International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Coal_MTCNN: Coalesed Multi Task CNN with residual network extractor for face liveness detection in biometric application

[1]M. Leelavathi, [2]Dr. D. Kannan

Submitted: 03/02/2024 **Revised**: 11/03/2024 **Accepted**: 17/03/2024

Abstract: Face liveness detection is the first step of the entire face detection technology, and it is critical to system security when using face identification technology. Biometrics has become a fascinating yet tricky field in the past ten years. Although one of the most hopeful biometrics methods is facial detection, it is susceptible to fake attacks. To shield biometric verification systems from false assaults using printed photographs, video recordings, etc., many academics concentrate on facial liveness detection. As a result, the Coalesced Multi Task Convolution Neural Network (Coal_MTCNN) is used in the present research. The appropriate characteristics are first compressed and extracted from the leftover with the attention layer using dimensionality reduction before being sent to the encoder. The encoder uses the residual network design to maximize model size. Three distinct convolutional layers merge concurrently in the centre of each residual network to gather extra information. The encoder is upsampled in the decoder to translate the input images pixel for pixel into the segmented output. The residual Attention Network is constructed explicitly by layering attention modules that produce attention-aware characteristics. The proposed Coal_MTCNN is examined regarding various factors on two datasets, including FDDB and WIDER FACE. It is discovered that it obtains 99.6% accuracy for FDDB and 98% accuracy for WIDER FACE.

Keywords- *face live detection, residual network, convolution neural network, feature extraction, biometric application*

1. INTRODUCTION

Biometric verification has repeatedly beaten standard password-based security methods [1]. In ancient days, personal identity was restricted. Currently, biometrics and computer vision can identify individuals without using documents or other identifying features [2]. Instead of using a person's connections, possessions, or private information to recognize them, biometrics can be used. With the help of modern technology, we were able to create fingerprints that greatly facilitate the identification and confirmation of people. , in response to the need for precise and machine-based identity. Written I.D.s have been replaced by biometric I.D.s, which can prove "who you are" without requiring a card or other piece of paper [3]. An essential stage in giving approved people access to the resources is verification. Conventional verification methods, such as PIN, card, and password combinations, cannot differentiate between genuine users and imposters who illegally gained entry to the system [4]. There are many possibilities for missing the PIN or password, misplacing, or losing the card. A biometric system is a tool that makes it possible to identify people automatically. The biometric identification method is user-friendly, so there is no need to remember a passcode, card, or PIN [5]. A biometric method called facial detection uses characteristics from a person's visage and matches them to information about others in a recognized image collection. Different approaches to facial detection have been created by researchers, who also overcame challenges like conflicting lighting, perspectives, and facial expressions[6].In the last ten years, it has expanded very quickly. It has been used in various industries, including purchases, facial detection in attendance systems, mobile device identification, investigations, and security access [7]. Face spoofing is one of the issues that coders encounter when installing a face detection system. When an intruder attempts to get past a facial detection system, that is called face spoofing [8]. The most well-known Face spoofing techniques involve printed pictures, videos, and 3D masks, which an assailant can use to obtain unauthorized access to a person with authorization and circumvent facial detection technology. One of many techniques used to stop facial spoofing assaults is face liveness detection. Since biometrics and passwords are the most popular security measures, facial liveness sensing is a comparatively novel technology [10]. However, many businesses must identify facial spoofing to stop any unauthorized entry to their networks. Intruders can be fooled into thinking they are a trusted staff member by holding a photo of a staff member in front of a security camera. Access to the images of real people is one of the

[1] *Research Scholar, Department of Computer Science, Pollachi College of Arts and Science, and Assistant Professor, Department of AI & ML, Sree Saraswathi Thyagaraja College,*
Coimbatore, Tamilnadu. India.
And Assistant Professor, Department of AI&ML, Sree Saraswathi Thyagaraja College, Pollachi.
[2]Professor, Department of Computer Science, Principal, Pollachi College of Arts and Science, Coimbatore, Tamilnadu, India.
Email: leelavathi.mphil86@gmail.com
Email: kd.khannan@gmail.com

security system's functions. Consequently, identifying liveness in the Face will be crucial in avoiding Face spoofing assaults [10]. Due to this, the following are the accomplishments of this work:

- To build a residual attention network using several attention modules. The actual implementation of the split focus strategy is the layered structure. As a result, various attentional patterns can be recorded in various attention modules.
- We can use this structure to create an end-to-end trainable network with top-down attention by simulating rapid bottom-up feedforward and top-down attention feedback in a singular feedforward process. In contrast to the traditional symmetric strategy used by the earlier approaches, we present an asymmetric encoder-decoder network in the merged model to reduce the model size further.

Existing works are briefly discussed in Section 2, the suggested strategy and methods are outlined in Section 3, and evaluation outcomes and discussion are presented and discussed in Section 4. The paper concludes with a summary and suggestions for future study in Section 5.

2. RELATED WORKS

According to the authors' understanding, surprisingly few comprehensive literature review articles discuss facial liveness detection. The suggested model in [11] uses Face mesh to find and identify faces. Face mesh allows the model to function under various circumstances, including changing backdrop and lighting conditions. The model can also handle non-frontal pictures of both sexes, regardless of their age or ethnicity. This yields a precision of 94.23%. Using the multi-task cascade neural network (MTCNN), a facial detection and classification system for illegal identity is created [12]. This technology will be able to identify offenders' features in real time and instantly discover their identities. The obtained precision is 86%. The quality evaluation is approached as a classification issue in [13], emphasizing challenging examples close to the categorization limit.

About this, paired binary quality pseudo-label is produced based on the facial similarity score alone, achieving 94% accuracy. With the help of image pairs, a deep Siamese network was trained [14]. Two real facial pictures or one real and one fake face image make up each image combination. The customer whose visage is represented in each set of photos. Joint Bayesian, contrastive, and softmax loss are used to train the deep Siamese network. It is suggested in [15] to use deep learning to implement an electronic polling system with facial detection. The anonymous signing method and blockchain technology are used to complete the voting procedure. The primary goal of the suggested plan is to examine the advantages of security and safety in an online polling system. The VGG-16 design created a patch-based convolutional neural network (CNN) with a deep component for face liveness detection for security improvement [16]. The method was examined using the REPLAY-ATTACK and CASIA-FASD databases. The findings show that, with decreased HTER and EER ratings of 0.71% and 0.67%, our method generated the outcomes for the CASIA-FASD dataset. As long as balanced and minimal HTER and EER values of 1.52% and 0.30% were maintained, the proposed strategy generated consistent findings for the REPLAY-ATTACK dataset. The lightweight convolutional neural network (CNN) architecture introduced in [17] is intended to identify features of persons in mines, avalanches, underwater, or other hazardous circumstances when their Faces may not be readily visible against a nearby backdrop.

No current comprehensive study concentrates on previously unknown presentation assault identification difficulties and issues. The present comprehensive literature analysis needs a complete assessment based on openly available datasets and concentrates on a specific subject, such as deep learning-based methods. The literature also needs a comprehensive analysis of the techniques for reliable and successful facial anti-spoofing systems.

3. SYSTEM MODEL

Initially, two datasets, such as FDDB and WIDER FACE, are adopted for face identification, and AFLW is adopted for face alignment. These datasets are pre-processed using linear image transform and modified adaptive thresholding technique. Then, feature extraction uses a Residual network with an attention mechanism. After feature extraction, face liveness is detected using a Coalesed Multi Task Convolution Neural Network.

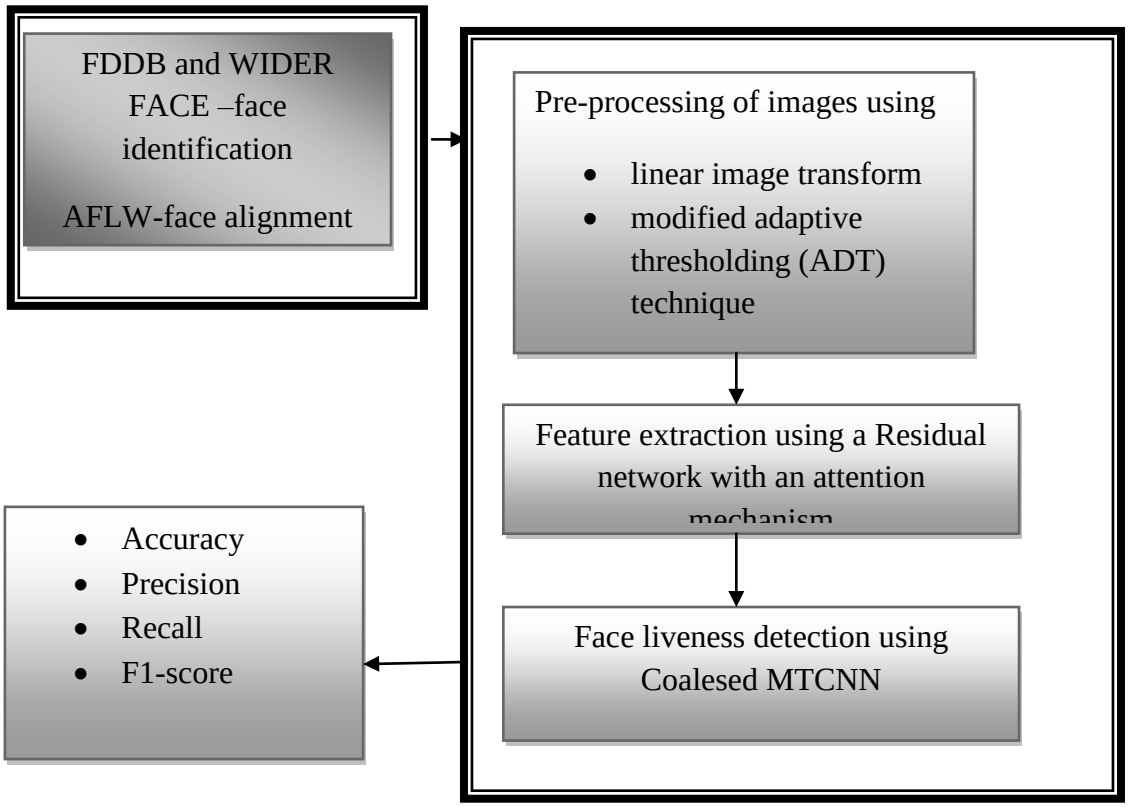

Figure-1 Overall architecture for face liveness detection

3.1 Pre-processing of images

Monotonic variations in the face picture are the most significant events that modify the outcomes of the facial detection program. One of these changes—illumination variation—shifts the image's pixel luminance, which alters the pixel difference, or $diff_{pix}$.

Here, we used modified adaptive thresholding (ADT) to switch out the zero thresholding for the pixel difference with an adaptive function, as is illustrated in the following example:

$$a_{pix}^{i} = \begin{cases} 1 & if diff_{pix} \geq ada\_Thres_i \\ 0 & if diff_{pix} > ada\_Thres_i \end{cases} pix = 1,2, \dots K \quad (1)$$

Here, $K$and I stands for the number of pixels and block number, respectively, and the term "ada_Thres_i" is the adaptive threshold function. In our suggested approach, the block dimensions and counts are determined.

However, the threshold adjusts to the local and worldwide information of the blocks and the picture. To have both robust and sensitive characteristics, we found the threshold function on the cumulative density function (CDF) of the Gaussian distribution function, as shown below:

$$feat_i = \frac{1}{2}[1 + \text{err\_fun}[\frac{\sigma_i}{\sigma\sqrt{2\sqrt{|\mu-\mu_i|}}}] \text{ where, } i = 1,2,\dots K \quad (2)$$

Where K is the total number of picture blocks, μ and σ are the global mean and standard deviation, and μ_i and σ_i are the mean and standard variation for each image frame. err_fundefines the error function as follows:

$$\text{err}_{\text{fun}} = \frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}\, dt \quad (3)$$

3.2 Feature extraction using Residual network

To accomplish multi-pose facial image recognition, a laser 3D scanning method is used to record the image of the subject's Face in various poses; this data is then merged with image imaging technology to determine the subject's biological features. Considering that the gathered multi-pose facial image has texture information and pixel feature components stored as $c_m = (R_m,\ G_m, B_m)$ and $aux_m(I_m, J_m, f_m, df_m, dl_m)$. The gray pixel values of the multi-pose face image are stored as $\varphi_{11}, \varphi_{12}, \varphi_{13}, \varphi_{14}, \varphi_{15}$ the multi-pose facial image's border features are broken down using the fuzzy feature recognition technique, and the edge pixel values are $\rho = \{\rho_{ij}(i,j)\epsilon S, \rho_{i,j}\epsilon S\}$. To derive features from multi-pose facial pictures, it is necessary to identify four kinds of points: solitary points, ends, straight points, and branching points. To address many issues, we suggest attention to residual learning. If a soft mask unit can be constructed as a specific mapping, as shown in fig. 2, its performance shouldn't suffer compared to its unattended counterpart. This idea is similar to that of residual learning.

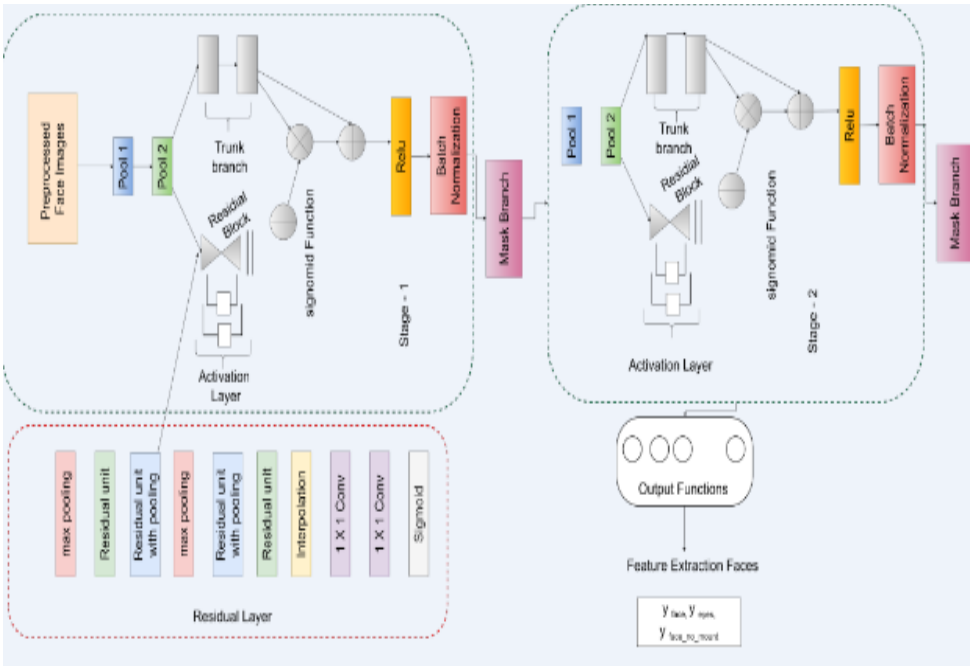

Figure-2 Residual network with an attention mechanism

As a result, we make the following adjustments to the Attention Module's output R:

$$R_{i,c}(y) = \left(1 + G_{i,c}(Y)\right) * D_{i,c}(y) \quad (4)$$

$G(y)$is in the interval [0, 1], and while its approximation function, $G(y)$, is close to zero, its reconstruction function, $R(y)$, will be close to the characteristics of the original, $D(y)$. T The term "attention residual learning" describes this strategy. The residual learning we present here, which we call attention-stacked residual learning, is unique. An approximation of the residual function, $D_{i,c}(y)$, is used in the original ResNet's formulation of residual learning as $R_{i,c}(y) = y + D_{i,c}(y)$. The characteristics produced by deep convolutional networks are denoted by the formula $D_{i,c}(y)$. We have the answer in our disguise of a tree's branches $mask(y)$.

They serve as feature evaluators, enhancing positive characteristics and muzzling undesirable ones. Increasing the density of the suggested Residual Attention Network can reliably enhance performance using attention residual learning. After only a few Residual Units, max pooling is applied several times from the input to expand the receptive field quickly. The worldwide information is then enlarged by a balanced top-down design to direct input features in each location after achieving the lowest resolution. Upsampling the result using linear interpolation after a few residual units. Maximum pooling equals the number of bilinear interpolations to maintain the output size constant with the input feature map. After two successive 1×1 convolution layers, the output range is normalized to [0; 1] by a sigmoid layer. We also inserted jump links between the bottom-up and top-down sections to include data from various dimensions. In our research, the focus offered by the mask branch adapts to the characteristics of the stem branch. However, by altering the normalization phase in the activation function before the soft mask output, constraints to focus can still be introduced to the mask branch.

We employ three activation functions to depict the three types of attention—combined, channel, and spatial. Use a basic sigmoid for each channel and geographic location when using mixed attention mix_att without additional restrictions. To eliminate spatial information, channel attention cha_att conducts $L2$ levelling across all channels for each spatial location. Spatial focus spa_t conducts normalization within each channel's feature map before applying a sigmoid to obtain a soft filter containing only spatial information.

$$mix_{att[y_{i,c}]} = \frac{1}{1+\exp(-y_{i,c})} \quad (5)$$

$$cha_{att[y_{i,c}]} = \frac{y_{i,c}}{\|y_{i,c}\|} \quad (6)$$

$$spa_{att[y_{i,c}]} = \frac{1}{1+\exp(-(y_{i,c}-\frac{mean_c}{std_c})} \quad (7)$$

where covers all channels, and i covers all spatial locations. The terms $mean_c$ and std_c indicate the feature map's mean and S.D. value from the c-th channel, respectively. The feature vector of the ith spatial location is denoted by y_i.

4. FACE LIVENESS DETECTION

Attackers who use photos, videos, or even 3D masks can fool the facial detection system when they engage in face spoofing. Regardless of the spoofing technique, the

attacker's sole objective is to make the facial detection system think the attacker is one of the system's clients. Face detection technology typically presents an accurate facial picture of its customers. We suggest a facial liveness detection technique based on client identification in light of the abovementioned finding. The foundation of our approach is depicted in Fig. 3.

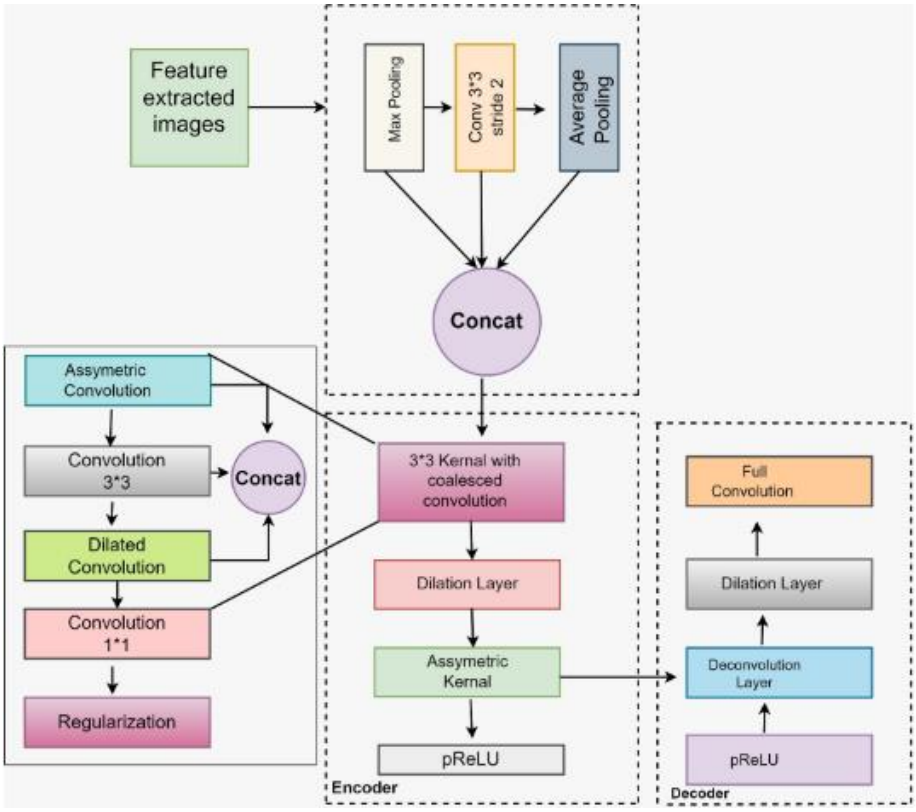

Figure-3 Architecture of Coalesced network

We present a kernel with a Coalesed convolution block, which ResNet designs imprudently initiated, as the central component of our network. The projection-receptive projection patterns with skip connections make up the RCC module. While the receptive section comprises three merged distinct convolutional layers, the projection portions are achieved by 1 X 1 convolution.

Let $h_j^i \in R^{a\times b}$ be the $a \times b$ c convolution kernel and let $h_j^i \in R^{a\times b}$ be the n-dimensional input of the merged convolution. The converged convolution's equivalent feature output can be represented by

$$L = [h_j^i * Y \cup h_j^i * Y \cup h_j^i * dil_Y] \qquad (8)$$

Where the broader convolution with the dilation factor dil is the final component, three steps comprise the complete encoder, and each level comprises five components. The standard convolution employs a centre of 3 X 3. The dilated convolutions' dilation factors vary from 2 to 32, while the values for the uneven kernels are 5 and 7. A batch n and a parametric rectified linear unit (PReLU) activation layer are inserted between the convolutional process inside the RCC modules. After that, we add a drop-out layer for regularization after RCC modules. The same RCC components used to build the encoder are used to build the decoder, except that a deconvolutional layer has been added instead of the merged convolutional layer, and there are fewer steps entirely. The motivation for this configuration is that the encoder should perform most of the image identification. The decoder's only responsibility is to upsample the encoder's output and correct the specifics. An ultimately linked convolutional ($Full\ Conv.$) layer is added behind the encoder to execute pixel-wise mapping. In summary, Table 1 displays the RCC-Net configuration with 3-channel input images and 11 face images. Table 1 provides a brief overview of the suggested network's RCC-Net setup, which includes three-channel input pictures and eleven facial image classifications.

Table-1 network configuration

stages	**convolution**		
	Ordinary	**Asymmetric**	**dilated**
input		4*245*369	
Encoder 1	5*5	12*325*240	4*57

Encoder 2	5*5	5*2;2*3	72*57
Encoder 3	5*5	8*2;2*5	8*45
Decoder 1	3*3	6*3deconv	
Decoder 2	3*3	7*3 deconv	
Full conv		45*23*124	

4.1 Performance analysis

Accuracy, precision, recall and F1-score were selected as the parameters for analysis. The proposed Coal_MTCNN is compared with three standard methods, namely Face mesh [11], MTCNN [12], and LightQNet[13], based on these parameters.

4.2 Dataset description

- A standard collection for facial detection is called WIDER FACE Dataset. With a high degree of diversity in size, posture, and shadowing, as shown in the example pictures, we select 32,203 images and identify 393,703 features. The 61 event types were used to arrange the WIDER FACE dataset. We arbitrarily choose 40%, 10%, and 50% of the data for each event type to serve as training, confirmation, and testing groups [18].
- Faces from the Faces in the Wild dataset are included in the Face Detection Dataset and Benchmark (FDDB) dataset. There are 5171 facial notes in total, and the pictures range in quality, e.g. 363x450 and 229x410 [19]

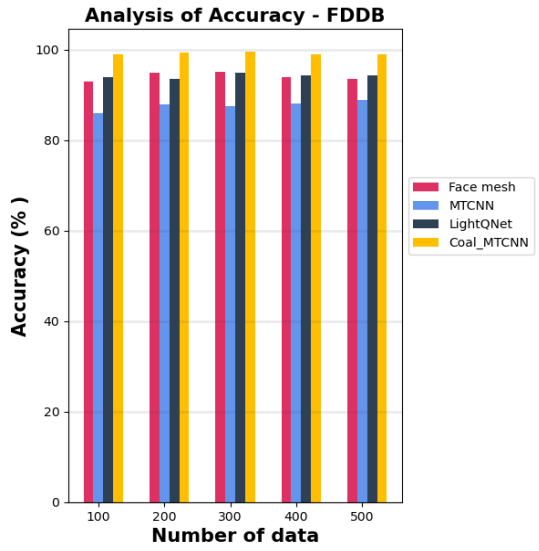

Figure-3 comparison of accuracy for the FDDB dataset

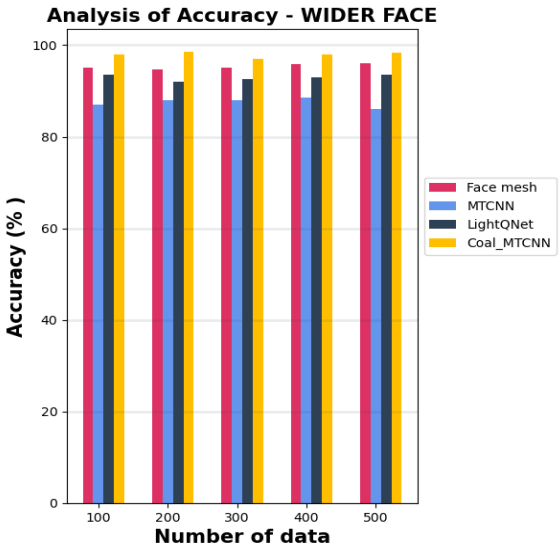

Figure-4 comparison of accuracy for the WIDER FACE dataset

Figures 3 and 4 depict the accuracy evaluation. When analyzing the FDDB dataset, the existing face mesh, MTCNN and LightQNet achieves 94.23%, 86% and 94% accuracy. In contrast, the proposed Coal_MTCNN achieves 99.6%, which is 2.43%, 13.6% and 5.6% better than the aforementioned existing methods; when analyzing the WIDER FACE dataset, we can see that the existing method achieves 95%, 87%,93.8%. In contrast, the proposed Coal_MTCNN achieves 98% of accuracy, which is 3%,11% and 5.8% better results.

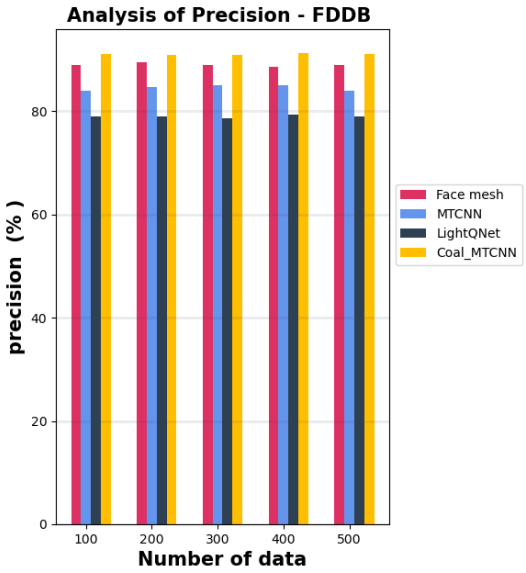

Figure-5 comparison of precision for the FDDB dataset

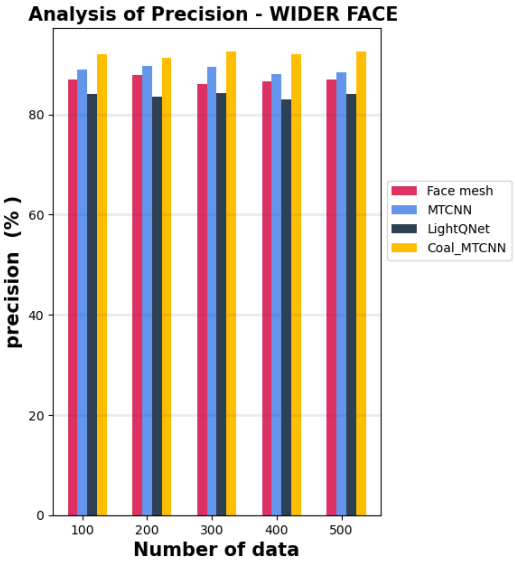

Figure-6 comparison of precision for the WIDER FACE dataset

Figures 5 and 6 depict the precision evaluation. When analyzing the FDDB dataset, the existing face mesh, MTCNN and LightQNet achieves 89%, 84% and 79% precision. In contrast, the proposed Coal_MTCNN achieves 91%, which is 2%, 7% and 12% better than the aforementioned existing methods; when analyzing the WIDER FACE dataset, we can see that the existing method achieves 87%, 89%,84%. In contrast, the proposed Coal_MTCNN achieves 92% precision, 5%,3% and 8% better results.

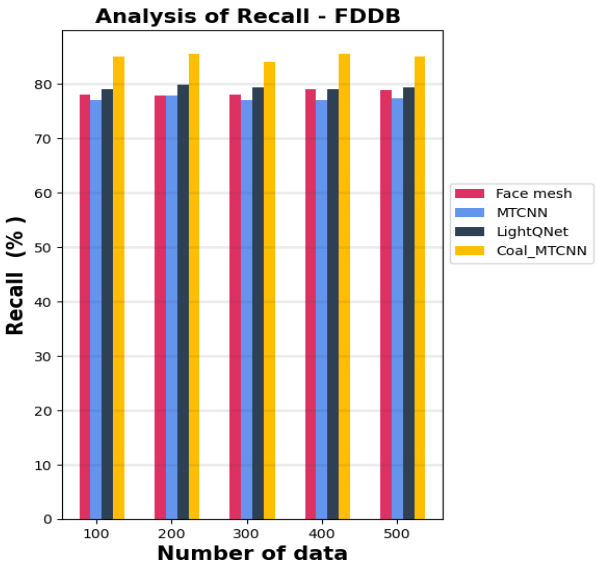

Figure-7 comparison of recall for the FDDB dataset

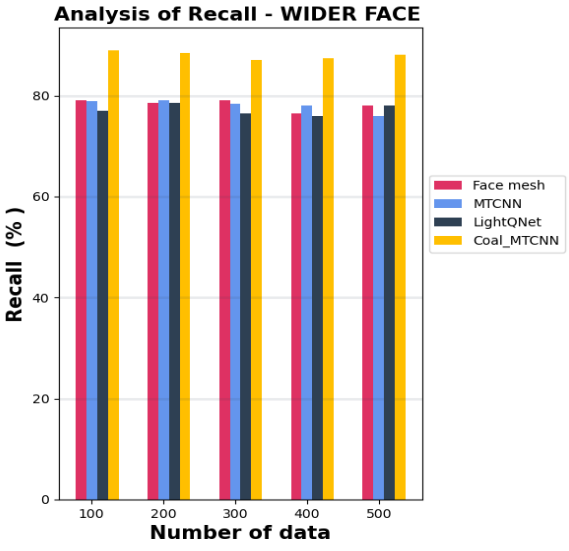

Figure-8 comparison of recall for the WIDER FACE dataset

Figures 7 and 8 depict the recall evaluation. When analyzing the FDDB dataset, the existing face mesh, MTCNN and LightQNet achieves 78%, 77% and 79.3% of recall. In contrast, the proposed Coal_MTCNN achieves 85%, which is 7%, 8% and 6.3% better than the aforementioned existing methods; when analyzing the WIDER FACE dataset, we can see that the existing method achieves 79%, 78.4%,77%. In contrast, the proposed Coal_MTCNN achieves 89% of recall, 10%,11.4% and 12% better results.

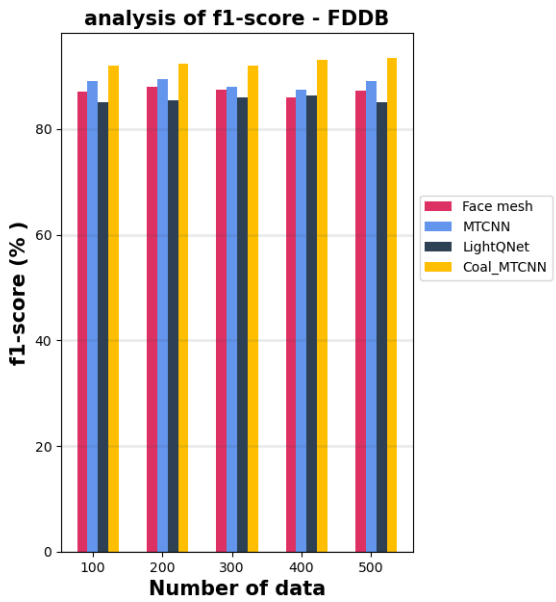

Figure-9 comparison of f1-score for FDDB dataset

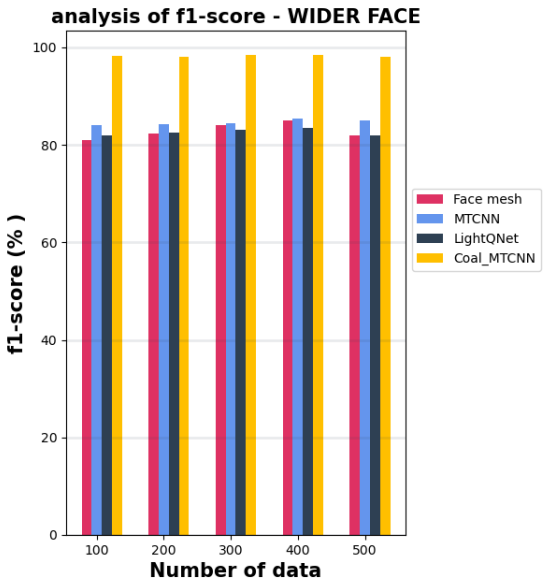

Figure-10 comparison of f1-sore for the WIDER FACE dataset

Figures 9 and 10 depict the f1-score evaluation. When analyzing the FDDB dataset, the existing face mesh, MTCNN and LightQNet achieves 87%, 89.5% and 85% of the f1-score. In contrast, the proposed Coal_MTCNN achieves 92%, which is 5%, 3.5% and 7% better than the aforementioned existing methods; when analyzing the WIDER FACE dataset, we can see that the existing method achieves 81%, 84%, and 82%. In contrast, the proposed Coal_MTCNN achieves 93.2% of f1-score, 12.2%,9.2% and 11.2% better results.

Table- 2 Comparison of existing and proposed methods for the FDDB dataset

method	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-score (%)**
Face mesh [11]	94.23	89	78	87
MTCNN [12]	86	84	77	89.5
LightQNet [13]	94	79	79.3	85
Coal_MTCNN [proposed]	99.6	91	85	92

Table- 3 Comparison of existing and proposed methods for the WIDER FACE dataset

method	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-score (%)**
Face mesh [11]	95	87	79	81
MTCNN [12]	87	89	78.4	84
LightQNet [13]	93.8	84	77	82
Coal_MTCNN [proposed]	98	92	89	93.2

CONCLUSION

Face liveness detection has been integrated into many security systems to prevent face fraud. The suggested Coal_MTCNN method attained a level of precision that is generally satisfactory. The feature extraction and detection stages comprise the first two phases of the

system. We exceed the current state-of-the-art techniques by achieving high-quality findings on facial liveness detection with a comparatively small model size by merging numerous convolutional layer types and stacking them in a deep residual network structure. Future research focuses on using ensemble convolution neural networks to increase accuracy and compute performance.

Reference

[1] Sébastien, M.S.; Nixon, J.F.; Marcel, N.E. Handbook of Biometric Anti-Spoofing, 2nd ed.; Springer: Cham, Switzerland, 2019.

[2] Sharma, S.B.; Dhall, I.; Nayak, S.R.; Chatterjee, P. Reliable Biometric Authentication with Privacy Protection. Adv. Commun. Devices Netw. 2023, 902, 233–249.

[3] Biometrics Recognition Using Deep Learning: A Survey. Available online: https://doi.org/10.1007/s10462-022-10237-x (accessed on 8 December 2022).

[4] Ross, A.; Jain, A.K. Biometrics, Overview. In Encyclopedia of Biometrics; Springer: Boston, MA, USA, 2015; pp. 289–294.

[5]]D. N. Parmar and B. B. Mehta, "Face recognition methods & applications," arXiv preprint arXiv:1403.0485, vol. 4, 2013.

[6] J. K. Sandeep Kumar and S. Singh, "A comparative study on face spoofing attacks," International Conference on Computing, Communication and Automation (ICCCA), 2017.

[7] T. O. Koichi Ito and T. Aoki, "Recent advances in biometric security: A case study of liveness detection in face recognition," Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 2017.

[8] J. Komulainen and M. Pietikainen, "Face spoofing detection from single images using texture and local shape analysis," International Conference on Computing, Communication and Automation (ICCCA), vol. 1, pp. 3– 10, 2012.

[9] B. Yaman Akbulut, Abdulkadir Sengur and S. Ekici, "Deep learning based face liveness detection in videos," international Artificial Intelligence and Data Processing Symposium (IDAP), 2017.

[10] Y. A. Abdulkadir Sengur, Zahid Akhtar and S. Ekici, "Deep feature extraction for face liveness detection," International Conference on Artificial Intelligence and Data Processing (IDAP), 2018.

[11] Hangaragi, S., Singh, T., & Neelima, N. (2023). Face Detection and Recognition Using Face Mesh and Deep Neural Network. *Procedia Computer Science*, *218*, 741-749

[12] Kumar, K. K., Kasiviswanadham, Y., Indira, D. V. S. N. V., & Bhargavi, C. V. (2021). Criminal face identification system using deep learning algorithm multi-task cascade neural network (MTCNN). *Materials Today: Proceedings*.

[13] Chen, K., Yi, T., & Lv, Q. (2021). Lightqnet: Lightweight deep face quality assessment for risk-controlled face recognition. *IEEE Signal Processing Letters*, *28*, 1878-1882.

[14] Pei, M., Yan, B., Hao, H., & Zhao, M. (2023). Person-Specific Face Spoofing Detection Based on a Siamese Network. *Pattern Recognition*, *135*, 109148.

[15] Revathy, G., Raj, K. B., Kumar, A., Adibatti, S., Dahiya, P., & Latha, T. M. (2022). Investigating E-voting system using face recognition using convolutional neural network (CNN). *Theoretical Computer Science*, *925*, 61-67.

[16] Muhtasim, D. A., Pavel, M. I., & Tan, S. Y. (2022). A patch-based CNN built on the VGG-16 architecture for real-time facial liveness detection. *Sustainability*, *14*(16), 10024.

[17] Wieczorek, M., Siłka, J., Woźniak, M., Garg, S., & Hassan, M. M. (2021). A lightweight convolutional neural network model for human face detection in risk situations. *IEEE Transactions on Industrial Informatics*, *18*(7), 4820-4829.

[18] http://shuoyang1213.me/WIDERFACE/

[19] V. Jain and E. Learned-Miller. FDDB: A benchmark for face detection in unconstrained settings. Technical report, University of Massachusetts, Amherst, 2010.