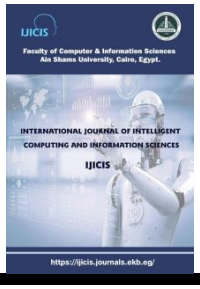
IJICIS, Vol.23, No.4, 96-111
DOI: 10.21608/ijicis.2023.254505.1309

International Journal of Intelligent Computing and Information Sciences

https://ijicis.journals.ekb.eg/



EVALUATING RESTRICTIVE VOTING AND MOVINET-BASED TRANSFER LEARNING FOR ROBUST FACE LIVENESS DETECTION

Mahmoud Omara*

Computer Systems Department,
Faculty of Computer and information sciences, Ain Shams University,
Cairo, Egypt
mahmoud.omara@cis.asu.edu.eg

Mahmoud Fayez

Computer Systems Department,
Faculty of Computer and information sciences, Ain Shams University,
Cairo, Egypt
mahmoud.fayez@cis.asu.edu.eg

H. Khalid

Computer Systems Department,
Faculty of Computer and information sciences, Ain Shams University,
Cairo, Egypt
heba.khaled@cis.asu.edu.eg

Said Ghoniemy

Computer Systems Department,
Faculty of Computer and information sciences, Ain Shams University,
Cairo, Egypt
ghoniemy1@cis.asu.edu.eg

Received 2023-12-10; Revised 2023-12-26; Accepted 2023-12-28

Abstract: *Face recognition technologies are rapidly becoming integral to a variety of applications, ranging from security systems to user authentication. This increasing reliance necessitates robust and accurate methods for face liveness detection. This paper presents and thoroughly evaluates two cutting-edge methods for face liveness detection: the Restrictive Voting approach and the Transfer Learning Approach. Our evaluation was performed using the Replay-Attack dataset. Various performance metrics were reported, including Accuracy, Precision, Recall, and F1-Score. Additionally, a comparative analysis was presented specifically for Half Total Error (HTER) and Equal Error Rate (EER), clearly indicating the superior performance of both methods compared to current state-of-the-art techniques. Remarkably, both methods achieved a zero False Acceptance Rate (FAR), thereby entirely negating the possibility of unauthorized access. These groundbreaking findings not only affirm the robustness of the introduced methods but also suggest their substantial potential for implementation in high-security, real-world scenarios, highlighting their unmatched excellence in face liveness detection.*

Keywords: *machine-learning, face anti-spoofing, SVM, transfer-learning, deep-learning.*

*__Corresponding Author__: Mahmoud Omara

Computer Systems Department, Faculty of Computer and Information Science, Ain Shams University, Cairo, Egypt

Email address: mahmoud.omara@cis.asu.edu.eg

1. Introduction

Face recognition has emerged as a cornerstone technology in biometric authentication, offering ease-of-use and non-intrusiveness across a wide array of applications from personal mobile devices to sophisticated security systems in commercial and governmental settings. As the adoption rate of this technology soars, so does its attractiveness as a target for malicious spoofing attempts. A person with ill intentions can exploit vulnerabilities in the system by presenting photos, videos, or even 3D masks to the face recognition sensors, thereby gaining unauthorized access to secured areas or confidential information.

The implications of such unauthorized access are far-reaching and could have disastrous consequences. For instance, it could lead to identity theft, fraudulent financial activities, or unauthorized alteration of sensitive data. Therefore, effective countermeasures, particularly robust face liveness detection algorithms, have become crucial for enhancing the security of such systems.

In this rapidly evolving landscape, one of the key challenges is to develop a face liveness detection algorithm that not only minimizes the number of false rejections but, most critically, eliminates false acceptances. Achieving a False Acceptance Rate (FAR) of zero is of paramount importance because it ensures that fraudulent attempts to spoof the system will not be successful. In fact, the criticality of achieving a zero FAR cannot be overstated for systems that guard against unauthorized access to sensitive or classified information.

With this imperative, we present two innovative methods focused on achieving a zero FAR. The first, coined as "Restrictive Voting" employs an ensemble of classifiers and incorporates a weighted voting system. Impressively, it achieved an Equal Error Rate (EER) of 2.5%, a Half Total Error (HTER) of 2.75%, and a zero FAR on the Replay-Attack dataset. The second method, known as the "Transfer Learning Approach for Face Liveness Detection", exploits a pre-trained MoviNet model [1] tailored for our use-case. This method also attained a zero FAR and has exceedingly low EER and HTER values of 5.0%.

Both methods demonstrate not only robust performance in discriminating genuine from spoofed faces but also offer the crucial benefit of ensuring maximum resistance to unauthorized access exemplified by their achievement of a zero FAR.

2. Related Work

Face recognition technology has gained substantial attention over the past few decades, leading to a plethora of research focusing on various aspects such as recognition accuracy, speed, and security. In this regard, face liveness detection emerges as a critical subdomain, aimed at thwarting spoofing attacks that can compromise the integrity of face recognition systems.

2.1. Traditional Approaches

Traditional methods for face liveness detection have played a seminal role in shaping the field, offering foundational techniques that have been broadly implemented. These conventional approaches typically

rely on handcrafted features and specifically engineered algorithms [2]–[6] to distinguish genuine faces from spoofed versions. Three primary clusters of traditional methodologies are as follows:

2.1.1. Texture Analysis

Texture analysis represents one of the earliest and most common methodologies. Methods such as Local Binary Patterns (LBP), Histogram of Oriented Gradients (HOG), and Scale-Invariant Feature Transform (SIFT) have been employed to extract unique facial texture features. These extracted features are subsequently classified using machine learning algorithms like Support Vector Machines (SVM) or Random Forests. While successful under controlled conditions, texture analysis approaches [7] are often vulnerable to environmental variabilities like lighting conditions and pose. Recent advancements have aimed to overcome these limitations by fusing traditional texture features with deep learning-based features to enhance robustness against such variabilities [8].

2.1.2. Motion Analysis

Another cornerstone in traditional liveness detection is motion analysis, which focuses on the dynamic characteristics of facial movement. Techniques include blink pattern analysis and micro-expression recognition. Notable works in this area include A. Fogelton et al.'s study on eye-blink detection [9] and Pfister et al.'s research on facial micro-expression analysis [10]. Although effective against basic spoofing attacks, these methods face challenges in detecting more sophisticated attempts, such as high-quality deepfake videos. In response to these challenges, newer approaches have been developed, such as those exploring motion blur as an indicator of replayed video attacks [11], and the use of motion and similarity feature analysis to improve the robustness of anti-spoofing systems under various spoof attacks [12].

2.1.3. Depth-based Methods

Depth-based approaches leverage 3D sensors like depth cameras or structured light scanners to capture the geometric properties of a face. This includes depth distribution and surface curvature, which are resistant to 2D image-based attacks. Noteworthy algorithms in this sector include Maatta et al.'s depth-based approach [13] and subsequent work on utilizing facial landmarks in depth recognition [14]. Additionally, innovative methods such as recovering 3D facial structures from a single camera have also been explored, contributing significantly to the versatility of depth-based methods [15] However, the requirement for specialized hardware may limit the widespread adoption of these techniques.

2.2. Deep Learning-Based Approaches

In the ever-evolving landscape of computer vision, deep learning technologies have emerged as instrumental disruptors, particularly in the realm of face liveness detection.

2.2.1. Convolutional Neural Networks (CNNs)

Among the deep learning architectures, Convolutional Neural Networks (CNNs) have gained considerable traction in face liveness detection. Specialized in image and video analysis, CNNs employ a combination of convolutional and pooling layers to capture local and hierarchical features. Leading

architectures like VGGNet[16], ResNet[17], and Inception[18] have set benchmarks in the field, attesting to the efficacy of CNNs in categorizing facial authenticity.

2.2.2. Recurrent Neural Networks (RNNs) and Variants

Recurrent Neural Networks (RNNs) [19] and their derivatives address the temporal facets of video-based face liveness detection. Equipped with memory cells, these networks can internalize past information, thereby enriching their contextual understanding. Variants such as Long Short-Term Memory (LSTM) [20] and Gated Recurrent Units (GRUs) [21] are specifically engineered to mitigate the vanishing gradient problem, thus enabling the modeling of long-term dependencies. These specialized architectures enhance the system's capability to accurately flag spoofing attempts based on temporal irregularities.

In the frontiers of deep learning-based face liveness detection, several innovative approaches have been proposed. For instance, a novel deep architecture combining image diffusion with CNN and LSTM effectively addresses both static and dynamic aspects of liveness detection, showing significant improvements in accuracy and error rates [22]. Hybrid architectures that meld CNNs and RNNs aim to exploit both spatial and temporal dimensions for improved performance[23]. Additionally, recent advancements focus on enhancing temporal consistency, further refining the robustness of online liveness detection systems [24]. Attention mechanisms have also been incorporated to refine the focus on critical facial regions and dynamic behaviors.

2.3. Pretrained Models and Transfer Learning Approaches

Transfer learning has emerged as a powerful approach in face liveness detection, particularly beneficial when dataset sizes are limited. Pre-trained models like VGG-Face [25] have been explored for this task[26]. Fine-tuning these pre-trained models specifically for liveness detection has been shown to be highly effective, especially when supported by diverse datasets [27]–[29]. While existing approaches have advanced the field significantly, they still face challenges in achieving a zero FAR, indicating room for further research.

While these methods have advanced the field, many still do not adequately address the pressing issue of achieving a zero FAR. A system vulnerable to even a minuscule number of false acceptances can pose significant security risks, especially in high-stakes environments. Recognizing this critical gap, our work introduces two novel methods aimed explicitly at minimizing FAR to the point of non-existence.

The first method, "Restrictive Voting", employs an ensemble approach with weighted voting, enhancing decision accuracy beyond traditional techniques. The second, "Transfer Learning Method", leverages the pre-trained MoviNet model [1] for efficient and highly accurate face liveness detection, harnessing advanced deep learning capabilities not fully exploited in prior studies.

Both methods not only push the boundaries of accuracy in differentiating genuine faces from spoofed ones but also meet the vital criterion of achieving a zero FAR. This is not a mere incremental advance but a significant leap forward, addressing a critical and previously unmet need in the literature. By achieving zero false acceptances, our methods substantially elevate the robustness and reliability of face liveness detection systems, setting a new standard for security measures in the field."

3. Methodology

The objective of this research is to enhance the security of face recognition systems through robust face liveness detection techniques. We employed two distinct methods to achieve this objective: a traditional machine learning-based approach and a transfer learning methodology.

3.1. Dataset

The Replay-Attack[30] Database is utilized to train, evaluate, and test both approaches. This database, designed for assessing face spoofing, contains 1,300 video clips that feature attempts to deceive facial recognition systems. These clips showcase 50 unique individuals and are recorded under varying lighting environments. Each clip is either an authentic access attempt by a person using a laptop's integrated webcam or a spoofing attempt using a photograph or video of that person, lasting a minimum of 9 seconds.

3.2. Restrictive Voting Mechanism with Traditional Machine-Learning Classifiers

3.2.1. Data Preprocessing

The employed dataset comprises a multitude of videos of variable durations. To facilitate effective analysis, a multi-step preprocessing pipeline is enacted, as outlined in Figure. 1.

Initially, a frame sampling technique is applied to the videos, segmenting them into two-second sub-clips with a one-second overlap between consecutive sub-clips. This segmentation serves dual purposes: it not only reduces the computational burden by limiting the data that needs immediate processing but also retains crucial temporal information, treating each sub-clip as an independent unit for subsequent analysis.

Following segmentation, each sub-clip is decomposed into its constituent frames using the OpenCV library. This operation transforms the continuous video data into a series of discrete frames, thereby enabling more granular analysis.

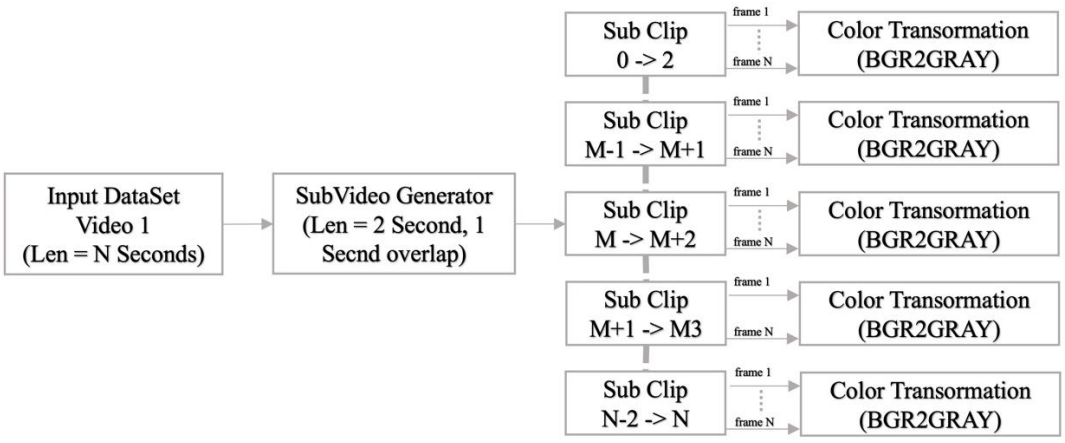


Figure 1: Single video pre-processing

Subsequently, a color transformation is applied to each frame, converting it from the Red-Green-Blue (RGB) color space to grayscale. This conversion serves multiple objectives. First, it minimizes visual noise and distractions inherent in the full-color spectrum, allowing for a more focused feature extraction process. Second, it enhances the efficacy of subsequent analytical steps by preserving meaningful information. Lastly, the transition to grayscale effectively reduces data dimensionality by eliminating chromatic variations, thereby streamlining the computational workload.

3.2.2. Feature Extraction

Upon completion of the data preprocessing steps, each segmented video is represented by N frames. The first analytical procedure involves calculating the differences between successive frames, yielding $N-1$ differential values. These frame differences serve as a metric for assessing variations in content between adjacent frames and are instrumental in capturing dynamic changes within the scene.

After the computation of frame differences, Local Binary Patterns (LBP) are applied to each differential value. LBP is a texture descriptor that quantifies the spatial configuration of pixels, and it has been widely employed in various image analysis tasks. When applied to frame differences, the LBP descriptor enables the extraction of salient features from the video clips that are pertinent to the study's objectives.

These individual LBPs are then amalgamated into a singular feature vector, as illustrated in Figure. 2. This composite feature vector offers a succinct yet comprehensive representation of the video, encapsulating essential information about its content. In essence, the feature extraction process distills key characteristics from the dataset's videos, thereby facilitating subsequent classification tasks.

The feature extraction pipeline, comprising frame differencing, LBP computation, and feature vector aggregation, results in a robust set of features that can serve various analytical requirements.

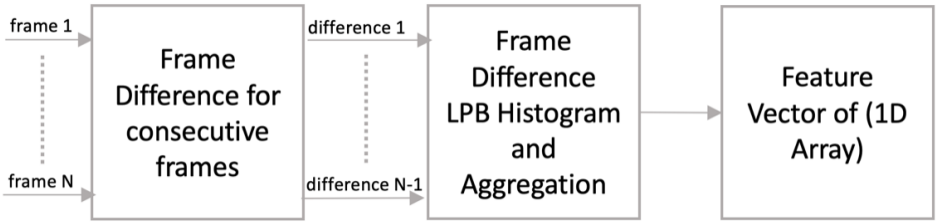


Figure 2: Single sub-clip feature extraction

3.2.3. Building Multiple AI Models

The training regimen incorporates a dataset of 360 primary samples which, subsequent to frame-sampling, expands to 3,240 sub-clips. Feature vectors extracted from each of these sub-clips serve as the input for this phase. The overarching goal is to train a variety of machine learning models to minimize the FAR.

In this phase, several classifiers are employed to discern between genuine and spoofed facial features. These classifiers include Support Vector Machine (SVM) with RBF Kernel, Nearest Neighbors, Gaussian Process, Decision Tree, AdaBoost, and Majority Voting.

The utilization of a diverse set of classifiers aims to engender a robust and comprehensive machine learning model adept at effectively identifying face liveness.

3.2.4. Constructing the Restrictive Voting Classifier

This research aims to eliminate the FAR while concurrently minimizing the False Rejection Rate (FRR). To achieve this objective, we developed an ensemble-based machine learning classifier known as "Restrictive Voting," which capitalizes on the combined efficacy of multiple classifiers.

Restrictive Voting amalgamates the predictions of these classifiers, weighting them based on their individual accuracies. The ensemble's aggregated prediction is then thresholded, as illustrated in Figure. 3. Specifically, each classifier operates on its dedicated OMP thread, independently processing sub-clips. Upon completion, a distinct thread consolidates the weighted outputs from all classifiers for each sub-clip set associated with a full video clip. A 65% threshold is then applied to the cumulated predictions to determine the video's final class. This threshold, chosen to further curtail the FAR, was established after iterative assessments of the system's performance under varying thresholds. A threshold of 65% was found optimal in achieving a zero FAR, bolstering the system's resilience against unauthorized access attempts.

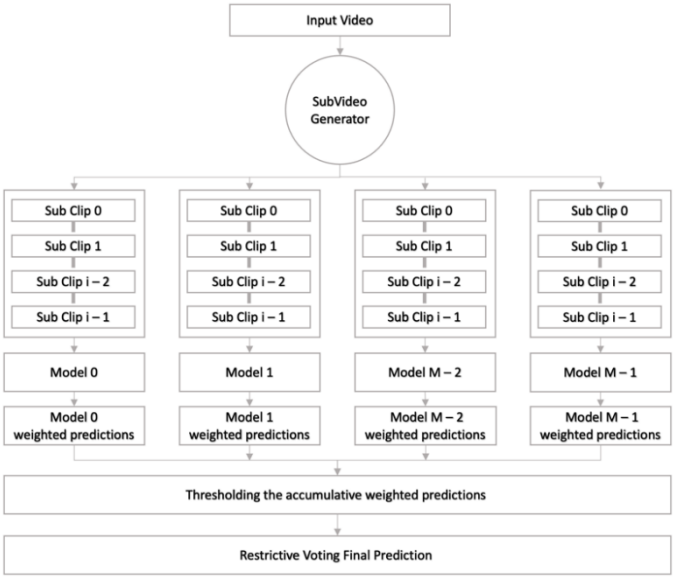


Figure 3: Restrictive Voting

Mathematically, the weighted average of classifications (WAC) across M classifiers for N sub-clip is given by Eq. (1):

$$WAC = \frac{\sum_{i=0}^{M} \sum_{j=0}^{N} w(i) * c(i,j)}{\sum_{i=0}^{M} w(i)} \qquad (1)$$

Herein,

- M: Total number of classifiers.
- N: Number of sub-videos.
- $w(i)$: Weight of the i^{th} classifier.
- $c(i,j)$: Classification by the i^{th} classifier for the j^{th} sub-video.

Based on the computed WAC, the final classification C is given by Eq. (2):

$$C = \begin{cases} 1, & if\ WAC > 0.65 \\ 0, & otherwise \end{cases} \quad (2)$$

The operational flow of the Restrictive Voting classifier is graphically depicted in Figure. 4.

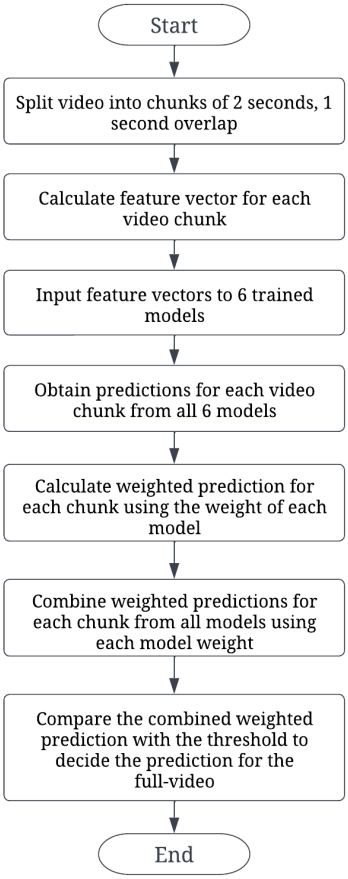


Figure 4: Restrictive Voting flowchart

Ensemble techniques, such as Restrictive Voting, harness the diversified strengths of individual classifiers, offering enhanced robustness. By integrating the decisions of these classifiers, the collective wisdom and specialization of each model are utilized, often leading to superior performance over

singular classifiers. As such, the Restrictive Voting classifier, as delineated in this study, promises robust and reliable performance across multifaceted datasets.

3.3. A Transfer Learning Approach with MoViNet

3.3.1. Data Preprocessing

In this study, the preprocessing phase is a critical precursor to ensure that the video inputs are both compatible and optimized for the MoViNet model. To achieve this, a well-defined sequence of steps is rigorously applied to the videos in the Replay-Attack dataset.

Initially, the OpenCV library is employed to break down each video file into individual frames. This operation converts the continuous nature of video data into a more manageable series of discrete frames, setting the stage for further processing.

Following this, a frame sampling technique is invoked to select a consistent number of frames from each video. Given that the Replay-Attack dataset comprises videos of varying lengths, it becomes imperative to standardize the number of frames used for training and evaluation. After iterative optimization, we settled on extracting 8 frames per video. This specific number was selected to strike a balance between computational efficiency and the richness of temporal information captured.

Subsequently, each extracted frame undergoes resizing and padding to fit the MoViNet model's specific input requirements. Frames are resized to a resolution of 224×224224×224 pixels. In instances where the original dimensions diverge from this target resolution, a zero-padding technique is implemented to maintain the aspect ratio, thus ensuring that all frames are consistently sized.

Moreover, a data normalization step is performed on the frames to facilitate the model's training convergence and stability. This involves scaling the pixel values to a common range of [0, 1], achieved by converting the pixel values to floating-point numbers and then dividing them by the maximum possible pixel value.

Finally, a color transformation step is enacted. Initially captured in the BGR color space, the frames are converted to the RGB format to align with MoViNet's input expectations, as well as the norms of most deep learning models.

Through the execution of these preprocessing steps, we ensure that the input data is tailored for both compatibility and performance optimization within the MoViNet model. This sets the foundation for effective face liveness detection, aiding in the capture of pertinent temporal information, ensuring spatial uniformity, and enhancing the model's overall performance.

3.3.2. MoViNet Model Architecture

To address the problem of face liveness detection, this study employs the MoViNet [1] architecture, a specialized deep learning framework designed for video classification tasks. MoViNet is distinct in its capability to amalgamate the computational efficiency of 2D frame-based classifiers with the rich temporal context captured by 3D convolutions.

One of the standout features of MoViNet is its use of causal convolutions along the temporal axis. This design choice ensures real-time processing capabilities by allowing the model to make predictions at any given time t based solely on data available up to that point. This is especially beneficial for streaming video data, where future frames are not available during the inference.

For the specific requirements of face liveness detection, the MoViNet architecture is particularly apt due to its incorporation of 3D convolutions. These allow the model to capture essential temporal dynamics, which is crucial for differentiating between genuine and spoofed faces.

To adapt the MoViNet architecture for this specialized task, we commence with a pre-trained MoViNet model to serve as the backbone of our system. Building upon this, we append a custom classification layer to fine-tune the model's capabilities towards our specific objective of face liveness detection.

By leveraging the MoViNet architecture in this manner, we aim to create a robust and efficient face liveness detection system that excels in capturing temporal dependencies while maintaining computational efficiency.

3.3.3. Model Training and Optimization

The training of the MoViNet model is a multi-step process geared towards fine-tuning the model's parameters to specialize in face liveness detection. This section delineates the comprehensive approach taken to train the model using the pre-processed data from the Replay-Attack dataset.

Fine-Tuning: Although MoViNet is a robust architecture, it was originally designed for general video analysis. To adapt it to the specific nuances of face liveness detection, we undertake a fine-tuning phase. This allows the model to adjust its internal parameters weights and biases to better recognize authentic versus spoofed faces.

Modification of the Output Layer: The original MoViNet model was configured for multi-class video categorization. However, our problem is a binary classification task: distinguishing genuine faces from spoofed ones. Accordingly, we modify the model's output layer to contain two neurons, thereby adapting it for binary classification.

Transfer Learning: Capitalizing on pre-existing knowledge, we employ transfer learning to initialize the MoViNet model with pre-trained weights from its original version. This strategy enables us to harness the model's well-established generalization capabilities and significantly hastens the training phase for our specific task.

Model Compilation: Before the actual training, we compile the model with appropriate loss and optimization functions. For this task, we opt for Sparse Categorical Cross-Entropy loss, given its suitability for classification tasks. The Adam optimizer, with a learning rate of 0.001, is chosen to iteratively update the model's parameters.

Training Loop: The training procedure involves several epochs wherein the pre-processed video frames and their corresponding labels are fed into the model. Each epoch comprises a forward pass through the

model to compute the loss, followed by a backward pass to propagate gradients and update the model's weights.

Early Stopping and Model Checkpointing: To mitigate overfitting and efficiently monitor model performance, early stopping, and model checkpointing strategies are employed. Early stopping ceases the training if the validation loss plateaus or worsens over a predefined number of epochs. Model checkpointing, on the other hand, preserves the model state that exhibits the lowest validation loss. In this study, the training was halted at the 16th epoch due to the absence of further improvements in validation loss.

Figure 5 presents the training and validation loss curves throughout the model training epochs. The graph shows a consistent decrease in both training and validation loss, which indicates that the model is learning and generalizing well. The proximity of the two loss curves throughout the process suggests that the model is not overfitting the training data.

Figure. 5. visualize the effectiveness of the training process. 'Model Behavior' graph which plots the training and validation loss across epochs. Initially, we observe a sharp decline in both losses, indicating that the model is learning effectively. As the number of epochs increases, the losses converge, which is a positive sign of the model's ability to generalize from the training data to the validation data. However, we note a slight uptick in the validation loss in the later epochs. This trend suggests the beginning of overfitting, which our early stopping strategy successfully mitigates by halting training before the model's performance on unseen data degrades. This graph serves as a crucial checkpoint in our model optimization process, ensuring that we maintain a balance between learning and the ability to generalize.

By meticulously following this structured training methodology, we aim to develop a MoViNet model that is both robust and specialized in the task of face liveness detection.

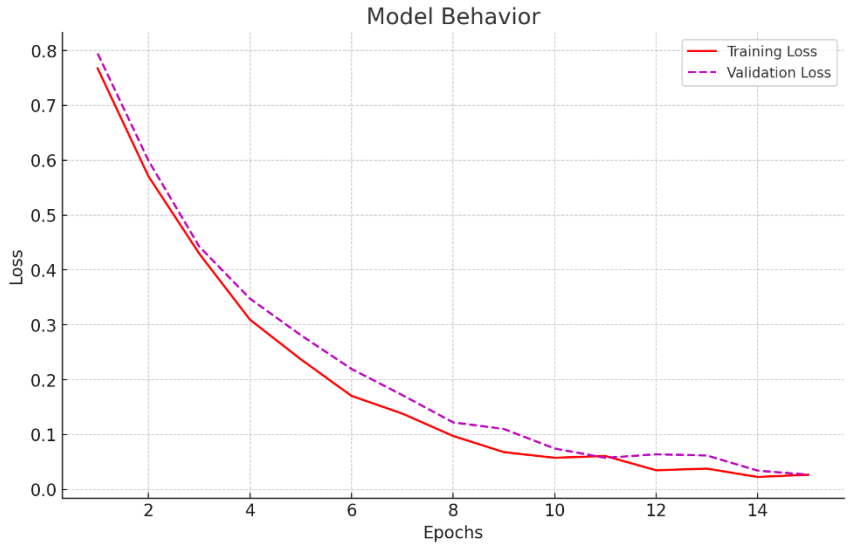


Figure 5: Training and Validation Loss Curves for the MoViNet Model

3.4. Model Evaluation

To rigorously assess the face liveness detection systems developed using both the Restrictive Voting and Transfer Learning methods, we employ a consistent evaluation framework.

The development and test subsets from the respective datasets are utilized for a comprehensive evaluation. The development subset is particularly instrumental in calibrating the decision thresholds, a crucial step for achieving our primary goal of zero FAR.

While a complete set of metrics, including FAR, accuracy, precision, recall, F1 score, EER, and HTER, are considered, special attention is given to achieving a zero FAR. This is aligned with our primary objective of ensuring the highest level of security by minimizing unauthorized access.

Our models' performances are benchmarked against existing state-of-the-art methods in face liveness detection.

The overarching goal of this evaluation strategy is to validate that both the Restrictive Voting and MoViNet models achieve zero FAR while maintaining reasonable FRR. Each model is trained on its corresponding preprocessed dataset, enabling the extraction of discriminative features crucial for distinguishing between genuine and spoofed faces.

4. Results and Discussion

This section offers a comparative evaluation of our two proposed methods for face liveness detection, specifically focusing on their performance metrics, comparative efficacy, and confusion matrices. Both methods were evaluated using the Replay-Attack dataset and aimed to achieve a zero FAR.

Both methods were evaluated using the same set of metrics: FAR, Accuracy, Precision, Recall, F1-Score, EER, and HTER. Table 1 consolidates these metrics for both methods, enabling a direct comparison.

Table 1: Comparative Evaluation Metrics

Metric	Restrictive Voting (%)	Transfer Learning Approach (%)
FAR	0	0
Accuracy	97.92	98.33
Precision	100	100
Recall	87.5	90
F1-Score	93.33	94.74
EER	2.5	5.0
HTER	2.75	5.0

To gauge the performance of our methods relative to existing techniques, both were compared based on EER and HTER metrics. As shown in Table 2, both methods exhibit competitive or superior performance compared to state-of-the-art approaches.

The confusion matrices for both methods, shown in Table 3 and Table 4, indicate a high rate of correct predictions for genuine and spoofed faces, further corroborating their robustness.

Table 2: Comparative Analysis on Replay-Attack Dataset

Method	EER (%)	HTER (%)
Fine-tuned VGG-Face[31]	8.40	4.30
DPCNN[31]	2.90	6.10
FASNet[32]	-	1.20
Multi-Scale[33]	2.14	-
Moire pattern[34]	-	3.30
Depth-based CNN[35]	3.78	2.52
Patch-based CNN[35]	4.44	3.78
Restrictive Voting	2.5	2.75
Transfer Learning Approach	5.0	5.0

Table 3: Confusion Matrix for Restrictive Voting

	Predicted Attack	Predicted Live
Actual Attack	400	0
Actual Live	10	70

Table 4: Confusion Matrix for Transfer Learning Approach

	Predicted Attack	Predicted Live
Actual Attack	400	0
Actual Live	8	72

Both the "Restrictive Voting" and the "Transfer Learning Approach" for face liveness detection have demonstrated effectiveness by achieving a zero FAR. However, the "Restrictive Voting" method shows a balance of high precision and computational efficiency, which may be advantageous for environments with computational constraints. It also offers flexibility due to its ensemble nature and requires less data preparation, facilitating quicker deployment compared to deep learning models that necessitate extensive training.

Conversely, the "Transfer Learning Approach" leverages the MoViNet architecture, providing a higher accuracy rate and better recall, albeit with a slightly higher EER and HTER compared to Restrictive Voting. This method is potent for scenarios where capturing temporal anomalies is critical, such as in advanced spoofing attempts involving movement. The deep learning foundation of this approach also suggests it is more scalable and adaptive to complex scenarios.

The selection between these methods should be guided by the specific operational demands, such as the necessity for either computational efficiency and rapid deployment, or for higher accuracy and the ability to discern complex spoofing tactics.

5. Future Work

While our study has demonstrated the efficacy of Restrictive Voting and MoViNet-based Transfer Learning in achieving robust face liveness detection with a zero FAR, several avenues for future research have emerged. Firstly, exploring the application of these methods in more diverse and challenging real-world scenarios would be invaluable. This includes environments with varied lighting, angles, and sophisticated spoofing techniques not represented in the Replay-Attack dataset.

Moreover, integrating the current approaches with other biometric systems, such as iris or fingerprint recognition, could offer a multi-modal defense mechanism, significantly enhancing security measures. Further investigation into reducing the computational overhead, especially for MoViNet-based Transfer Learning, would make these methods more accessible for real-time applications and devices with limited processing power.

Additionally, delving deeper into adversarial attacks and developing countermeasures will be crucial. As spoofing techniques become more advanced, continuously testing, and updating the models to detect new types of attacks will be paramount.

Lastly, as new deep learning architectures and optimization techniques are developed, continually refining, and adapting our models will be essential to maintain their effectiveness and efficiency. By pursuing these research directions, we aim to contribute further to the field of face liveness detection, ensuring the security and reliability of biometric authentication systems.

6. References

[1] D. Kondratyuk *et al.*, “MoViNets: Mobile Video Networks for Efficient Video Recognition,” *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 16015–16025, Mar. 2021, doi: 10.1109/CVPR46437.2021.01576.

[2] “Eye blink detection based on multiple Gabor response waves | Semantic Scholar.” Accessed: Feb. 28, 2023. [Online]. Available: https://www.semanticscholar.org/paper/Eye-blink-detection-based-on-multiple-Gabor-waves-Li/d8f6be181f6d8f508e77d73b4cb1284c27f5e178

[3] “Generalized face anti-spoofing by detecting pulse from face videos — Hong Kong Baptist University.” Accessed: Feb. 28, 2023. [Online]. Available: https://scholars.hkbu.edu.hk/en/publications/generalized-face-anti-spoofing-by-detecting-pulse-from-face-video-2

[4] T. de Freitas Pereira, A. Anjos, J. M. de Martino, and S. Marcel, “LBP-TOP based countermeasure against face spoofing attacks,” *Computer Vision - ACCV 2012 Workshops*, vol. 7728 LNCS, no. PART 1, pp. 121–132, 2013, doi: 10.1007/978-3-642-37410-4_11.

[5] “Secure Face Unlock: Spoof Detection on Smartphones | IEEE Journals & Magazine | IEEE Xplore.” Accessed: Feb. 28, 2023. [Online]. Available: https://ieeexplore.ieee.org/document/7487030

[6] "A liveness detection method for face recognition based on optical flow field | IEEE Conference Publication | IEEE Xplore." Accessed: Feb. 28, 2023. [Online]. Available: https://ieeexplore.ieee.org/document/5054589

[7] Z. Boulkenafet, J. Komulainen, and A. Hadid, "FACE ANTI-SPOOFING BASED ON COLOR TEXTURE ANALYSIS," 2015.

[8] F. M. Chen, C. Wen, K. Xie, F. Q. Wen, G. Q. Sheng, and X. G. Tang, "Face liveness detection: fusing colour texture feature and deep feature," *IET Biom*, vol. 8, no. 6, pp. 369–377, Nov. 2019, doi: 10.1049/IET-BMT.2018.5235.

[9] A. Fogelton and W. Benesova, "Eye blink detection based on motion vectors analysis," *Computer Vision and Image Understanding*, vol. 148, pp. 23–33, Jul. 2016, doi: 10.1016/J.CVIU.2016.03.011.

[10] T. Pfister, X. Li, G. Zhao, and M. Pietikäinen, "Recognising spontaneous facial micro-expressions," *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1449–1456, 2011, doi: 10.1109/ICCV.2011.6126401.

[11] L. Li, Z. Xia, A. Hadid, X. Jiang, H. Zhang, and X. Feng, "Replayed video attack detection based on motion blur analysis," *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 9, pp. 2246–2261, Sep. 2019, doi: 10.1109/TIFS.2019.2895212.

[12] A. Bakshi and S. Gupta, "Face Anti-Spoofing System using Motion and Similarity Feature Elimination under Spoof Attacks," *The International Arab Journal of Information Technology*, vol. 19, no. 5, 2022, doi: 10.34028/iajit/19/5/6.

[13] J. Määttä, A. Hadid, and M. Pietikäinen, "Face spoofing detection from single images using micro-texture analysis," *2011 International Joint Conference on Biometrics, IJCB 2011*, 2011, doi: 10.1109/IJCB.2011.6117510.

[14] 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, Apr. 2015, doi: 10.1109/TIFS.2015.2400395.

[15] T. Wang, J. Yang, Z. Lei, S. Liao, and S. Z. Li, "Face liveness detection using 3D structure recovered from a single camera," *Proceedings - 2013 International Conference on Biometrics, ICB 2013*, 2013, doi: 10.1109/ICB.2013.6612957.

[16] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, Sep. 2014, Accessed: Jul. 12, 2023. [Online]. Available: https://arxiv.org/abs/1409.1556v6

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2016-December, pp. 770–778, Dec. 2015, doi: 10.1109/CVPR.2016.90.

[18] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision." pp. 2818–2826, 2016.

[19] K.-L. Du and M. N. S. Swamy, "Recurrent Neural Networks," *Neural Networks and Statistical Learning*, pp. 337–353, 2014, doi: 10.1007/978-1-4471-5571-3_11.

[20] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Comput*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, doi: 10.1162/NECO.1997.9.8.1735.

[21] K. Cho *et al.*, "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation," *EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, pp. 1724–1734, Jun. 2014, doi: 10.3115/v1/d14-1179.

[22] R. Koshy and A. Mahmood, "Enhanced Deep Learning Architectures for Face Liveness Detection for Static and Video Sequences," *Entropy 2020, Vol. 22, Page 1186*, vol. 22, no. 10, p. 1186, Oct. 2020, doi: 10.3390/E22101186.

[23] Z. Xu, S. Li, and W. Deng, "Learning temporal features using LSTM-CNN architecture for face anti-spoofing," *Proceedings - 3rd IAPR Asian Conference on Pattern Recognition, ACPR 2015*, pp. 141–145, Jun. 2016, doi: 10.1109/ACPR.2015.7486482.

[24] X. Xu, Y. Xiong, and W. Xia, "On Improving Temporal Consistency for Online Face Liveness Detection System." pp. 824–833, 2021.

[25] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "VGGFace2: A dataset for recognising faces across pose and age," *Proceedings - 13th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2018*, pp. 67–74, Jun. 2018, doi: 10.1109/FG.2018.00020.

[26] Y. Kumar Sharma, A. Professor, and S. Pandurang Patil, "Deep Transfer Learning for Face Spoofing Detection An Empirical Study of Outlook difference among Indian Students towards ICT for Demography and Educational Standards View project Deep Transfer Learning for Face Spoofing Detection," vol. 22, no. 5, pp. 16–20, doi: 10.9790/0661-2205031620.

[27] O. Lucena, A. Junior, V. Moia, R. Souza, E. Valle, and R. Lotufo, "Transfer learning using convolutional neural networks for face anti-spoofing," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 10317 LNCS, pp. 27–34, 2017, doi: 10.1007/978-3-319-59876-5_4/COVER.

[28] R. Quan, Y. Wu, X. Yu, and Y. Yang, "Progressive transfer learning for face anti-spoofing," *IEEE Transactions on Image Processing*, vol. 30, pp. 3946–3955, 2021, doi: 10.1109/TIP.2021.3066912.

[29] H. Samma, S. A. Suandi, H. Samma, and S. A. Suandi, "Transfer Learning of Pre-Trained CNN Models for Fingerprint Liveness Detection," *Biometric Systems*, Aug. 2020, doi: 10.5772/INTECHOPEN.93473.

[30] "On the effectiveness of local binary patterns in face anti-spoofing | IEEE Conference Publication | IEEE Xplore." Accessed: Mar. 13, 2023. [Online]. Available: https://ieeexplore.ieee.org/document/6313548

[31] L. Li, X. Feng, Z. Boulkenafet, Z. Xia, M. Li, and A. Hadid, "An original face anti-spoofing approach using partial convolutional neural network," *2016 Sixth International Conference on Image Processing Theory, Tools and Applications (IPTA)*, Jan. 2016, doi: 10.1109/IPTA.2016.7821013.

[32] O. Lucena, A. Junior, V. Moia, R. Souza, E. Valle, and R. Lotufo, "Transfer Learning Using Convolutional Neural Networks for Face Anti-spoofing," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 10317 LNCS, pp. 27–34, 2017, doi: 10.1007/978-3-319-59876-5_4.

[33] J. Yang, Z. Lei, and S. Z. Li, "Learn Convolutional Neural Network for Face Anti-Spoofing," Aug. 2014, doi: 10.48550/arxiv.1408.5601.

[34] K. Patel, H. Han, A. K. Jain, and G. Ott, "Live face video vs. spoof face video: Use of moiré patterns to detect replay video attacks," *2015 International Conference on Biometrics (ICB)*, pp. 98–105, Jun. 2015, doi: 10.1109/ICB.2015.7139082.

[35] Y. Atoum, Y. Liu, A. Jourabloo, and X. Liu, "Face anti-spoofing using patch and depth-based CNNs," *IEEE International Joint Conference on Biometrics, IJCB 2017*, vol. 2018-January, pp. 319–328, Jan. 2018, doi: 10.1109/BTAS.2017.8272713.