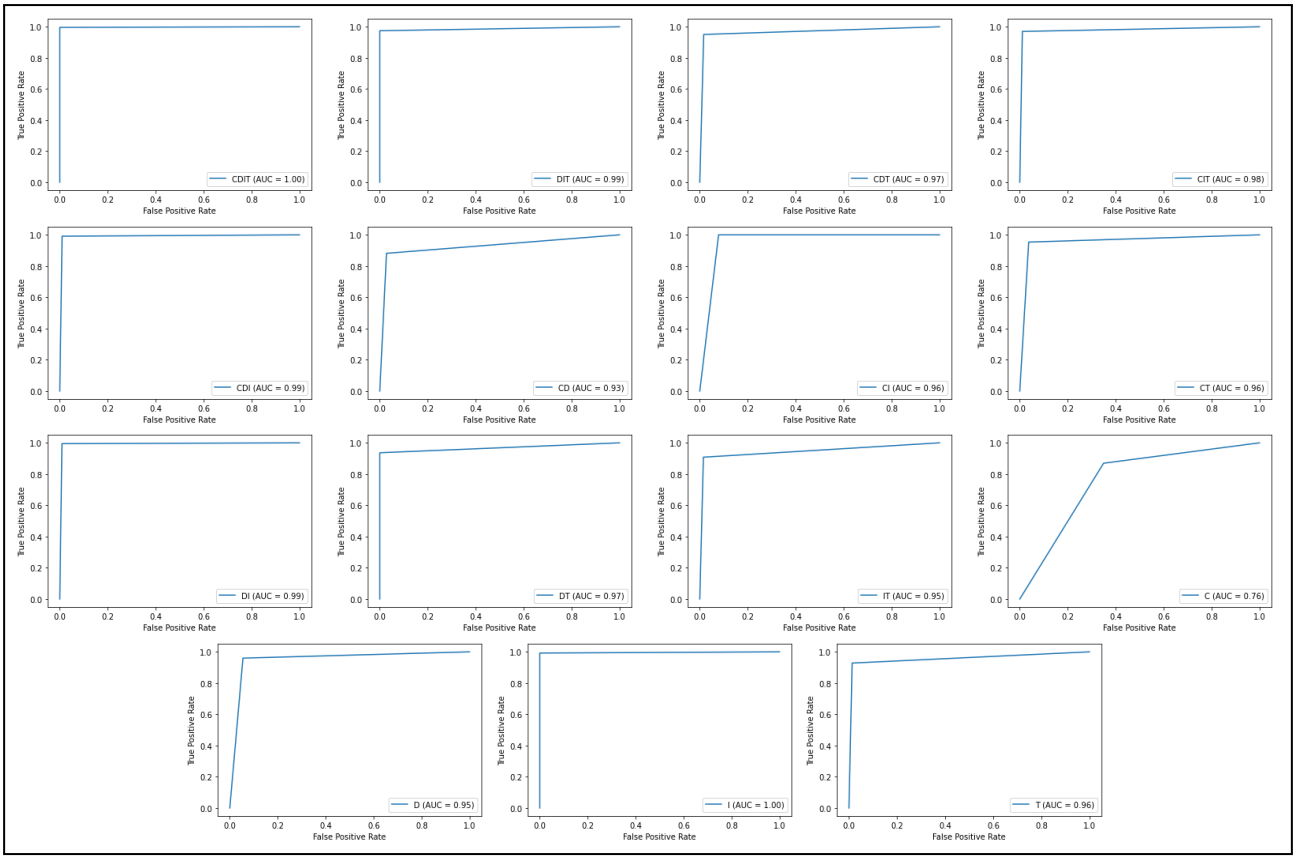

Fig. 5. ROC curve of pre-fusion scenario

TABLE III.
THE TEST SET IN THE CASE OF CDIT DATA AND FULL DATASET OF 83950 IMAGES

Method	APCER (%)	BPCER (%)	ACER (%)
RDWT-Haralick+ SVM [12]	6.39	0.49	3.44
MC-CNN [12]	0.60	0	0.30
Basic feature set + SVM [13]	0.11	2.24	1.18
Basic feature set + RF [13]	3.23	6.50	4.87
Pre-fusion (Proposed)	0	0.39	0.19
Post-fusion (Proposed)	0.07	0	**0.03**

model exhibits excellent discriminative power, effectively distinguishing between genuine and spoofed faces. The high AUC value of 1 indicates that the proposed multi-modal fusion method achieved near-perfect performance in distinguishing between real and fake facial images. Typically, a well-performing model's ROC curve shows a steep ascent towards the top-left corner, indicating a high true positive rate and a low false positive rate, further supporting the robustness and accuracy of the proposed fusion method.

VI. Conclusion

In this paper, a face presentation attack detection PAD method is proposed in this study, employing multi-modal images and a CNN-based anomaly detection. The multi-modal images offer rich information to differentiate between different attack modes, and the anomaly detection technique ensures generalization performance. The effectiveness of the approach is evaluated using the wide multi-modal presentation attack WMCA dataset, showcasing its applicability within real-world scenarios. Introducing a new face anti-spoofing method that effectively utilizes multi-modal data, including RGB, IR, depth, and thermal modalities, to enhance system performance and exploration of different techniques for combining the results from each modality, including majority voting, weighted voting, average/ pooling, and a stacking classifier, to capitalize on the diverse information available. Extensive evaluation of

the proposed approach using the newly published face anti-spoofing dataset WMCA, which includes multiple modalities and significantly expands the dataset size and subject diversity.

For future work, it is possible to use more than one dataset and apply the proposed algorithm to them, further exploring the potential of multi-modal face anti-spoofing methods.

Conflict of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared influence the work reported in this paper.

References

[1] P. J. Phillips, A. N. Yates, Y. Hu, C. A. Hahn, E. Noyes, K. Jackson, J. G. Cavazos, G. Jeckeln, R. Ranjan, S. Sankaranarayanan, *et al.*, "Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms," *Proceedings of the National Academy of Sciences*, vol. 115, no. 24, pp. 6171–6176, 2018.

[2] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "Ms-celeb-1m: A dataset and benchmark for large-scale face recognition," in *European conference on computer vision*, pp. 87–102, Springer, 2016.

[3] O. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *BMVC 2015-Proceedings of the British Machine Vision Conference 2015*, British Machine Vision Association, 2015.

[4] A. Hadid, "Face biometrics under spoofing attacks: Vulnerabilities, countermeasures, open issues, and research directions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 113–118, 2014.

[5] A. Kadhim and S. Al-Darraji, "Face recognition system against adversarial attack using convolutional neural network.," *Iraqi Journal for Electrical & Electronic Engineering*, vol. 18, no. 1, 2022.

[6] U. Scherhag, R. Raghavendra, K. B. Raja, M. Gomez-Barrero, C. Rathgeb, and C. Busch, "On the vulnerability of face recognition systems towards morphed face attacks," in *2017 5th international workshop on biometrics and forensics (IWBF)*, pp. 1–6, IEEE, 2017.

[7] Z. Boulkenafet, J. Komulainen, L. Li, X. Feng, and A. Hadid, "Oulu-npu: A mobile face presentation attack database with real-world variations," in *2017 12th IEEE international conference on automatic face & gesture recognition (FG 2017)*, pp. 612–618, IEEE, 2017.

[8] A. Costa-Pazo, S. Bhattacharjee, E. Vazquez-Fernandez, and S. Marcel, "The replay-mobile face presentation-attack database," in *2016 international conference of the Biometrics Special Interest Group (BIOSIG)*, pp. 1–7, IEEE, 2016.

[9] D. Wen, H. Han, and A. K. Jain, "Face spoof detection with image distortion analysis," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 4, pp. 746–761, 2015.

[10] I. Chingovska, N. Erdogmus, A. Anjos, and S. Marcel, "Face recognition systems under spoofing attacks," in *Face Recognition Across the Imaging Spectrum*, pp. 165–194, Springer, 2016.

[11] N. Erdogmus and S. Marcel, "Spoofing in 2d face recognition with 3d masks and anti-spoofing with kinect," in *2013 IEEE sixth international conference on biometrics: theory, applications and systems (BTAS)*, pp. 1–6, IEEE, 2013.

[12] A. George, Z. Mostaani, D. Geissenbuhler, O. Nikisins, A. Anjos, and S. Marcel, "Biometric face presentation attack detection with multi-channel convolutional neural network," *IEEE transactions on information forensics and security*, vol. 15, pp. 42–55, 2019.

[13] A. Denisova, "An improved simple feature set for face presentation attack detection," 2022.

[14] Y. Zheng and E. A. Essock, "A local-coloring method for night-vision colorization utilizing image analysis and fusion," *Information Fusion*, vol. 9, no. 2, pp. 186–199, 2008.

[15] J. Kittler, M. Hatef, R. P. Duin, and J. Matas, "On combining classifiers," *IEEE transactions on pattern analysis and machine intelligence*, vol. 20, no. 3, pp. 226–239, 2002.

[16] L. I. Kuncheva, *Combining pattern classifiers: methods and algorithms*. John Wiley & Sons, 2014.

[17] D. H. Wolpert, "Stacked generalization," *Neural networks*, vol. 5, no. 2, pp. 241–259, 1992.

[18] I. Chingovska, A. Anjos, and S. Marcel, "On the effectiveness of local binary patterns in face anti-spoofing," in *2012 BIOSIG-proceedings of the international conference of biometrics special interest group (BIOSIG)*, pp. 1–7, IEEE, 2012.

[19] Z. Zhang, J. Yan, S. Liu, Z. Lei, D. Yi, and S. Z. Li, "A face antispoofing database with diverse attacks," in *2012*

5th IAPR international conference on Biometrics (ICB), pp. 26–31, IEEE, 2012.

[20] Y. Liu, A. Jourabloo, and X. Liu, "Learning deep models for face anti-spoofing: Binary or auxiliary supervision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 389–398, 2018.

[21] M. Fang, M. Huber, and N. Damer, "Synthaspoof: Developing face presentation attack detection based on privacy-friendly synthetic data," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1061–1070, 2023.

[22] N. Erdogmus and S. Marcel, "Spoofing face recognition with 3d masks," *IEEE transactions on information forensics and security*, vol. 9, no. 7, pp. 1084–1097, 2014.

[23] H. Steiner, A. Kolb, and N. Jung, "Reliable face anti-spoofing using multispectral swir imaging," in *2016 international conference on biometrics (ICB)*, pp. 1–8, IEEE, 2016.

[24] N. Kose and J.-L. Dugelay, "Countermeasure for the protection of face recognition systems against mask attacks," in *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pp. 1–6, IEEE, 2013.

[25] L. Feng, L.-M. Po, Y. Li, X. Xu, F. Yuan, T. C.-H. Cheung, and K.-W. Cheung, "Integration of image quality and motion cues for face anti-spoofing: A neural network approach," *Journal of Visual Communication and Image Representation*, vol. 38, pp. 451–460, 2016.

[26] K. Patel, H. Han, and A. K. Jain, "Cross-database face antispoofing with robust feature representation," in *Chinese Conference on Biometric Recognition*, pp. 611–619, Springer, 2016.

[27] J. Yang, Z. Lei, and S. Z. Li, "Learn convolutional neural network for face anti-spoofing," *arXiv preprint arXiv:1408.5601*, 2014.

[28] W. Luo, P. Sun, F. Zhong, W. Liu, T. Zhang, and Y. Wang, "End-to-end active object tracking and its real-world deployment via reinforcement learning," *IEEE transactions on pattern analysis and machine intelligence*, vol. 42, no. 6, pp. 1317–1332, 2019.

[29] L. Li, X. Feng, Z. Boulkenafet, Z. Xia, M. Li, and A. Hadid, "An original face anti-spoofing approach using partial convolutional neural network," in *2016 Sixth international conference on image processing theory, tools and applications (IPTA)*, pp. 1–6, IEEE, 2016.

[30] Y. Kim, J. Na, S. Yoon, and J. Yi, "Masked fake face detection using radiance measurements," *Journal of the Optical Society of America A*, vol. 26, no. 4, pp. 760–766, 2009.

[31] S. Bhattacharjee, A. Mohammadi, and S. Marcel, "Spoofing deep face recognition with custom silicone masks," in *2018 IEEE 9th international conference on biometrics theory, applications and systems (BTAS)*, pp. 1–7, IEEE, 2018.

[32] Y. Li and Y. Yuan, "Convergence analysis of two-layer neural networks with relu activation," *Advances in neural information processing systems*, vol. 30, 2017.

[33] A. C. Marreiros, J. Daunizeau, S. J. Kiebel, and K. J. Friston, "Population dynamics: variance and the sigmoid activation function," *Neuroimage*, vol. 42, no. 1, pp. 147–157, 2008.

[34] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[35] M. W. Browne, "Cross-validation methods," *Journal of mathematical psychology*, vol. 44, no. 1, pp. 108–132, 2000.

[36] J. D. Rodriguez, A. Perez, and J. A. Lozano, "Sensitivity analysis of k-fold cross validation in prediction error estimation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, no. 3, pp. 569–575, 2009.

[37] J. Brownlee and M. L. Mastery, *Deep Learning with Python: Develop Deep Learning Models on Theano and TensorFlow Using Keras*. Machine Learning Mastery, 2017.

[38] I. Standard, "Information technology–biometric presentation attack detection–part 3: testing and reporting," *International Organization for Standardization: Geneva, Switzerland*, vol. 7, 2017.

[39] K. E. Ewald, L. Zeng, C. B. Mawuli, H. S. Abubakar, A. Victor, *et al.*, "Applying cnn with extracted facial patches using 3 modalities to detect 3d face spoof," in *2020 17th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)*, pp. 216–221, IEEE, 2020.

[40] X. Wu, R. He, Z. Sun, and T. Tan, "A light cnn for deep face representation with noisy labels," *IEEE transactions on information forensics and security*, vol. 13, no. 11, pp. 2884–2896, 2018.