

Deep Segmentation Assisted Residual Spatio-Textural Ensemble Feature Learning Framework For Multimodal Person Authentication System

Chethana J¹, Ravi J²

¹ECE, Assistant Professor, Global Academy of Technology, Bangalore, Karnataka, India
RNS Institute of Technology, Bangalore, Karnataka, India

²ECE, Professor Global Academy of Technology, Bangalore, Karnataka, India
Corresponding author: Ravi J

Abstract- The last few years have witnessed exponential rise in global populations and allied demands; amongst which personal authentication and verification systems have gained widespread attention. The high-pace increasing intrusion cases and replication attacks too have alarmed industries to develop a robust security system for person identification and authentication (PAS). Unlike traditional cryptographic methods, biometrics driven PAS have played vital role due to its reliance towards imitability and forgery. Yet, merely reliance onto standalone biometrics like face or fingerprint can broaden space for intrusion, especially under the threat of generative artificial intelligence driven mimicking. Unlike traditional methods, touchless biometrics like touchless fingerprint, retina etc. can be vital, where their feature level fusion to perform AI-based classification can enable more attack-resilient solution for PAS. Despite the fact that the use of multimodal biometrics like fingerprint and iris can enable a robust PAS solution; yet, ensuring computational optimality by addressing at hand challenges like local minima and convergence over large feature space, instance-level class imbalance, gradient vanishing, gradient explode, accuracy degradation, etc. remains a challenge. Considering it as motivation, in this paper a robust deep segmentation assisted residual spatio-textural ensemble feature learning framework is proposed for multi-modal PAS solution. The proposed method intends to optimize both modalities with the suitable set of pre-processing, segmentation, feature extraction to be followed by feature level fusion and ensemble classification. To achieve it, at first the touchless fingerprint images obtained from the subjects were processed for the pre-processing tasks including image normalization, local orientation estimation, region masking and Gabor-filtering based smoothening. It helped to optimize local spatial features in touchless domain for better feature extraction and learning. The pre-processed touchless fingerprint images were then processed for deep spatio-textural feature extraction including gray-level cooccurrence matrix (GLCM) and DenseNet121. These features were locally fused to yield fingerprint specific composite feature vector. On the other hand, iris images were processed for UNet segmentation that retained ROI-specific color retinal images from each subject, which was later processed for the hybrid deep spatio-textural feature extraction by using GLCM and DenseNet121. Once extracting GLCM and DenseNet121 features from both modalities from each subject, horizontal concatenation was done to achieve feature level fusion. The fused composite features from both modalities (i.e., fingerprint and iris) were processed for principal component analysis (PCA) followed by z-score normalization and ensemble-of-ensemble classification. The proposed method applied bagging, AdaBoost, Random Forest, Extra Tree Classifier and XGBoost ensemble classifiers as base classifier that with the maximum voting ensemble method yielded consensus-based binary as well as multi-class classification for person authentication and verification. The simulation results confirmed that the proposed multi-modal PAS yielded person identification and classification accuracy of 98.84%, precision 98.91%, recall 99.21% and F-Measure 0.99, which is higher than the other state-of-art solutions. It confirms superiority, robustness and run-time efficacy of the proposed PAS solution.

Keywords— Personal Identification and Verification, Iris, Touchless Fingerprint, UNet segmentation, Deep Residual Features, GLCM, Ensemble Classification.

I. INTRODUCTION

The last few years have witnessed exponential rise in advanced computing and sophisticated hardware solutions to enable productive, scalable and time-efficient decision making. To cope up with real-time demands, different smart and intelligent systems have been developed including the IoT, IoT-assisted Smart Home, HMI, data systems and system access control units (ACUs), etc. [1-3]. The decentralized computing with mobile technologies, interoperable data infrastructures, online analytics etc. have emerged decisively over the years, which is still on rise. The exponentially rising smart-tools like IoTs, Smart manufacturing, telemedicine, cloud computing, edge computing etc. too are on rise [3][4][6][7]. In such diverse applications or environment, guaranteeing personal authentication and seamless data or system access becomes inevitable. Though, in the past different cryptographic methods have been applied for personal authentication; however, allied computational costs, latency and threat of breaches make most of the cryptographic methods confined towards real-time authentication and personal verification [1][2]. Unlike aforesaid cryptographic methods, biometrics techniques have performed better towards scalable personal authentication and personalized security tasks. Particularly, the different biometrics security solutions are designed by using physiological features (i.e., fingerprint, Iris, DNA, palm-print, veins, etc.) and behavioral features (gait, signature, voice etc.) [2]. Unlike traditional cryptographic methods, the ability to achieve minimum duplication or mimic ability makes biometrics methods robust over traditional cryptographic approaches. The reduced computational cost too makes biometrics methods more efficient towards real-world personal security or authentication solution [2-4]. In the last few years, numerous efforts have been made towards biometrics-based personal authentication and personalized security solutions. Amongst the major biometrics-based person authentication and/or personal security solutions, fingerprint-based security systems have gained widespread attention [4][9-13]. It exploits the different spatio-textural features from ridge and furrow, etc. to perform learning-based personal verification [2][4]. However, any inappropriate region-of-interest (ROI) detection and allied learning might yield false positive or reduced accuracy, thus confining its suitability towards scalable authentication services. Additionally, the last few years have witnessed repudiation and/or prototyping attacks by mimicking original fingerprints (i.e., replicating or mimicking user's minutia distribution or ridge-furrow distribution of the genuine user) [9-13]. This as a result makes overall security system vulnerable and suspicious. There are a number of e-infrastructures or ACUs where merely applying fingerprint biometrics can't guarantee attack-resilience. This is mainly because of the repudiation probability, replication likelihood by intruders [2]. To alleviate this problem, recently multi-modal biometrics concept has surfaced which exploits multiple types of biometrics measures including physiological and behavioral. In other words, multimodal biometrics methods employ features (combinations) of the different types to achieve high-accuracy and reliability for personal authentication [4-8]. In multimodal biometrics the use of secondary biometric cue helps improving verification results and hence makes overall system robust to any repudiation attacks or mimicking-based intrusion.

Amongst the major biometric (enterprise) solutions in e-infrastructures, fingerprint-based systems have remained dominant. Though, the recent pandemic of COVID-19 has forced industry to reassess efficacy of fingerprint biometrics to alleviate touch-based infections [1]. It has broadened the horizon for the discussions advocating touchless fingerprint detection methods. Yet, guaranteeing its robustness over the different local conditions such as the impact of lighting, contrast, viewing angle, gradience etc. can't be ruled out. Though, its efficacy with certain touchless biometrics such as iris, or gaits etc. can be the viable approach. For example, in India the UADI Aadhar verification system which applies both fingerprint as well as iris detail for personal authentication and/or

verification tasks is acknowledged to be one of the largest centralized verification solution serving billions of users concurrently. In function, fingerprint biometrics models require enhancement for related minutia detection, ridge and furrow detection and segmentation to achieve high accuracy [2][4][9-23]. On the other hand, unlike other touchless biometrics iris-based models have exhibited superior due to its non-repudiation ability and minimum likelihood of replication. Despite aforesaid robustness iris-biometrics which often involve edge detection, segmentation and allied feature learning demands more efficient and accuracy segmentation models as well as feature learning environments. The traditional iris detection and classification models often undergo inaccurate segmentation (say, ROI segmentation) due to non-linear spatio-textural feature presence. The traditional approaches employing edge detection and static threshold-based segmentation can't be hailed due to probable false positive over non-linear feature space. To alleviate such issues, performing automated ROI segmentation and ROI-specific feature extraction (and learning) can be a viable approach [21][22][36]. On the contrary, being reliant on ROI-specific feature learning, iris-based biometrics require more efficient and robust segmentation and feature learning methods. More specifically, it requires more effective fingerprint's minutia detection followed by ROI-specific feature extraction and learning from iris datasets [21]. As multimodal biometrics it needs feature level fusion to improve learning for person authentication [52].

Despite numerous efforts in the past majority of the state-of-art solutions failed to address certain challenges including instance level class-imbalance, non-linearity caused premature convergence and over-fitting problems, false positive and false negative evidences over large samples. Additionally, the deep network-based solutions become vulnerable of the gradient vanishing, gradient explode, accuracy degradation challenges that can have decisive impact on the overall reliability of the (biometric) authentication system. Considering such limitations and challenges developing a robust vision computing framework is inevitable which could address aforesaid challenges while strengthening reliability of the solution under scalable application environment. Considering discussed inferences and allied motivations, in this research a robust and efficient multi-modal person authentication system is proposed. Unlike standalone biometric driven authentication or person verification solutions the proposed method exploits touchless fingerprint and iris images to perform multi-modal person authentication. Unlike traditional 2D fingerprint-based solutions which requires human fingerprint physical imprints, the touchless fingerprints serve 3D representation of the user's input and hence serves better feature space for learning and prediction. It also avoids physical contact demands to serve touchless authentication solution. On the other hand, iris modality due to its resilience towards mimic provides a robust input to enable flaw-less authentication solution. Considering computational efficacy, being a multi-modal solution distinct pre-processing, feature extraction and learning mechanism is proposed. In this paper, a highly robust deep segmentation assisted residual spatio-textural ensemble feature learning framework for multi-modal PAS for run-time applications. Being a multi-modal PAS solution, the proposed method executes separate pre-processing and feature extraction, whose outputs are fused to yield composite feature set for further learning and classification towards PAS purpose. Initially, the proposed method made use of a robust information sensitive pre-processing tasks such as image normalization, local orientation estimation, region masking and Gabor-filtering based smoothening on Iris images, which was later processed for GLCM and DenseNet121 feature extraction. In this method, GLCM feature serves information rich feature set embodying energy, entropy, correlation, homogeneity, contrast, mean, standard deviation, and variance, while DenseNet121 yielded residual feature set, where are fused together to yield a composite feature set for further feature level fusion with the Iris features. Unlike touchless fingerprint modality,

iris images are more complex and can have the non-region-of-interest (non-ROI) details as well based on the camera acquisitions. The features from non-ROI regions can impact overall learning and prediction results thus impacting overall reliability of the solution. To address it, the proposed method at first executes semantic segmentation that retains ROI-specific images for further feature extraction and learning. To achieve it, the proposed method used UNet deep segmentation technique that retained ROI-specific (iris) details from each user, which were subsequently processed for GLCM and DenseNet121 feature extraction. Extracting GLCM and DenseNet121 features, horizontal concatenation is performed to generate deep spatio-textural feature (DSTF) set. The DSTF features from both fingerprint and iris were fused to generate composite feature set, which was later processed for PCA feature selection and z-score normalization to address local minima, convergence and over-fitting issues. The normalized multi-modal DSTF feature is trained by applying E2E ensemble learning model encompassing bagging, AdaBoost, Random Forest, Extra Tree Classifier and XGBoost ensemble classifiers as base classifiers. The proposed method yielded person identification and classification accuracy of 98.84%, precision 98.91%, recall 99.21% and F-Measure 0.99, which is higher than the other state-of-art solutions. The simulations confirmed superiority, robustness and run-time efficacy of the proposed PAS solution to meet real-world application demands.

The remaining sections of the manuscript are given as follows. Section II presents the related work, which is followed by research questions in Section III. Section IV presents the research method, while results and discussion are given in Section V. Conclusion is given in Section VI. The references used in this work are given at the end of the manuscript.

II. RELATED WORK

This section discusses some of the recent literatures pertaining to the fingerprint and iris-based PASs.

A. Fingerprint Biometrics for Person Authentication

Goel et al. [9] contributed CNN-based dual-identity fingerprint identification system. This method used CNN method for patch-learning to measure the cutline and the region-of-joint to perform dual-identity fingerprint detection for PAS. Anand et al. [10] exhibited pore-detection in high-resolution fingerprint inputs for PAS in cross-sensor authentication mechanism. The authors used residual learning-based CNN for pore detection for fingerprint detection and PAS. This model yielded the highest PAS identification accuracy of 88.12% with the F-Measure of 83.82%. Feng et al. [11] proposed an end-to-end fingerprint minutiae detection for PAS. The authors designed a lightweight pixel-wise local dilated neural network (DNN) to extract local features. It applied patch-wise global network for global feature extraction. The local as well as global features were applied for minutiae detection by using recursive connected component technique. It applied a loss-function technique for minutiae detection towards PAS solution. Zhang et al. [12] proposed FLDNet, designed as a light dense CNN to detect fingerprint liveness towards spoof presentation attack detection. The FLDNet model employed an attention pooling layer to alleviate the problem of global average pooling (GAP) in fingerprint anti-spoofing detection. Ali et al. [13] proposed CNN-based fingerprint pore detection for PAS. Here, initially CNN was applied to generate a pore intensity map, which is later processed for pore's centroid detection. Unlike traditional pore-detection techniques, the pore's centroid as feature enabled accurate PAS solution. Yuan et al. [14] designed multimodal feature fusion-driven fingerprint liveness detection for PAS. They applied weighted multimodal CNN (MCNN) for deep feature extraction and learning. Bhattacharya et al. [15] proposed standalone multi-attention fusion network for double-identity fingerprint detection system (SAMAF). SAMAF employed multi-instance attention technique to aggregate local and global discriminative contexts and/or attentive features across

multiple scales towards fingerprint detection-based PAS. González-Soler et al. [16] used local feature encoding technique on fingerprint images for presentation attack detection (PAD). It applied local as well as global features from fingerprint images for PAS, where it showed the highest accuracy of 96.17%. Spinoulas et al. [17] applied fingerprint-based PAS solution; yet, the cost and scalability remained a challenge. Yuan et al. [18] applied deep CNN (DCNN) for the ROI-specific feature extraction and learning for fingerprint liveliness detection. In order to improve local textural features, local gradient pattern was applied, which was later trained to detect fingerprint liveliness. Ji et al. [19] designed holistically-nested edge detection technique for edge detection, to be followed by maximum discrimination HOG feature extraction and XGBoost classification. The authors used k-NN algorithm to measure the best match within the dictionary by using mean-shift algorithm which reduced irrelevant feature elements. Hung et al. [20] emphasized on improving fingerprint detection accuracy under the different lighting conditions. They used CNN deep network to enable mobile phone light assisted fingerprint identification. The authors applied CNN with multi-class greedy autoencoder and variational autoencoder with the domain adaptation method to detect fingerprint accurately. With the use of the Bayesian optimization technique, they claimed to achieve detection accuracy of 99.67%. Lei et al. [21] applied FCM clustering with genetic algorithm (GA) to segment ROI, which was later processed for feature extraction and classification for person identification. Xun et al. [22] applied CNN for feature extraction from input fingerprint images, where extracted features were trained by using support vector machine (SVM) for object detection and verification. Galbally et al. [23] proposed a touchless full-3D fingerprint detection and classification method (3D-FLARE). The proposed techniques use photometric stereo 3D reconstruction from the different angles to improve PAS accuracy. The authors applied real depth (x,y,z) coordinate information pertaining to the ridge pattern to improve error-free fingerprint-based person identification. Bera et al. [24] initially applied hand contour details by applying finger-level shape presentation, which was followed by normalization that performed subtraction of the transformed binary image from binary hand contour image to achieve left-side of finger profiles (LSFPs). Later, it employed XOR to the LSFP image and hand contour image to generate the right side of finger profiles. This method made use of 30 geometric features for each normalized fingerprint, while in later phase the authors applied the rank-based forward-backward greedy algorithm to retain significant features to improve fingerprint classification accuracy. The authors applied RF classifier on the derived geometric features to perform PAS decision that achieved accuracy of 96.92%.

B. Iris-Based Person Authentication

Fang et al. [25] used both 2D and 3D textural features from the input iris images for iris detection and classification. They applied binary statistical image features (BSIF) for 2D feature extraction. The 3D iris features were extracted by applying photometric stereo method from input images. The feature vectors were used to examine the convexity of the measured iris surface. It helped to decide whether the user has worn contact lens. Khade et al. [26] designed iris-based attacks detection in which fragmental coefficients from the input iris images were applied to perform person identification. They applied discrete cosine transform (DCT) and Haar transform to extract hybrid features. The hybrid features were trained by applying seven machine learning algorithms including SVM, Naive Bayes (NB), RF, and decision tree (J48) and ensembles of SVM+RF+NB, SVM+RF+RT, and RF+SVM+MLP (multi-layer perceptron) for person identification. The authors applied RF classifier on hybrid features to perform person detection and identification that resulted accuracy of 99.95%. Choudhary et al. [27] exhibited score-level fusion of two distinct domain-specific features including multiple binarized

statistical image feature (BSIF) and DenseNet-based features for iris liveness detection. Tapia et al. [28] used cascaded deep learning model for spoofing resilient iris liveness detection. They used MobileNetV2 to learn features from the input iris images. Agarwal et al. [29] applied 3D contact lens for iris-based person identification. Yildiz et al. [30] applied webcam-based eye tracking system. They performed Hough transformation over circular shape of the iris data, while the central coordinates of the iris image were processed by k-NN classifier for iris-based person detection and classification. It showed average accuracy of 99.54%; though, the same method under ambient light showed the accuracy of 94.74%. Daniel et al. [31] applied Euclidean classifier, MLP classifier and the Hybrid Classifier for iris-based person identification. Choudhary et al. [32] detected the real and the forged iris. The authors used multi-scale BSIF and Fine-tuned DenseNet features together for person detection and identification. Hsiao et al. [33] applied YOLO deep network to detect iris position, which was later processed for VGG-16 feature extraction and learning. This method showed the highest accuracy of 98%. Sueishi et al. [34] developed a dynamic iris authentication system by using high-speed gaze and focus control by high-speed image processing. Prasanth et al. [35] applied iris along with other periocular biometrics for person detection and identification. This method applied eye region along with the periocular biometric region (eye regions like eyelids, eyelashes, and eyebrows) for person detection and classification. The authors applied CNN model to extract feature and learn from the input iris images. Hsiao et al. [36] applied UNet deep segmentation followed by EfficientNet deep model for feature extraction. The adaptive histogram equalization and Gabor filtering technique achieved the highest detection accuracy of 98%. In [37], EfficientNet deep network to perform iris-based person detection and identification. This method achieved person identification accuracy of 98.2%. Hachim et al. [38] applied CNN to extract features from the iris images, which were later used to generate key by applying linear discriminant analysis (LDA) technique to perform data encryption in cloud. The 3DES cryptographic method encrypted the generated key (from LDA output) and yielded accuracy of 100%. Ouda et al. [39] developed local ranking-based cancelable biometrics (LRCB) technique to protect iris-codes. In function, the irreversibility of the LRCB was revisited, while a new attack for reversing the LRCB transform was proposed. This attack model exploited the distribution of order statistics for discrete random variables to reverse the protected rank values which was used to retrieve a close approximation of the original iris template. Later, the ranking-inversion attack was applied to realize authentication against the LRCB transform. It can be exhaustive to meet run-time application demands. Jardine et al. [40] used iris features for person identification, where it showed the accuracy of 88.94%. Chen et al. [41] designed DBANet, a dual branch attention-based DNN for iris recognition. The authors at first extracted spatial feature with a small stride to preserve lower-level spatial detail, which was later processed for context feature extraction towards high-level feature extraction. With these feature sets, training and classification was done over UBIRIS. v2, CASIA-V4-Distance, and MICHE-I datasets towards person identification. It showed reduced error rates of 38.7%, 53.4%, and 71.9%, with the aforesaid input datasets, correspondingly. Jalal et al. [42] proposed SegNet for the joint semantic segmentation (iris and pupil), which was followed by AlexNet deep network-based feature extraction. It showed the highest accuracy of 97.7% with the MMU Iris dataset.

C. Multi-Metric Biometrics-Based Person Authentication

Bagwan et al. [43] developed a multi-metric biometric authentication system by applying face and iris images. The authors [44] applied multi-metric biometrics encompassing iris, face and fingerprint features for person identification. They applied VGG16 and Grassmann techniques for feature extraction and learning towards person identification.

This method showed the highest person identification accuracy of 92%. Chang et al. [45] proposed multi-metric biometrics by applying iris and fingerprint features. A cancelable template protection technique was proposed on the basis of the fingerprint and iris features. Subsequently, PCA over fused features from the fingerprint and the iris features. Here, applying random scrambling, wavelet transform and DFT operations the fused features were obtained, which were later used to perform person authentication.

III. RESEARCH QUESTIONS

In sync with the overall research goals and allied methodological scopes, this work defines certain questions (RQ). The justifiable answers to these questions can put foundation to a robust multi-modal (touchless) PAS solution for scalable applications. These research questions (RQ) are given as follows:

- RQ1: Can the strategic amalgamation of touchless fingerprint and iris biometric modality provide a robust AI solution for person identification and verification under scalable application environment?
- RQ2: Can the strategic implementation of GLCM and DenseNet121 feature learning with PCA feature selection and ensemble-of-ensemble (E2E) driven MVE classification enable highly accurate and reliable prediction toward person identification and verification?
- RQ3: Can the strategic implementation of UNet segmentation followed by (Iris) ROI specific deep spatio-textural feature extraction (GLCM + DenseNet121) and learning be effective towards multimodal person identification and verification?
- RQ4: Can the strategic amalgamation of Bagging, AdaBoost, RF, ETC and XGBoost base classifiers towards MVE consensus driven prediction enable accurate and reliable person identification and verification for run-time applications?

The overall research intends to achieve justifiable answers for these key questions so as to contribute a robust PAS solution.

IV. SYSTEM MODEL

This section discusses the overall proposed model and allied implementation. The proposed multi-modal feature extraction encompassing touchless fingerprint and iris modality followed by feature selection and classification are discussed in this section. The overall proposed method is depicted in Fig. 3. As depicted in Fig. 3, the overall process embodies the following steps:

1. Multi-Modal Data Acquisition
2. Feature Engineering
 - 2.1 Modality-1: Touchless Fingerprint
 - 2.2 Modality-2: Iris
3. Multi-Modal Feature Fusion
4. PCA Feature Selection
5. Z-Score Normalization
6. E2E MVE Ensemble Classification.
7. Performance Characterization.

The detailed discussion of the overall proposed method is given in the subsequent sections.

A. Multi-Modal Data Acquisition

Undeniably, multi-modality driven bioinformatics can enable a robust person authentication system or verification system; however, there exist no benchmark dataset available encompassing both iris as well as fingerprint dataset. On the other hand, this research proposes touchless fingerprint driven multimodal biometrics, where there

exists no benchmark available towards touchless benchmark data. The available datasets are mainly for the standalone biometrics driven person authentication including the iris [53], 2D fingerprint [54] etc. In this case, there is no publicly available benchmark data pertaining to the Iris-(touchless) fingerprint images for multiple subjects towards person identification or verification. Considering it as challenge, in this work a synthetic dataset was prepared encompassing both Iris as well as 3D (touchless) fingerprint images. The data was prepared in such manner that the proposed method can be executed on any available Iris- fingerprint dataset(s) towards person verification. In this work, the collected datasets from CASIA-v2, LivDet, Kaggle [53][54] were taken from both genders (Male as well as female). Though, the collected data doesn't confirm age of the subjects. We consider the data as a strategically prepared synthetic samples possessing both genders in varied age categories, where the overall PAS is trained to perform person authentication. The proposed method can be executed in any multimodal PAS setup with aforesaid diversities, where it can learn over the different image modalities to perform binary classification or person classification. The gender-specific or age-specific prediction is beyond the scope of this research. In addition to the Kaggle IRIS dataset, we also considered IITD Iris dataset, which can be publicly accessed with the prior request [53]. The IIT Delhi Iris Database mainly consists of the iris images collected from the students and staff at IIT Delhi, New Delhi, India. This database has been acquired in Biometrics Research Laboratory during Jan - July 2007 using JIRIS, JPC1000, digital CMOS camera [2]. The image acquisition program was written to acquire and save these images in bitmap format and is also freely available on request. The available database is from 218 different subjects; all the images are in bitmap (*.bmp) format. All the subjects in the database are in the age group 14-55 years comprising of 175 males and 43 females. The database of 2180 images are organized into 218 different folders each associated with the integer identification/number. The resolution of these images is 320×240 pixels and all these images were acquired in the indoor environment. From these set of images, we considered only 200 subjects images for further processing. A snippet of the input images or sample is given in Table I.

Table I Data Illustration

Modality	Samples			
Fingerprin t				
Iris				

B. Feature Engineering

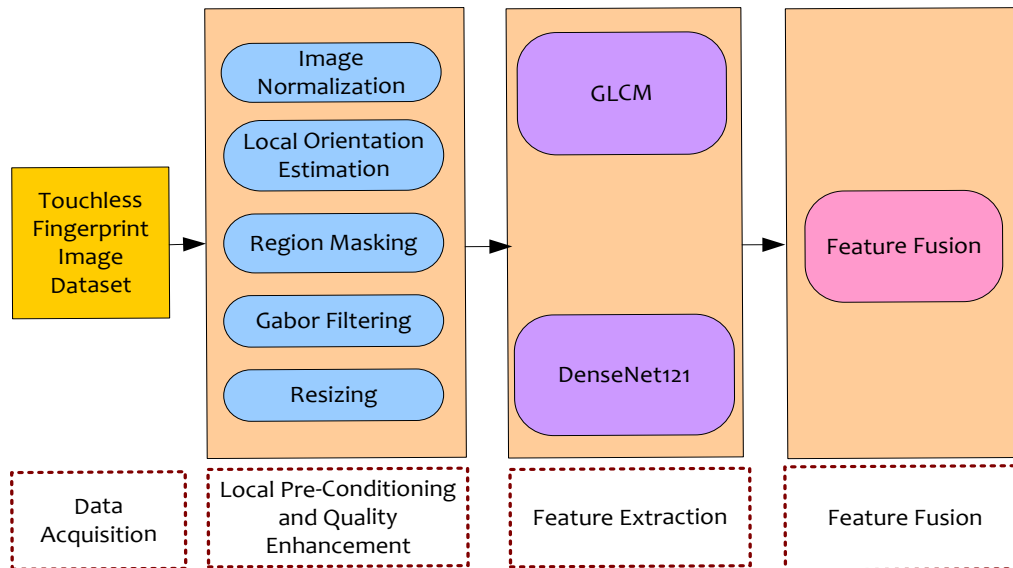
Since, in this work two different image modalities including the touchless 3D fingerprints and Iris images are taken into consideration towards a multi-modal person authentication system, the pre-processing and feature extraction measures are used differently. For

instance, the touchless fingerprint images require focusing on the effective ridge-furrow bifurcation followed by adaptive feature extraction. On the other hand, Iris images require extracting features pertaining to the Iris only. On the contrary, the real-world iris images can have the non-ROI (say, non-Iris image sections) image details, whose feature extraction and learning can impact accuracy and reliability of the person authentication system. It requires retaining ROI specific features for learning and prediction, for which semantic segmentation method can be well-suited. Considering data heterogeneity and disparity, in this work the different types of feature engineering techniques including data specific pre-processing, feature extraction and fusion are applied. Unlike traditional 2D fingerprint images, where there can be more linear and perceptible ridge-furrow details (say, spatio-textural details), the 3D touchless fingerprint images can have relatively poor image quality, surface quality and complex local conditions. Additionally, the touchless fingerprint images can have the impact of lighting conditions, orientation etc. Extracting features from such ambiguous and complex image inputs can impact overall feature learning efficiency, and hence can impact classification or verification results in scalable run-time environment. To address such challenge, it requires distinct pre-processing techniques for the different modalities. In this reference, the proposed method at first performs local pre-conditioning and fingerprint quality enhancement, which is followed by feature extraction and fusion. On the other hand, to alleviate possible instance-level class-imbalance in iris dataset, UNet segmentation is applied that retains ROI-specific features for learning and classification. The detailed discussion of the pre-processing, feature extraction and fusion methods is given in the subsequent sections.

a. Modality-1 Touchless Fingerprint

The overall proposed feature engineering methods involved in touchless fingerprint images is given in Fig. 3.

Fig. 3: Touchless fingerprint Feature Modality



The detailed discussion of the overall proposed feature model (Fig. 3) is given in the subsequent sections.

1. Data Acquisition

In this work, publicly available touchless fingerprint images are taken into consideration. Some of the input images were collected from open-source datasets like Kaggle, and MNIST. Though, due to the lack of benchmark multi-modal data in this work primary samples were also collected. A total of 200 subjects were considered for touchless fingerprint image collection. The images were acquired by normal camera placed at the different local lighting conditions. The camera armored with the Zeiss lens acquired five distinct fingerprint images from each subject. Thus, obtaining a total of 1200 images, including 1000 primary collections, a detailed pre-processing technique was applied that ensured that the training takes place effectively and can yield accurate prediction results under run-time environment. The proposed method intended to ensure optimal spatial (feature) distribution of each input image so as to achieve higher reliability and robustness. Before performing the proposed local pre-conditioning over the input images, a snippet of the image conditioning method (say, image spatial feature enhancement method) is given as follows:

2. Preliminary

Let, I be the input fingerprint image with $N \times N$ dimensional matrix, with $I(i, j)$ as the pixel intensity for the i - th row and the j - th column. In sync with touchless input, we hypothesize that the input images possess minimum resolution of 600 dots per inch, which is not difficult in contemporary high-definition camera. Thus, for the input images with aforesaid specification, the mean and the variance of the fingerprint image I in its gray-level form are derived as per the equation (1) and (2), respectively.

$M(I) = \frac{1}{N^2} \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} I(i, j)$	(1)
$Var(I) = \frac{1}{N^2} \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} (I(i, j) - M(I))^2$	(2)

Consider that the orientation image be O , with the dimension $N \times N$. Moreover, $O(i, j)$ be the local ridge representation at the pixel instant (i, j) . Noticeably, over touchless multi-dimensional image, the local ridge orientation can be characterized for a block in comparison to each pixel-based analysis. In this reference, an input fingerprint image can be split into multiple non-overlapping blocks (say, multiple $w \times w$), where each local ridge orientation is obtained for each block. Noticeably, in case of a typical fingerprint image the local ridge orientation at 900 and 2700 remains to be the same and can't be distinguished. This is mainly because the ridge oriented at both 900 and 2700 remains the same. Similarly, over a frequency image (say, F), which is defined with $N \times N$ dimensional image, let $F(i, j)$ be the local ridge frequency. Here, local ridge frequency refers the total number or the frequency of the ridge and furrow structure in adjacency, especially in orthogonal to the local ridge orientation. To be noted, the ridge and furrow structure in adjacency often shows a feature where minutiae doesn't constitute any specific and well-structured sinusoidal-shaped wave. In this case, the frequency is characterized as the mean frequency of its neighbors. Similar to the orientation image, frequency image is also characterized in block-wise manner. The proposed model also applies region-masking concept, which is also performed over each block sequentially to improve intrinsic features per image so as to improve better feature extraction and learning over touchless fingerprint images. In this work, region masking R is defined as $N \times N$ image possessing $R(i, j)$ signifying the type of the pixel. Thus, each pixel can be classified into two broad categories; first, non-recoverable non-ridge-and-furrow pixel

and second, recoverable ridge-and-furrow pixels, which are labelled as 0 and 1, respectively. Here, region masking was performed in block-wise manner, where each block is classified into aforesaid two-types. Now, once identifying the labelled ridge-furrow details the further processes like normalization, orientation estimation, frequency estimation and region masking, filtering etc. were done, sequentially. The details of these methods are given in the following sub-sections.

3. Local Pre-conditioning for Spatial Feature Optimization

The proposed local pre-conditioning model encompasses the following key processes:

1. Image Normalization,
2. Local Orientation Estimation,
3. Local Frequency Estimation,
4. Region Marking, and
5. Gabor Filtering and Smoothing.

As a pre-conditioning or pre-processing step, the proposed model at first performed normalization over each input fingerprint image. In normalization step the proposed model normalized in such manner that it retains a pre-defined mean and variance characterized. Once normalizing the input fingerprint image, the proposed model executed local orientation estimation that enabled measurement of the image's orientation, which is then followed by local frequency estimation. Subsequently, over the estimated frequency image, region masking was performed so as to classify each block as non-recoverable or recoverable ridge-and-furrow structure. Finally, region masked image blocks filtering was done to smoothen the input image. Unlike classical filter structures, in this work, modified Gabor filter was designed, which was tuned as per the local ridge orientation and frequency for each ridge-and-furrow pixel (over each block). The above stated approaches enabled solving major local adversaries including the impact of orientation, viewing angle, contrast etc. on local spatio-temporal feature distribution.

3.1 Image Normalization

Consider $I(i, j)$ be the gray-level value for the input touchless fingerprint image where (i, j) be the corresponding pixel values. Moreover, let M and Var be the measured mean and variance of the input image I . In this case, the normalized gray-level image for the input $I(i, j)$ can be obtained as $G(i, j)$, which is mathematically derived as per (3).

$$G(i, j) = \begin{cases} M_0 + \sqrt{\frac{Var_0(I(i, j) - M)^2}{Var}}, & \text{if } I(i, j) > M \\ M_0 - \sqrt{\frac{Var_0(I(i, j) - M)^2}{Var}}, & \text{Otherwise} \end{cases} \quad (3)$$

In (3), M_0 and Var_0 represents the expected mean and the variance values, correspondingly. Normalization was performed as a pixel-wise function and therefore it retained the native image clarity, especially ridge-and-furrow structure for further feature extraction and learning. Here, the key motive was to minimize the variations in the gray -level values in the direction of ridges and furrows to enable further processes more efficient without losing any intrinsic information.

3.2 Ridge Orientation Estimation

In reference to the touchless fingerprint detection problem, where the input image can have different spatio-temporal differences caused because of varying light conditions, change in orientation, spatial and temporal feature non-linearity, etc., to ensure optimal feature learning, the proposed model emphasized on improving ridge STF. To achieve improved ridge information and allied intrinsic values, ridge orientation estimation was performed, where orientation image (often called ridge orientation image) represents the intrinsic characteristic of the touchless fingerprint images defining invariant coordinates for the encompassing ridge and furrow structures (in a vicinity). Hypothesizing fingerprint image as certain oriented textural cue, numerous approaches have been suggested to perform orientation field in sample image. In this work, least-mean square (LMS) image orientation estimation method was applied for orientation image estimation. The proposed ridge orientation image estimation (ROIE) model is accomplished in multiple sequential steps. A snippet of the involved algorithm and allied (step-wise) implementation is given as follows:

Step-1: Split the input Gray-level image in $w \times w$ dimension or size. The proposed model considered 64×64 dimension to split input image into multiple grids.

Step-2: Estimate the gradient information in x and y directions for each pixel element (i, j) . Here, the gradient in x and y directions were, $\delta_x(i, j)$ and $\delta_y(i, j)$, respectively for the input pixel elements (i, j) . In this work, to ensure low computational overheads over a large input image, Sobel operator was applied to perform gradient estimation.

Step-3: Measure the local orientation values for each input block (Step-1), especially centred at the pixel element (i, j) by applying following mathematical formula.

$V_x(i, j) = \sum_{u=i-\frac{w}{2}}^{i+\frac{w}{2}} \sum_{v=j-\frac{w}{2}}^{j+\frac{w}{2}} 2\delta_x(u, v)\delta_y(u, v)$	(4)
$V_y(i, j) = \sum_{u=i-\frac{w}{2}}^{i+\frac{w}{2}} \sum_{v=j-\frac{w}{2}}^{j+\frac{w}{2}} (\delta_x^2(u, v) - \delta_y^2(u, v))$	(5)
$\theta(i, j) = \frac{1}{2} \tan^{-1} \left(\frac{V_y(i, j)}{V_x(i, j)} \right)$	(6)

In (6), $\theta(i, j)$ represents the LMS value of the local ridge orientation for the block centred at the pixel location (i, j) . In fact, ridge orientation signifies the direction which is orthogonal to the dominant direction of the Fourier spectrum over $w \times w$ window.

Step-4: Unlike touch-based fingerprint detection models, touchless image driven approaches might undergo more noise, reflections, dust related problems. In addition, touchless images can have the likelihood of the more damaged or corrupted ridge values or orientation, which can also be given rise due to the change in orientation or light intensity, contrast etc. Non-uniform skin surfaces too can show different spatio-temporal distribution for the ridge and furrow values (as well as minutiae) in touchless fingerprint images. In sync with such complexities and allied challenges, the estimated values of the local ridge orientation (i.e., $\theta(i, j)$) (6) can become inaccurate as well at certain time. In reference to these issues, we recall a hypothesis stating that as the local ridge orientation values vary gradually in local vicinity, especially in those neighboring localities where there is no singular point takes place or appear. In this reference, the use of a low-pass

filter (LPF) can be employed to manipulate or modify the incorrect local ridge estimation (6). Now, to achieve it (i.e., LPF) the proposed model converted the orientation image into a continuous vector field (CVF), which is mathematically derived as per (7) and (8). In (7) and (8), the variables Φ_x and Φ_y represent the x and y components of the vector fields, correspondingly.

$\Phi_x(x, y) = \cos(2\theta(i, j)), \text{ and}$	(7)
$\Phi_y(x, y) = \sin(2\theta(i, j))$	(8)

Thus, estimating above stated components (7) -(8), the proposed model performed LPF filtering by applying following mathematical approaches (9) -(10).

$\Phi'_x(i, j) = \sum_{u=-\frac{w_\phi}{2}}^{\frac{w_\phi}{2}} \sum_{v=-\frac{w_\phi}{2}}^{\frac{w_\phi}{2}} W(u, v) \Phi_x(i - uw, j - vw), \quad \text{and}$	(9)
$\Phi'_y(i, j) = \sum_{u=-\frac{w_\phi}{2}}^{\frac{w_\phi}{2}} \sum_{v=-\frac{w_\phi}{2}}^{\frac{w_\phi}{2}} W(u, v) \Phi_y(i - uw, j - vw)$	(10)

In (9) -(10), W states the two-dimensional LPS possessing single integral where the size of the filter is considered as $w_\phi \times w_\phi$. The proposed model performed smoothing at the block level where the filter size was fixed as 5×5 .

Step-5: Update the local ridge orientation at the pixel position (i, j) by using (11).

$O(i, j) = \frac{1}{2} \tan \left(\frac{\Phi'_y(i, j)}{\Phi'_x(i, j)} \right)$	(11)
---	------

Thus, applying above stated approach of smoothening and allied orientation image estimation, the proposed method obtained a uniformly smoothened (orientation) field image, which was further processed for frequency estimation.

3.3 Ridge Frequency Estimation

As stated in the previous sections, in case of touchless fingerprint images, especially when there are no minutiae in certain neighborhood, the gray-level values along ridges can be reconstructed as a sinusoidal wave. Noticeably, these sinusoidal-shaped waves are modelled towards the direction orthogonal to the local ridge orientation. Because of this reason, another key intrinsic feature from the input fingerprint images can be obtained in the form of local ridge frequency estimation. In other words, similar to the ridge orientation, ridge frequency can be modelled as an intrinsic feature for the touchless fingerprint images. In the proposed work, to estimate the ridge frequency information in a neighborhood we employed the pre-estimated measures like normalized image and the ridge orientation images. Let, G and O be the normalized and the orientation images, respectively. Then, with these values, ridge frequency was measured by using following sequential implementation approach.

Step-1: Split the input Gray-level image in $w \times w$ dimension or size. Here, the proposed model considered 64×64 dimension to split input image into multiple grids.

Step-2: Estimate the orientation window with fixed size $l \times w$ (i.e., 128×64) over each block, centred at the pixel information (i, j) .

Step-3: In reference to the Step-2, estimate the x –signature (say, $X[0], X[1], \dots, X[l - 1]$) of the ridges within the window, conditioned at:

$X[k] = \frac{1}{w} \sum_{d=0}^{w-1} G(u, v), k = 0, 1, \dots, l-1$	(12)
$u = i + \left(d - \frac{w}{2}\right) \cos O(i, j) + \left(k - \frac{l}{2}\right) \sin O(i, j)$	(13)
$u = j + \left(d - \frac{w}{2}\right) \sin O(i, j) + \left(\frac{l}{2} - k\right) \cos O(i, j)$	(14)

In case there exists no minutiae in the oriented window, the x-signature constitutes a discrete sinusoidal-shape wave, possessing the similar frequency as that of the ridges in (oriented) window. This as a result, enables estimation of the ridge frequency from x-signature. Consider that $T(i, j)$ be the mean pixel counts in between the two subsequent peaks in the x-signature, then the ridge frequency $\Omega(i, j)$ is measured as per (15).

$\Omega(i, j) = \frac{1}{T(i, j)}$	(15)
------------------------------------	------

In case there is no consecutive peak(s) available in x-signature, then the frequency is assigned a fixed value -1 that helps in differentiating it from the genuine frequency values.

Step-4: In case, the fingerprint images are taken over a predefined and definite resolution, then the value of frequency of the ridges within certain vicinity remains within a definite range. In case of 600 dots per inch resolution (often called DPI), this range remains within the level of $\left[\frac{1}{3}, \frac{1}{25}\right]$. In this manner, in case the measured value of the frequency becomes higher than the above stated range, the frequency is assigned a value -1, signifying that no genuine frequency could be estimated or observed.

Step-5: In touchless fingerprint images and corresponding block(s) where the minutiae or ridges are corrupted due to any local or personal regions, it doesn't constitute any well-structured sinusoidal wave. In this case, it becomes inevitable to interpolate those frequency values (of those specific blocks) from the frequency of the adjacent blocks possessing well-structured frequency. Here, we applied the following measures to perform interpolation, over each block centered at the pixel location (i, j) .

$\Omega'(i, j) = \begin{cases} \Omega(i, j) & \text{if } \Omega(i, j) \neq -1 \\ \frac{\sum_{u=w_{\Omega/2}}^{w_{\Omega/2}} \sum_{l=w_{\Omega/2}}^{w_{\Omega/2}} W_g(u, v) \mu(\Omega(i - uw, j - uw))}{\sum_{u=w_{\Omega/2}}^{w_{\Omega/2}} \sum_{l=w_{\Omega/2}}^{w_{\Omega/2}} W_g(u, v) \delta(\Omega(i - uw, j - uw) + 1)} & \text{Otherwise} \end{cases}$	(16)
---	------

where,

$\mu(x) = \begin{cases} 0, & \text{if } x \leq 0 \\ x, & \text{Otherwise} \end{cases}$ $\delta(x) = \begin{cases} 0, & \text{if } x \leq 0 \\ 1, & \text{Otherwise} \end{cases}$	(17)
--	------

In (16), W_g refers the discrete Gaussian kernel with mean and variance are assigned as 0 and 9, correspondingly. Here, the other components w_Ω be the size of kernels which was fixed at 7. In case there exists minimum one block possessing the frequency value of -1, then the value of Ω is swapped to Ω' , and the above stated process is repeated (Step-5). Step-6 Considering the gradual change in the inter-ridge distance variation, the proposed model applies LPF to eliminate the outliers in f' .

$F(i, j) = \sum_{u=w_{\Omega/2}}^{w_{l/2}} \sum_{v=w_{\Omega/2}}^{w_{l/2}} W_t(u, v) \Omega'(i - uw, j - uv)$	(18)
---	------

In (18), W_t represents the 2D LPF with single integral, while $W_t = 7$ be the filter's size.

3.4 Region Masking

As stated above, in real-time touchless fingerprint image a block or allied pixel can be either in non-recoverable or recoverable region. And therefore, classification of the blocks (or pixels) in above stated categories can be done on the basis of the wave's shape analysis (i.e., the shape of the waves formed by the ridges). This work employed three distinct features including amplitude (α), frequency (β) and variance (γ). Consider that, $X[1], X[2], \dots, X[l]$ be the x-signature of a specific block centered at the pixel position (i, j) , then the aforesaid three different features pertaining to that block are obtained as per the following approach.

Step-1: Assign the value of α as the mean height of the peak and the mean depth of the valley.

Step-2: Define β as $\frac{1}{T(i, j)}$, where $T(i, j)$ refers the number of pixels in between the two consecutive peaks (average value).

Step-3: Estimate the value of variance γ , as per (4.19).

$\gamma = \frac{1}{l} \sum_{i=1}^l \left(X[i] - \left(\frac{1}{l} \sum_{i=1}^l X[i] \right) \right)^2$	(4.19)
--	--------

Thus, applying this method the proposed model estimated a large number of three-dimensional patterns for each input image. Moreover, k-NN classifier was applied to classify each block of $w \times w$ dimension that classifies each input block as recoverable or non-recoverable so as to help identifying the most suitable set of feature blocks for future segmentation and allied feature extraction. In case a block (centered at (i, j)) was found recoverable, the corresponding region was estimated. In case, the fraction of the recoverable region was lower in comparison to a predefined threshold (here, $T_{Threshold} = 40$), the proposed model dropped that image from further feature extraction. Finally, an image with higher recoverable image with sufficiently large intrinsic features were considered for further feature extraction and learning, so as to improve fingerprint detection and classification more efficiently. Here, the proposed model labeled the recoverable region $R(i, j)$ as 1, while non-recoverable region is labelled as 0. Now, once identifying the optimal set of intrinsically enriched images, we performed filtering to improve spatio-temporal feature distribution. The details of the filtering method applied is given in the subsequent section.

3.5 Gabor Filtering and Smoothing

This is the matter of fact that the structure of the parallel ridges in fingerprint image, especially possessing well-structured orientation and frequency can provide sufficiently large intrinsic information to drop irrelevant and noisy components. On the other hand, the sinusoid waves pertaining to the ridges too change gradually in the local fixed orientation. Because of this reason, a bandpass filter can be designed in such manner (by tuning with respect to the frequency and orientation) that it would eliminate all unexpected or undesired noise components, while retaining the true ridge information for further learning. In reference to this scope, Gabor filter can be a viable solution as it possesses both orientation-selective characteristics as well as frequency-selective characteristics in both frequency as well as spatial domains. Considering this fact, we applied Gabor filter as the bandpass filter to eliminate noise components while preserving genuine ridge structures in fingerprint images. The Gabor filter can typically be presented as (20).

$h(x, y: \phi, f) = \exp \left\{ \frac{1}{2} \left[\frac{(x \cos \phi)^2}{\delta_x^2} + \frac{(y \sin \phi)^2}{\delta_y^2} \right] \right\} \cos(2\pi f \cos \phi)$	(20)
--	------

where ϕ refers the (Gabor) filter's orientation, while f represents the frequency of a sinusoidal wave. The components δ_x and δ_y be the space constants pertaining to the Gaussian envelope towards x and y , correspondingly. Here, the modulation transfer function of the considered filter is stated as per (21).

$H(u, v: \phi, f) = 2\pi\delta_x\delta_y \exp \left\{ -\frac{1}{2} \left[\frac{[(u - 2\pi/f)\sin\phi]^2}{\delta_u^2} + \frac{(u\cos\phi)^2}{\delta_v^2} \right] \right\} \\ + 2\pi\delta_x\delta_y \exp \left\{ -\frac{1}{2} \left[\frac{[(u - 2\pi/f)\sin\phi]^2}{\delta_u^2} + \frac{(u\cos\phi)^2}{\delta_v^2} \right] \right\}$	(21)
---	------

In (21), $\delta_u = 1/2\pi\delta_x$ and $\delta_v = 1/2\pi\delta_y$.

To implement Gabor filters over each input touchless fingerprint image, three different parameters including the frequency of the sinusoidal wave u_0 , filter orientation and standard deviation of the Gaussian envelope in the different directions, δ_x and δ_y . Here, the frequency characteristics of the filter, f is estimated by employing local ridge frequency and the ridge orientation values. In the proposed model, the selection of trade-off between δ_x and δ_y is maintained in such manner that higher the trade-off, more noise tolerant. However, it might cause spurious ridges information. On the contrary, smaller the values, the lower the Gaussian envelope, δ_x and δ_y . However, it might be less effective towards noise elimination. In this work, δ_x and δ_y values were assigned as 4.0, each. Now, consider that the input gray-level input fingerprint image be G , O be the ridge orientation image, while F be the ridge frequency image, and R be the recoverable mask. Then, the improved fingerprint image ε is obtained using the following equation.

$= \begin{cases} \varepsilon(i, j) & \text{if } R(i, j) = 0 \\ \sum_{u=-w_g/2}^{w_g/2} \sum_{v=-w_g/2}^{w_g/2} h(u, v: O(i, j), F(i, j)G(i - u, j - v)), & \text{Otherwise} \end{cases}$	(22)
--	------

Thus, the improved images are processed for the proposed hybrid feature extraction method (HDSTF) where it applies GLCM and DenseNet121 feature extraction method, distinctly. A brief of the proposed feature extraction methods is given in the following sections.

4. Residual Deep Spatio-Textural Feature Extraction

Once performing pre-conditioned local conditioning as pre-processing task, as discussed in the previous section, the proposed model executed hybrid feature extraction by applying two distinct feature extraction models, GLCM and DenseNet121. Here, GLCM serves as a spatio-textural feature extraction method, where different spatio-temporal descriptive statistical features are obtained over the pre-processed or the improved touchless fingerprint images. On the other hand, it applied DenseNet121 as a second feature extraction model that yields highly significant and information rich residual features from the input fingerprint image. Unlike classical fingerprint detection models, it avoided the traditional minutiae detection and learning, rather it employed GLCM-DenseNet121 as a hybrid feature extraction model (i.e., HDSTF) to retrieve deep spatio-temporal features. Here, the only key motive was to retain optimal set of features with different statistical feature distribution along with residual features that cumulatively could achieve superior feature vector for further learning and classification. The discussion of the proposed feature selection method is given in the subsequent sections.

4.1 GLCM Feature Extraction

In the last few years, GLCM method has emerged as a vital descriptive statistical feature distribution model that examines the likelihood of the pixel's gray scale values over an input image (here, fingerprint and iris images). In function, it extracts high-dimensional statistical features including entropy, energy, contrast, homogeneity, mean, standard deviation, variance, etc. encompassing statistical features, orientational features and textural distributions or features. In this work, we hypothesize that once performing ROI-specific color space reconstruction, the different spatio-textural and orientational features are distributed uniformly across the input images. Thus, over each input image the different GLCM features were obtained, which were later horizontally concatenated to yield a composite feature vector for learning and classification. Here, the extracted spatio-textural features were obtained as a matrix signifying pixel intensities $I(x, y)$, centered on the pixels (x, y) . These features eventually construct a probability matrix $P_{i,j}$ for each input image. Noticeably, $P_{i,j}$ states the disparities of the intensity between the pixel i and j that subsequently enables detection of the motion pattern(s). Typically, in GLCM method the gray scale value signifies the pair relatedness along a specific direction or orientation, and hence obtaining aforesaid gray-scale values a matrix called association matrix can be obtained for each image. With this motive, we retrieved a symmetric matrix S by combining gray-scale values and corresponding transpose values. We obtained the cumulative associations amongst pixels in a direction (say, symmetric matrix) where the normalized matrix S was obtained to yield $P_{i,j}$ (23).

$P_{i,j} = \frac{S_{i,j}}{\sum_{i,j=0}^{N-1} S_{i,j}}$	(23)
--	------

Once obtaining $P_{i,j}$ matrix for each input image, the different GLCM features were obtained. A total of seven GLCM features encompassing statistical distribution features, textural features and orientation features were obtained. These features were:

- Entropy,
- Contrast,

- Energy,
- Homogeneity,
- Correlation,
- Mean,
- Standard deviation, and
- Variance.

A brief definition of these GLCM features is given as follows:

— **Contrast**

Contrast represents the spatial variation of the gray-scale value across an input image. In (24), the pixel pairs being the diagonal element of $P_{i,j}$ signifies higher disparity in contrast. In terms of textural feature distribution, contrast refers the overall change in pixel intensities over input image. The allied pixel disturbances across the spatial regions can be done by statistical assessment and allied uniformity (distribution) analysis. Mathematically, we applied equation (25) to measure contrast for each input image (i.e., ROI specific fingerprint and iris images).

— **Homogeneity**

Dissimilar to the contrast feature, homogeneity refers the Inverse Different Moment (IDM). Typically, contrast and homogeneity features are defined inversely related features where higher contrast indicates lower homogeneity in an image. We used equation (23) to measure homogeneity.

$HOM = \sum_{i,j=0}^{N-1} \frac{P_{i,j}}{1 + (i - j)^2}$	(24)
--	------

GLCM features embody the values distributed across the grids. In case, GLCM embodies a meagre fraction of the large magnitude, the image used to have high contrast and therefore low homogeneity. On the contrary, for higher uniform magnitude across the image, it signifies high homogeneity and hence low contrast, where the later (i.e., contrast) is measured as per (25).

$CONT = \sum_{i,j=0}^{N-1} P_{i,j} (i - j)^2$	(25)
---	------

— **Energy**

In order to measure energy distribution across an input image, angular second moment information (ASM) is obtained by using equation (26). Here, ASM method measures the rotational acceleration across the input image (feature) space (26). Typically, the value of ASM increases with the rise in the uniform distribution of the gray-scale values within an image.

$ASM = \sum_{i,j=0}^{N-1} P_{i,j}^2$	(26)
--------------------------------------	------

Obtaining the value of ASM (26), corresponding energy value (for each image) was obtained by using equation (27).

$ENR = \sqrt{ASM_{i,j}}$	(27)
--------------------------	------

— Entropy

Generally, entropy feature states the extent to which the gray-scale values are randomly distributed across the input feature space (per image). With rising irregularity in pixel distribution, the entropy value increases. We used equation (28) to measure energy GLCM feature for each input ROI-specific image.

$ENT = \sum_{i,j=0}^{N-1} P_{i,j} (-\ln P_{i,j})$	(28)
---	------

— Correlation

In addition to the aforesaid textural features, to ensure fairly distributed heterogenous feature, we obtained descriptive statistical features as well. In this work, as statistical features, we obtained mean, standard deviation, variance and correlation for each input image. More specifically, the equation (29) was used to measure mean, while variance and standard deviation values were obtained by using (30) and (31), respectively. Because of the symmetric characteristics of (23), the mean values (i.e., (29)) remains equal (i.e., $\mu_i = \mu_j$).

$\mu_i = \sum_{i,j}^{N-1} i(P_{i,j})$ $\mu_j = \sum_{i,j}^{N-1} j(P_{i,j})$	(29)
$\sigma_i^2 = \sum_{i,j}^{N-1} P_{i,j} (i - \mu_i)^2$ $\sigma_j^2 = \sum_{i,j}^{N-1} P_{i,j} (j - \mu_j)^2$	(30)
$\sigma_i = \sqrt{\sigma_i^2}$ $\sigma_j = \sqrt{\sigma_j^2}$	(31)

By using mean and variance features, we obtained correlation feature for each input image. We applied equation (32) to estimate correlation feature for each input image.

$CORR = \sum_{i,j}^{N-1} P_{i,j} \left[\frac{(i - \mu_i)(j - \mu_j)}{\sqrt{(\sigma_i^2)(\sigma_j^2)}} \right]$	(32)
---	------

Now, once extracting aforesaid GLCM features (i.e., contrast, energy, entropy, homogeneity, correlation, mean, standard deviation and variance), horizontal concatenation was done to obtain a composite spatio-textural feature, as given in (33).

$F_{GLCM} = [ENE, ENT, CONT, Homogeneity, CORR, VAR, Mean, SD]$	(33)
---	------

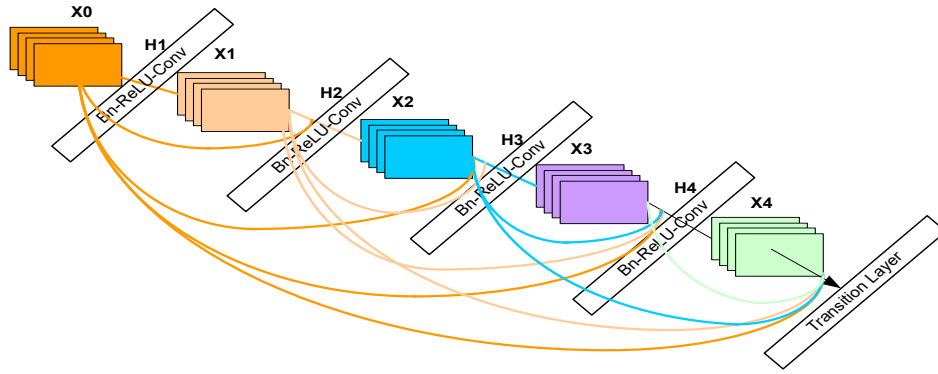
4.2 DenseNet121 Deep Feature

The dense convolutional networks, mostly known as DenseNet represents a kind of residual deep network that with a large number of filters can generate high dimensional feature possessing more intrinsic details for learning and classification. Same as other residual networks, in DenseNet121 deep network, each layer retrieves additional input(s) from the descendent layer, which is later passed as input to the subsequent layer. The architecture of the deployed DenseNet121 deep network is given in Fig. 4. As shown in Fig. 4, the DenseNet121 deep network embodies Dense Blocks with multiple convolution layers, where the FC layers are connected to the 121 convolution layers. In this work, DenseNet121 features are merely applied as a feature extraction layer, the output from FC layer is fused to the other feature sets for learning and classification. The proposed model designed DenseNet121 as hierarchical structure, in which the features from the $(l - 1)th$ layer is inputted to the $l - th$ layer by applying feed-forward network. It performed source to the next layer transition (i.e., $X_l = H_1(X_l - 1)$), with l as the layer index. Additionally, H_1 represents the composite function embodying sub-layers such as the batch normalization, ReLU, pooling etc. As shown in the deployed model (Fig. 5), the output features from the previous layer are mapped to certain target layer or defined layer, where feature reuse ability helped to retain features while minimizing computational cost and alleviating gradient vanishing and gradient explode issues, altogether. The features from the previous layer ($X_0, X_1, \dots, X_{l-1}$) are fused as input to the X layer. The $l - th$ layer with the inherited feature maps from the descendent layers is defined as (34).

$X_l = H_1([X_0, X_1, \dots, X_{l-1}])$	(34)
---	------

In (34), $[X_0, X_1, \dots, X_{l-1}]$ states the cumulative feature map obtained at the 0, 1..., $l - 1$ layers. As depicted in Fig. 4, the inputs of $H_1(\)$ are concatenated to a single tensor generating composite dense feature for further learning and classification. Fig. 4 shows the deployed DenseNet121 model and corresponding unit transition layer connecting each Dense Block. Each of the Dense Block embodies a convolution layer, transition layer interfacing the sub-layers like activation layer, batch normalization and pooling layer. Here, clipped rectified linear unit (cReLU) was applied as an activation layer in which any value lower than the clipping ceiling was assigned as zero. On the contrary, the values higher than the clipping threshold are assigned to the ceiling value. It can help to improve learning over non-linear feature space obtained from a large image. The feature at the end of the model (Fig. 4) (say, $F_{DenseNet121}$) is passed to the average pooling layer (GAP) to be further inputted to the flatten layer to generate learnable feature set.

Fig. 4: DenseNet121 deep architecture



Now, once extracting the DenseNet121 feature ($F_{DenseNet121}$) from the model (Fig. 4), it was horizontally concatenated to yield a composite feature set, which is defined in equation (35).

$Fused_{Feat_Fingerprint} = [F_{GLCM_Fingerprint}; F_{DenseNet121_Fingerprint}]$	(35)
---	------

b. Modality-2: Iris Dataset

As the second image modality, the proposed multi-factor authentication system applies iris images obtained for the same subjects (i.e., users). Recalling that in the proposed work both fingerprint as well as iris images were prepared for each user. Unlike touchless fingerprint images, the spatial (and textural) distribution of the iris images varies, where based on camera location and acquisition patterns the non-iris regions can be of majority class. In other words, in iris images there can be instance-based data imbalance with non-ROI as majority class. Training any machine learning model(s) over such imbalanced features/data can yield inaccuracy or false positive or false negative. To address this challenge, retaining ROI specific features by performing segmentation can be of paramount significance. Extracting ROI-specific features can not only alleviate class-imbalance problem but can reduce redundant computation as well, thus can make overall system more efficient in run-time environment. Unlike traditional segmentation problems with clear edges (of object) the iris regions can be more complex, where the non-linearity can impact overall feature efficiency. To address it, the use of semantic segmentation techniques can be vital. Considering it as motivation, the proposed method applied UNet (universal network) deep model for learning-based semantic segmentation.

1. Pre-processing

Before executing the proposed UNet segmentation method, the standard pre-processing methods including intensity equalization, histogram equalization, resizing and z-score normalization were performed. A brief of the feature engineering methods pertaining to the iris modality is given in the subsequent sections.

For pre-processing at first pixel values were measured by using equation (36).

$p_n = \left(\frac{[u_w(p) - u_w(Min)]}{[u_w(Max) - u_w(Min)]} \right)$	(36)
--	------

In (36), p refers pixel value, where Min and Max indicate the minimum and maximum pixel's values, correspondingly. In (36), u_w be the expected minimum and maximum intensity values, which is measured using equation (37). The other parameters, f_w and σ_w

refer the mean and standard deviation values measured for each window, calculated as per the equation (38) and (39), respectively.

$u_w = \left[1 + \exp\left(\frac{f_w - p}{\sigma_w}\right) \right]^{-1}$	(37)
$f_w = \frac{1}{N^2} \sum_{(i,j) \in w(x,y)} p(i,j)$	(38)
$\sigma_w = \frac{1}{N^2} \sum_{(i,j) \in w(x,y)} (p(i,j) - f_w)^2$	(39)

In (39), N states the size of window, with x and y be the pixel's position on each input window. To retain linearly distributed spatial feature distribution and pixel's homogeneity, the proposed method model applied histogram equalization. Here, MATLAB's function named "*imhist()*" was applied. Subsequently, the images were resized to the pixel dimension of 340×340 . The resized images were processed for z-score normalization, as per the equation (40). To normalize pixel's distribution, the mean (μ) and standard deviation (σ) values were measured for each image or window. With the measured μ and σ values, it intended to maintain average feature distribution to be near zero, while maintaining standard deviation at 1. For an input image i , the normalized z-score value (x) was obtained as (40).

$x^i = \frac{(x - \mu)}{\sigma}$	(40)
----------------------------------	------

The normalized images are mapped as input to the UNet segmentation method for ROI segmentation, which is discussed in the subsequent sections.

2. UNet Segmentation

In this work, UNet semantic segmentation method is applied to locate and segment iris region from input images. Here, UNet deep model was designed with the fully convolutional network that performs pixel-wise learning to locate and segment iris in each input image. In this network, the native contracting network was modified by using layered structure, where the pooling function was substituted with an up-sampling function, which improved image's resolution. In the deployed UNet model, the extracted high-resolution features from the contracting path are merged to the up-sampled outputs (Fig. 5). Here, it makes use of a convolutional layer dedicated to learn the fused features so as to detect patches in each input image. It also applies up-sampling function to retain multiple channels that strengthened the network to pass contextual features or information to the high-resolution layers to improve learning efficiency. As depicted in Fig. 5, the expansion path looks same as the contracting path and thus forms a U-shape structure. It avoided fully connected layers and hence employed the components of each convolution, depicting the segmentation map of the pixels for which the total context is available within the input image. In reference to the spatial non-linearity over input image(s), the proposed method designed UNet with weighted loss (cost) function to detect patches in each input image.

2.1 UNet Architecture

The deployed UNet encompasses two separate paths, called contracting path and expansion paths. These paths are also called left path and right paths, correspondingly. As

depicted in Fig. 5, the left path makes use of convolutional neural network (CNN) architecture, specially designed with two 3×3 convolution filters that function iteratively. As depicted in Fig. 5, the CNN feature extraction layer is connected to the regularization action (ReLU) layer and Max-Pooling layer in sequences. Here, to perform down-sampling, the max-pooling layer was designed with the stride of two. During down sampling, in each stage the number of feature channels were doubled. On the contrary, each step of the expansion path exhibited up-sampling of the feature-map, which was followed by 2×2 up-convolutions. This mechanism reduced the number of feature channels by 50%. The UNet model also applied a concatenation layer to fuse (say, concatenate) the cropped feature map obtained from the contraction path by applying two 3×3 convolutions and single ReLU layer. Finally, an 1×1 convolution layer was applied to map each 64-dimensional feature to the expected number of classes distinguishing the pixel(s) to their corresponding best possible class (i.e., patch or background). In this work, UNet embodied a total of 23 convolutional layers to enhance segmentation accuracy.

2.2 UNet Training

In order to train the network, the input images with their corresponding segmentation maps were given as input to the UNet. Considering feature non-linearity, UNet was designed with ADAM learning function. The batch size was fixed at 8, while initial learning rate and epochs were fixed at 0.0001 and 100, correspondingly. Due to the unpadded convolutions, the proposed design maintained the output images smaller than the input image. The proposed model applied pixel-wise Softmax layer to derive an energy function, as given in equation (41). Here, the Softmax layer functions as the final feature map to perform learning. More specifically, cross-entropy loss function was applied to classify patches (say, region of interest) and non-ROI regions in each input image.

$$p_k(x) = \frac{\exp(a_k(x))}{(\sum_{k'}^k \exp(a_{k'}(x)))} \quad (41)$$

In (41), $a_k(x)$ states the activation function defined for the k -th feature channel at certain pixel position given as $x \in \Omega$ for $\Omega \in \mathbb{Z}^2$. In (41), K be the total number of classes, while $p_k(x)$ states an approximation function. In this work, $p_k(x) \approx 1$ was maintained for k having the highest activation value of $a_k(x)$. In case of other values of k , it retains $p_k(x) \approx 0$. As stated earlier, the UNet model was trained with the help of cross-entropy function (41) that penalizes at each step for the disparities of $p_{l(x)}(x)$ from 1.

$$E = \sum_{x \in \Omega} w(x) \log(p_{l(x)}(x)) \quad (42)$$

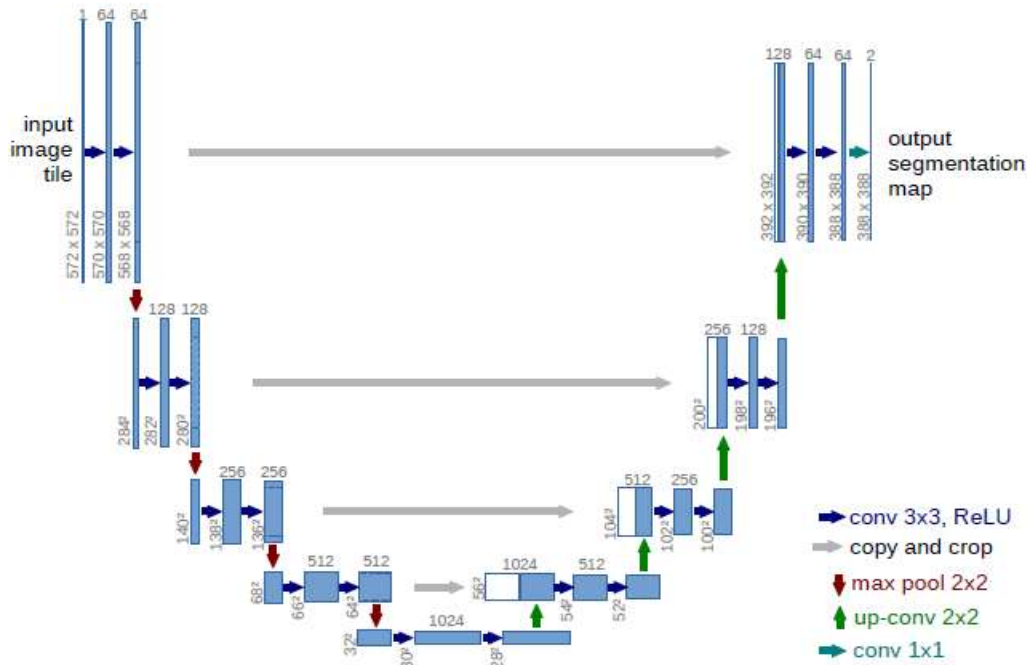
In (42), $l: \Omega \rightarrow \{1, \dots, K\}$ represents pixel's label, while $w: \Omega \rightarrow \mathbb{R}$ states the weight map to measure pixel's significance during training.

The UNet model initially retrieves the weight map for each ground truth input and tries to reward the disparities between the diverse pixel frequencies, depicting a specific class for an input image. It enables UNet to learn minute and fine-grained differences of the edges that eventually enhances segmentation accuracy despite with the severely non-linear and complex edges. To measure the separating border (i.e., ROI and non-ROI), it applies a morphological function calculating the weight map (43).

$w(x)$ $= w_c(x) + w_0 \cdot \exp\left(-\frac{(d_1(x) + d_2(x))^2}{2\sigma^2}\right)$	(43)
---	------

In (43) $w_c: \Omega \rightarrow \mathbb{R}$ represents the weight map that balances the class-frequencies, while $d_1: \Omega \rightarrow \mathbb{R}$ and $d_2: \Omega \rightarrow \mathbb{R}$ be the distance to the border of the nearest cell and the second nearest cell, respectively. For UNet realization, the proposed method assigned $w_0 = 10$ and $\sigma = 5$ pixels values. The UNet model was designed with an alternating convolution followed by a ReLU layer, and hence for a 3×3 convolution, the total number of channels used were 64, while the total incoming nodes for each neuron was $N = 9 * 64 = 576$. Once segmenting patches in each input image, the pixel-wise multiplication was done (between the segmented regions and the original three RGB images) to retain ROI-specific images, which were later processed for feature extraction.

Fig. 5 UNet deep network architecture



3. Residual Deep Spatio-Textural Feature Extraction

Once segmenting the iris regions in each input images, the proposed model executed residual deep spatio-textural feature extraction module in which GLCM and DenseNet121 were applied distinctly to extract features, which were later fused to yield composite feature set. The details of the GLCM and DenseNet121 feature models is given in the previous section. To save space and avoid repeat, their discussion is not given in this section. The GLCM and DenseNet121 features were obtained as F_{GLCM_Iris} and $F_{DenseNet121_Iris}$, correspondingly. These features were later fused to yield composite feature set for the iris images (44).

$Fused_{hybrid_Iris} = [F_{GLCM_Iris}; F_{DenseNet121_Iris}]$	(44)
--	------

4. Multi-Modal Feature Fusion

Once extracting the proposed residual deep spatio-textural features from both fingerprint ($Fused_{hybrid_Fingerprint}$) as well as iris images ($Fused_{hybrid_Iris}$), horizontal concatenation was performed as per the equation (45).

$Fused_{HybridFingIris} = [Fused_{hybrid_Fingerprint}; Fused_{hybrid_iris}]$	(45)
--	------

The fused multimodal features were fed as input to the PCA feature selection method to retain the most significant feature instances for learning and classification.

5. PCA Feature Selection and Normalization

Undeniably, the amalgamation of GLCM and DenseNet121 feature provide sufficiently large information rich feature space with diversities to improve learning and classification efficiency; however, at the cost of increased search space and hence computation. The increased search space with decisive non-linearity can even force machine learning models product false positive or false negative due to unavoidable local minima and convergence. To alleviate it, the proposed method applies PCA feature selection on the extracted composite feature (45). Functionally, the PCA method measured principal component along with the eigen values of the covariance of each feature set. It also calculated the Eigen distance of each feature with reference to the mean of the principal component (here, 0.5). Thus, those data elements or features with the Eigen distance more than the principal component (i.e., 0.5) were dropped, while those elements with smaller distance (<0.5) was retained for further tasks. In this manner, applying PCA feature selection, an optimal set of feature vector was retained for further normalization and classification. Though, the selected ensemble features (45) can have significant non-linearity and hence training any machine learning model over such non-linear feature space can cause over-fitting as well, thus impacting classification reliability. To address it, the proposed method applied z-score normalization on the selected feature set. It estimates the mean (μ) and standard deviation (σ) values (of feature distribution over search space) from the selected feature set. To normalize feature space or values, it intended to maintain mean as 0, while retaining standard deviation as 1. Same as pixel normalization (40), the feature instances were also normalized by performing z-score normalization, where mean of feature instances were intended to maintain near 0, while standard deviation was retained near 1.

6. E2E Ensemble-based Classification

In the past, the authors have applied standalone machine learning classifier to learn and predict best possible class. However, the literatures indicate performance diversity amongst the state-of-art solutions. The lack of generalizability too confines run-time significance of the major state-of-art solutions. Unlike aforesaid existing approaches or standalone classifier-based solutions, the ensemble learning methods like Random Forest (RF) and XGBoost have performed better than the traditional machine learning models like SVM, neural networks. Despite superiority over the traditional machine learning models, the efficacy of ensemble learning algorithms remain debatable. This is because these algorithms often use same type of multiple base classifiers to make consensus or voting-based decisions. To address such limitations, in this work an E2E learning model was developed that embodied five different classifiers including Bagging, AdaBoost, RF, Extra Tree Classifier (ETC) and XGBoost. Being a consensus driven solution, each of the base classifier performs distinct learning and prediction for an input sample and

annotates it with the best possible class to perform person detection or authentication. In this reference, the proposed Agro-Analytics problem was solved as a multi-class classification problem for which the base classifiers predicted each input image with their respective class and annotated respectively. Once annotating each image, the consensus (say, maximum voting by five base classifiers) was measured, and with the highest voting score the final prediction was done. Thus, unlike traditional standalone classifier-based solutions, the use of consensus-driven learning and prediction improved reliability of the analytics and hence made it suitable for the run-time significance. As stated, this research made use of the following ensemble base classifiers.

- Bagging,
- AdaBoost,
- Extra Tree Classifier (ETC),
- Random Forest (RF),
- XGBoost Ensemble.

A snippet of the base ensemble classifiers used in this work is given as follows.

— **Bagging**

The bagging ensemble learning method represents a meta-estimator, particularly designed to fit with the base classifiers such as the Naïve Bayes, k-nearest neighbor (k-NN) on random subsets of the input features. In this work, k-NN algorithm was applied as a base classifier with which it measures its corresponding predictions by applying maximum voting or average voting criteria to predict final class. It is applied in such manner that it minimizes variance of a black-box estimator such as the k-NN or decision tree (DT) algorithm, by performing randomization over tree growth. It constitutes an ensemble to perform consensus-driven learning and classification. In this work, the proposed model applied bagging algorithm as a base classifier to predict and label input image for its best related class or user (for authentication). It annotates each image with the class label (i.e., user's identity), which was later applied by MVE method to perform final prediction decision.

— **AdaBoost**

This method is also called as the adaptive boosting algorithm that has the ability to enhance feature learning ability iteratively by providing suitable reasoning for data classification. In order to make it an adaptive learner, the corresponding prerequisite assessments are given equal weight to detect and classify weak learners which often demands additional preparation. Functionally, the AdaBoost method acts by calculating error rate for the applied weak classifier once completing each computation cycle. The weight of the accurately classified sample is increased by minimizing the weights of the inaccurately clustered samples. Finally, the weak learner evolves into a strong learner that classifies and labels input feature into corresponding best possible class type. Accordingly, it labels each sample with the corresponding class value, which is later applied to perform MVE-based prediction decision.

— **Random Forest**

It is an ensemble learning algorithm that makes use of multiple tree-based classifiers to perform voting-based prediction decision. In comparison to the other machine learning classifiers, the RF algorithm demands relatively smaller number of hyperparameters to perform learning and classification. This ability makes it suitable for a large number of object or data classification problems serving run-time applications. In RF algorithm, each tree-based learner provides unit vote for the specific input data. Applying a number of

trees over the input feature space, it calculates the highest vote (score) for each input image, and thus with the maximum score value it predicts the best possible class for the targeted Agro-Analytics. Consider, N be the total number of images, then a unit image with N cases is arbitrarily selected from the pool of images. The selected images are applied as training sample to form a new tree. For M input images, the optimal split on m images are performed to divide the total number of nodes within tree. In this work, m value is fixed to grow the forest. The equation (46) defines the RF algorithm in the form of the amalgamation of the distinct tree-structures, with h as the classifier and $\{\theta_k\}$ as the random vector spread linearly. Here, the individual trees applied provides single vote depicting the most possible class for an input image x . In (46), θ value depends on its level of significance towards tree development.

$\{h(x, \theta_k), k = 1, 2, \dots i \dots\}$	(46)
---	------

In this work, for RF-based classification a total of 80% samples were used for training, while remaining 20% were employed as the out-of-bag (OOB) samples for inner cross-validation to perform classification.

— *Extra Tree Classifier*

The extra tree classifier (ETC) ensemble learner forms a cluster constituting unpruned decision tree by applying traditional top-down approach. Unlike RF classifier, the ETC algorithm applies attribute's randomization and cut-point selection during node-split. This approach enables it to constitute randomized trees having structures which are not depending on the outputs of the training data. This method distinguishes itself from other tree-based algorithms like DT, RF in two ways. First, it splits nodes by selecting cut-points randomly, while the second factor uses the training data for forest development. The classification results from the trees are used to make final prediction decision by performing consensus driven decision. Noticeably, the key approach behind ETC classifier is its ability to randomize the cut-point and elements altogether by applying ensemble averaging method, which reduces variance more effectively than the weaker randomization techniques, as used by other machine learning classifiers. In addition, the image's use as training samples reduces the probability of bias and thus achieves higher accuracy.

— *XGBoost Ensemble*

The extreme gradient boosting ensemble algorithm, also called (XGBoost) represents a distributed gradient boosted decision tree library, which is often applied for the large data classification. The XGBoost algorithm contributes parallel tree boosting technique, where classification output of each decision tree yields votes to make the final prediction decision. The potential to enhance gradient boosting, especially on non-linear data strengthens XGBoost to meet run-time analytical demands or classification demands, especially when higher accuracy is expected. Consider, the input feature space has n features representing m features $X = \{x_i\}_{i=1}^n, x_i \in R^m$ with labels $y = \{y_i\}_{i=1}^n, y_i \in \{\omega_j \in (1, 2, \dots, c)\}$, where ω_j be the j – th class from C total classes, the ensemble of decision tree using K additive functions is defined as (47).

$\hat{y}_i = \sum_{k=1}^K f_k(x_i), f_k \in F$	(47)
--	------

In (47), $F = \{f(x) = w_q(x)\} (q: R^m \rightarrow T, w \in R^T)$ be the space of the CART, while q and T represent the structure and the total leaves within the tree. Every tree f_k belongs to a structure q and leaf weight w . For an input data, it uses decision rules within the tree structure given as q to classify it into leaves, which is then followed by the measurement of the final prediction output by adding the score in the belonging leaves w . Here, the regularization function (48) is used to learn the set of functions used for the ensemble learning.

$\begin{aligned} \ell &= \sum_{i=1}^n l(\hat{y}_i, y_i) \\ &+ \sum_{k=1}^K \Omega(f_k), \text{ where } \Omega(f_k) \\ &= \xi T + \frac{1}{2} \xi \ w\ ^2 \end{aligned}$	(48)
---	--------

In (48), l be a differential convex loss function, especially used to measure the difference between the predicted value $\hat{y}_i$ and the targeted value y_i . $\Omega(f)$ represents the complexity of tree f_k . The other parameters ξT and $\xi \|w\|^2$ are applied to penalize each tree and maximum weights, respectively. Here, equation (49) comprises parametric functions that can be enhanced by applying traditional enhancement methods like Euclidean space. Despite this, since it is a type of additive training, the objective function for the t iteration in terms of the prediction in $t - 1$ iteration can be updated with the updated tree f_k .

$\begin{aligned} \ell^{(t)} &= \sum_{i=1}^n l(\hat{y}_i^{(t-1)} + f_t(x_i), y_i) \\ &+ \Omega(f_t) \end{aligned}$	(49)
---	--------

Applying Taylor expansion method on (49), particularly with 1st and 2nd -order gradients loss function, a simplified cost function can be derived as (50).

$\begin{aligned} \ell^{(t)} &\cong \sum_{i=1}^n \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] \\ &+ \Omega(f_t) \end{aligned}$	(50)
---	--------

In (50), $g_i = \delta_{\hat{y}_i^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)})$ and $h_i = \delta_{\hat{y}_i^{(t-1)}}^2 l(y_i, \hat{y}_i^{(t-1)})$.

A decision tree algorithm classifies or predicts fixed values in a leaf, hence the tree $f_k(x)$ can be defined as per $w_q(x)$, where w states the score vector for the individual leaf. In this process, $q(x)$ exhibits data mapping of x to a leaf. Now, expanding the 2nd term of (50), a sum over the tree leaves was calculated, where the regularization function outputs the output as (51).

$$\begin{aligned} \ell^{(t)} \cong \sum_{j=1}^T \left[G_j w_i + \frac{1}{2} (H_j + \xi) w_j^2 \right] \\ + \lambda T, \text{ where } G_j \\ = \sum_{i \in I_j} g_i, H_j = \sum_{i \in I_j} h_i \end{aligned} \quad (51)$$

In (51), $I_j = \{i | q(x_i) = j\}$ represents an instance at leaf j . For a definite tree structure, the cost function is reduced as $\frac{\delta \ell^{(t)}}{\delta w_i} = G_j + (H_j + \lambda) w_i = 0$, while the optimal weight of leaf node j is calculated as (52).

$$w^* = -\frac{G_j}{H_j + \xi} \quad (52)$$

Now, replacing the value of (52) in (51), the cost function to calculate the optimal tree structure turns out to be (53).

$$\ell^{(t)} = -\frac{1}{2} \sum_{j=1}^T \frac{G_j^2}{H_j + \xi} + \xi T \quad (53)$$

In this work, equation (53) was applied to examine the split candidates in XGBoost learner, where to achieve the best split, the greedy algorithm and global and local variant approximation methods were applied over possible splits pertaining to the input features. Despite that the local variant approximation technique requires smaller number of candidates in comparison to the global algorithm, the efficacy by the global method remains same as the local method. The proposed method designed distributed gradient boosting-based XGBoost learner and hence it supports the exact greedy method for a unit machine set and both local and global variant approximation methods for all settings. Here, DART booster is used while k trees were dropped during m -th training round. Let, $D = \sum_{k \in K} F_k$ be the leaf scores of the dropped trees and $F_m = \eta F_m$ be the leaf scores of a new tree, the cost function (54) is reframed as (54).

$$\begin{aligned} \ell^{(t)} = \sum_{i=1}^n l(\hat{y}_i^{(m-1)} - D_i + \tilde{F}_m, y_i) \\ + \Omega(\tilde{F}_m) \end{aligned} \quad (54)$$

In (54), D and $\tilde{F}_m$ represents the overshooting factors, were first normalized by using forest-based normalization technique. Thus, applying the above discussed XGBoost method, each input sample is classified with the best probable class. Based on the labelled output, the proposed E2E-MVE ensemble technique was applied to make final decision towards person identification or verification. Thus, applying above discussed E2E-MVE ensemble learning method each input image was classified into best possible category, representing user's identity. The simulation results obtained and allied inferences are discussed in the subsequent section.

V. RESULTS AND DISCUSSION

This research proposed a robust touchless multimodal biometrics driven person identification and verification system. As a touchless multimodality solution, touchless fingerprint and iris images from subjects were taken into consideration, where person verification problem was solved as an image classification problem. Unlike traditional biometrics driven person identification systems, where standalone biometric features are applied to detect or identify an individual, in this research paper a multimodal biometric driven person authentication system or person identification system is proposed. More specifically the proposed method made use of touchless fingerprint and iris images where feature level fusion was performed to train a robust E2E ensemble learner to detect and identify person for PAS purposes. In this work, multimodal biometric images were collected from a total of 200 subjects. Noticeably, there exists no benchmark dataset with iris and touchless fingerprints from the same subject, and therefore the data collected from Kaggle (touchless fingerprint), UCI data repository along with the IITD Iris datasets were processed to prepare or synthesize a set of multimodal biometric datasets. In the proposed multimodal biometrics driven PAS solution, there are the differences in local spatio-textural patterns of the input images (i.e., touchless fingerprint and iris), where touchless fingerprint images can have the impact of local lighting conditions, illumination, textural diversities etc. On the other hand, iris images can also have the non-ROI regions (say, background) having no decisive significance towards feature learning and classification performance. On the contrary, the non-ROI features can not only increase computational costs but can also broaden the horizon for local minima and convergence and over-fitting problems. It could make overall efficacy susceptible and therefore this research focused on performing feature level fusion from both modalities. In this process each of the modality (images) were processed for pre-processing and feature extraction, distinctly. For the proposed touchless fingerprint images, the inputs (images) were at first processed for pre-processing tasks, where different techniques like normalization, local orientation estimation, region masking and Gabor-filtering based smoothening were done in sequence. This method improved local spatial and textural details of each fingerprint images which helped to ensure scalability under diverse operating (and acquisition) conditions. Once processing pre-processing tasks on each input fingerprint images, deep spatio-textural feature extraction was performed by applying GLCM and DenseNet121 feature extraction methods. Here, the prime motive was to extract spatio-textural features with information rich residual feature for learning and classification. Particularly, the proposed method extracted energy, entropy, correlation, homogeneity, contrast, mean, standard deviation, and variance features as GLCM features, which were horizontally concatenated with the DenseNet121 deep features to yield composite feature. To extract deep residual features DenseNet121 was designed with the initial learning rate of 0.001, while the number of epochs were fixed at 200. The DenseNet121 deep learning was executed with ADAM learner to perform adaptive learning. Thus, extracting GLCM and DenseNet121 features horizontal concatenation was done to yield composite deep spatio-textural feature ($Fused_{hybrid_Fingerprint}$) for further learning and classification.

Unlike touchless fingerprint images, iris images were processed for UNet-based deep semantic segmentation to retain ROI specific feature extraction. To execute UNet deep network, the initial learning rate was fixed at 0.0001, while number of epochs were assigned as 100. The deep model applied ADAM non-linear optimization function to cope-up complex spatial feature learning, which helped to achieve accurate ROI segmentation. The segmented ROIs from each iris images (from each subject) were passed to DSTE feature extraction, where GLCM and DenseNet121 methods were applied to extract spatial and residual deep features, correspondingly. The GLCM and DenseNet121 features were fused to yield composite DSTE feature set ($Fused_{hybrid_iris}$). The extracted DSTE

features from both touchless fingerprint and iris were fused together (say, feature level fusion) by performing simple horizontal concatenation. To be noted, as a multimodal DSTE feature set, the features from each subject pertaining to the touchless fingerprint and iris were fused together to yield multimodal feature set for further learning and classification; however, such composite features might be humongous over large input data in run-time environment. Training any machine learning model(s) over such humongous data might impact learning efficiency and hence reliability. More specifically, training machine learning models over exceedingly large non-linear features might give rise to the local minima and convergence issues. To alleviate it, the proposed method multimodal feature ($Fused_{Hybrid_{FingIris}}$) was processed for PCA-based feature selection, followed by z-score normalization. Finally, the normalized feature set ($Fused_{Hybrid_{FingIris}}$) was mapped as input to the proposed E2E ensemble learning model, which was designed by using five base (ensemble) classifiers, encompassing bagging, AdaBoost, RF, ETC and XGBoost classifiers. The proposed method performed MVE ensemble technique to perform classification towards person identification and verification. In order to improve classification reliability 5-fold cross-validation was performed. Functionally, the proposed method performed binary classification as well as multi-class classification where at first it classifies whether the input subject (i.e., person) is recognized or not. Once identifying the user, the proposed method predicts it with corresponding user identity or user name. Thus, the proposed method serves a scalable PAS purpose for run-time significance.

To assess efficiency, confusion matrix was obtained in terms of the true positive (TP), true negative (TN), false positive (FP) and false negative (FN). These parameters were applied to derive performance parameters like accuracy, precision, recall and F-Measure. The mathematical expression for these different performance parameters is given in Table I.

Fig. 3 Confusion matrix

		Actual Class	
		Positive Class	Negative Class
Predicted Class	Positive Class	TP	FP
	Negative Class	FP	TN

Table I: Performance Parameters

Parameter	Mathematical Expression
Accuracy	$\frac{(TN + TP)}{(TN + FN + FP + TP)}$
Precision	$\frac{TP}{(TP + FP)}$

Recall	$\frac{TP}{(TP + FN)}$
F-Measure	$2 \cdot \frac{Recall \cdot Precision}{Recall + Precision}$

The overall proposed model was developed by using MATLAB 2022b software (scripting) tool, while the simulations were done over the central processing unit (CPU) armored with 16 GB Memory, operating over Intel-i5 processor with 3.2 GHz frequency. The simulation results and inferences are discussed in the subsequent sections.

The overall performance analysis was done in two phases, intra-model assessment and inter-model assessment. The intra-model assessment focused on assessing relative efficacy with the different feature models, classifiers, with and without feature selection and segmentation methods. Here, the prime motive was to assess whether the proposed DSTE features with multimodality and E2E learning and classification serves superior performance or not. On the other hand, inter-model assessment focused on comparing the performance of the proposed multimodal touchless biometrics driven PAS solution with the other state-of-art PAS solutions. The detailed discussion of the simulation results and allied inferences is given in the subsequent sections.

A. Intra-Model Assessment

In this method, the simulations were done with the different feature set(s) distinctly over the different ensemble classifiers. The simulation results with and without UNet segmentation and PCA feature selection are obtained to assess whether these feature engineering methods improves the targeted performance.

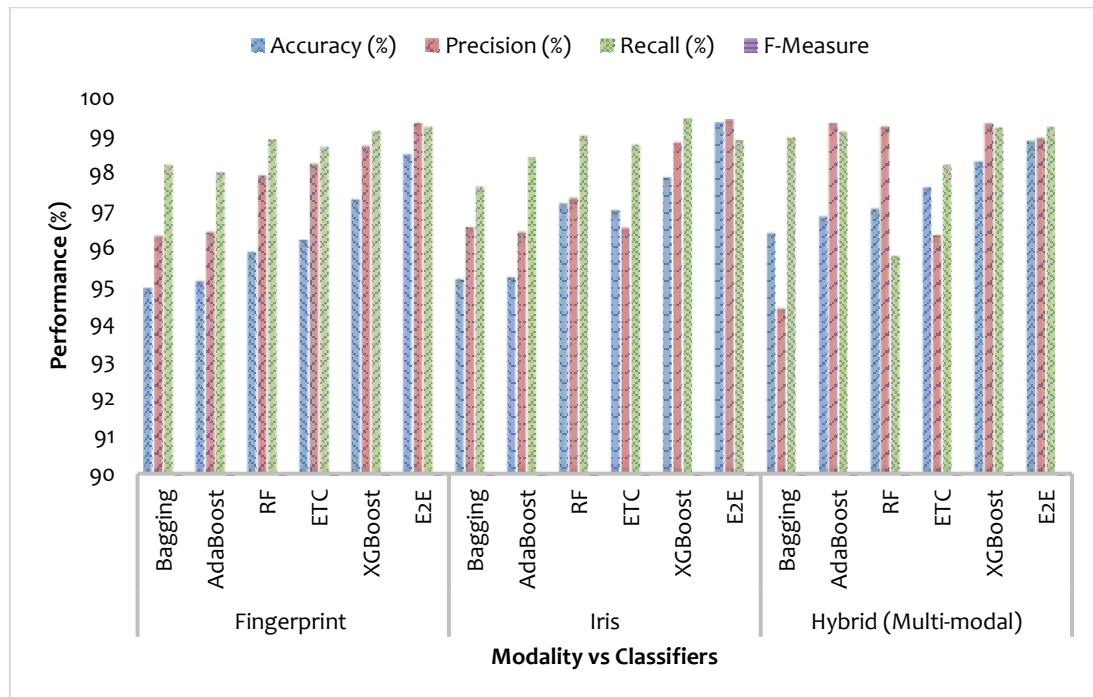
As depicted in Table II, the simulations were done with standalone touchless fingerprint features processed with the PCA feature selection and z-score normalization followed by the different machine learning classifiers for person identification. Similarly, the simulation was done with the iris modality as well, where the different standalone as well as E2E ensemble classifiers were applied to perform person detection and verification (say, identification). As a multimodal biometrics, the composite DSTE features from both fingerprint and iris images were trained with the different classifiers and corresponding performances were obtained. To be noted, in aforesaid simulation cases pre-processing, feature selection and normalization methods were remained common for each simulation. The results (Table II) shows that the touchless fingerprint modality yielded accuracies of 94.92%, 95.11%, 95.88%, 96.21%, 97.29% and 98.49% with the bagging, AdaBoost, RF, ETC and XGBoost classifiers, correspondingly. The F-Measure values too indicate that the fingerprint biometrics with the bagging, AdaBoost, RF, ETC and XGBoost classifiers were measured as 0.97, 0.97, 0.98, 0.98 and 0.98, correspondingly. The simulation results also show that the proposed E2E ensemble learner with the fingerprint modality resulted the superior accuracy (98.49%) and F-Measure (0.99). The results confirm that the proposed E2E ensemble learner shows superior performance over other state-of-art standalone machine learning classifiers. The simulation results also indicate that the iris modality with the bagging, AdaBoost, RF, ETC and XGBoost classifiers showed 95.17%, 95.21%, 97.18%, 96.99% and 97.87%, correspondingly. On the other hand, the iris images resulted the accuracy of 99.33%, while F-Measure too was measured as 0.99. The results confirm that similar to the fingerprint driven biometrics, the E2E ensemble classifier exhibited superior performance over other state-of-art machine learning classifier. The hybrid multimodal biometric solution encompassing touchless fingerprint and iris showed the PAS classification accuracy of 96.39%, 96.82%, 97.03%, 97.61%, 98.29% and 98.29% with the bagging, AdaBoost, RF, ETC, XGBoost and E2E ensemble classifiers, correspondingly. The results clearly indicate that the proposed hybrid

multimodal feature model performs superior PAS prediction accuracy. Unlike standalone feature driven PAS solutions, the accuracy and reliability of the proposed E2E ensemble learning model with the proposed DSTE multi-modal biometrics are superior and can serve run-time PAS purposes.

Table II: Relative Modality Performance

Modality	Classifier	Accuracy (%)	Precision (%)	Recall (%)	F-Measure
Fingerprint	Bagging	94.92	96.31	98.21	0.97
	AdaBoost	95.11	96.40	98.01	0.97
	RF	95.88	97.91	98.89	0.98
	ETC	96.21	98.22	98.68	0.98
	XGBoost	97.29	98.69	99.11	0.98
	E2E	98.49	99.31	99.21	0.99
Iris	Bagging	95.17	96.55	97.63	0.97
	AdaBoost	95.21	96.40	98.40	0.97
	RF	97.18	97.31	98.99	0.98
	ETC	96.99	96.52	98.74	0.97
	XGBoost	97.87	98.79	99.44	0.99
	E2E	99.33	99.40	98.87	0.99
Hybrid (Multimodal)	Bagging	96.39	94.38	98.94	0.97
	AdaBoost	96.82	99.31	99.09	0.99
	RF	97.03	99.21	95.79	0.98
	ETC	97.61	96.33	98.21	0.97
	XGBoost	98.29	99.30	99.20	0.99
	E2E	98.84	98.91	99.21	0.99

Fig. 4: Relative modality performance

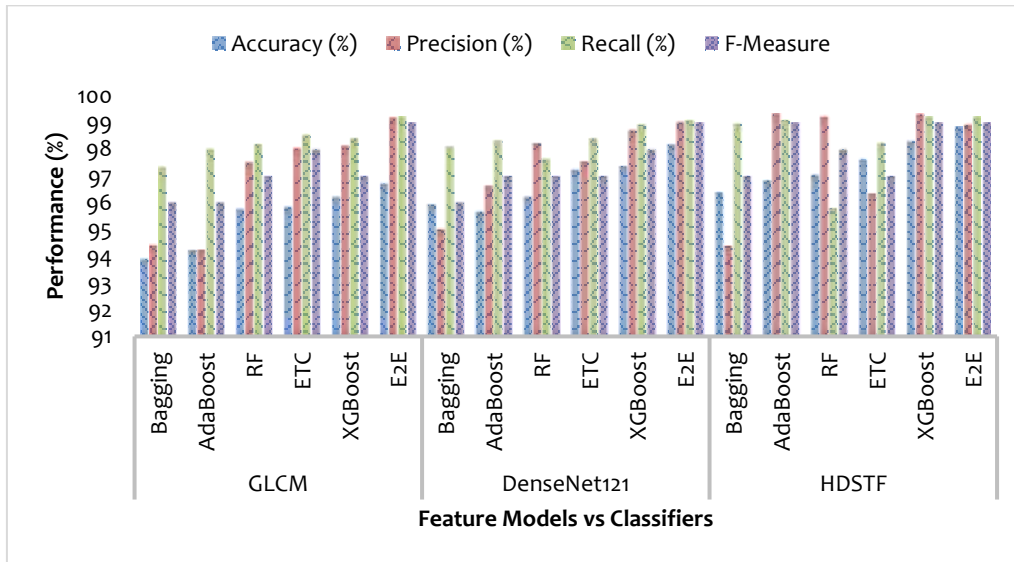


Realizing that the proposed E2E ensemble learning method performs better, the simulations were done with the different feature sets including GLCM, DenseNet121 and the proposed DSTF hybrid or composite (multimodal) feature set. Here, the prime motive was to assess whether the proposed GLCM and DenseNet features with the fingerprint and iris modalities (together) can yield accurate PAS prediction results. The simulation results obtained are given in the Table III. As depicted in Table III, the GLCM feature (it is the feature obtained from both modality, i.e., fingerprint and iris) and DenseNet121 features (feature obtained from both modality, i.e., fingerprint and iris) were trained over the different ensemble classifiers. The simulation results (Table III) show that the GLCM (multimodal) features exhibits accuracies of bagging, AdaBoost, RF, ETC, XGBoost and proposed E2E ensemble classifiers 93.89%, 94.22%, 95.76%, 95.83%, 96.21% and 96.72%, correspondingly. On the other hand, the F-Measure performance too confirms that the proposed E2E ensemble learner exhibits the F-Measure of 0.96, 0.96, 0.97, 0.98, 0.97 and 0.99, respectively with the bagging, AdaBoost, RF, ETC, XGBoost and proposed E2E ensemble classifiers. These results confirm that the proposed GLCM (multimodal) features with the E2E ensemble learner shows better performance than the statealone ensemble classifiers. The DenseNet121 (multimodal) features on the other hand show the PAS prediction accuracies of 95.92%, 95.66%, 96.21%, 97.24%, 97.37%, and 98.17%, with the bagging, AdaBoost, RF, ETC, XGBoost and proposed E2E ensemble classifiers, respectively. Though, the results indicate that the DenseNet121 deep feature shows better PAS prediction accuracy (98.17%) than the GLCM (multimodal) feature model (96.72%). The similar performances were obtained in terms of the precision, recall and F-Measure as well (Table III). The simulations towards the proposed hybrid deep spatio-textural feature ensemble (HDSTF) model indicated that the proposed feature model with the bagging, AdaBoost, RF, ETC, XGBoost and proposed E2E ensemble classifiers showed the average person identification accuracies of 96.39%, 96.82%, 97.03%, 97.61% 98.29% and 98.84%, correspondingly. The similar performances were obtained in terms of precision, recall, and F-Measure as well (Table III). The results indicate that the proposed multimodal HDSTF feature performs superior and hence more reliable PAS prediction to meet run-time scalable application's demands.

Table III: Feature Assessment

Feature Models	Classifier	Accuracy (%)	Precision (%)	Recall (%)	F-Measure
GLCM	Bagging	93.89	94.39	97.31	0.96
	AdaBoost	94.22	94.22	97.99	0.96
	RF	95.76	97.49	98.19	0.97
	ETC	95.83	98.02	98.51	0.98
	XGBoost	96.21	98.11	98.40	0.97
	E2E	96.72	99.17	99.21	0.99
DenseNet121	Bagging	95.92	94.99	98.09	0.96
	AdaBoost	95.66	96.62	98.31	0.97
	RF	96.21	98.21	97.61	0.97
	ETC	97.24	97.53	98.40	0.97
	XGBoost	97.37	98.69	98.90	0.98
	E2E	98.17	98.99	99.09	0.99
HDSTF	Bagging	96.39	94.38	98.94	0.97
	AdaBoost	96.82	99.31	99.09	0.99
	RF	97.03	99.21	95.79	0.98

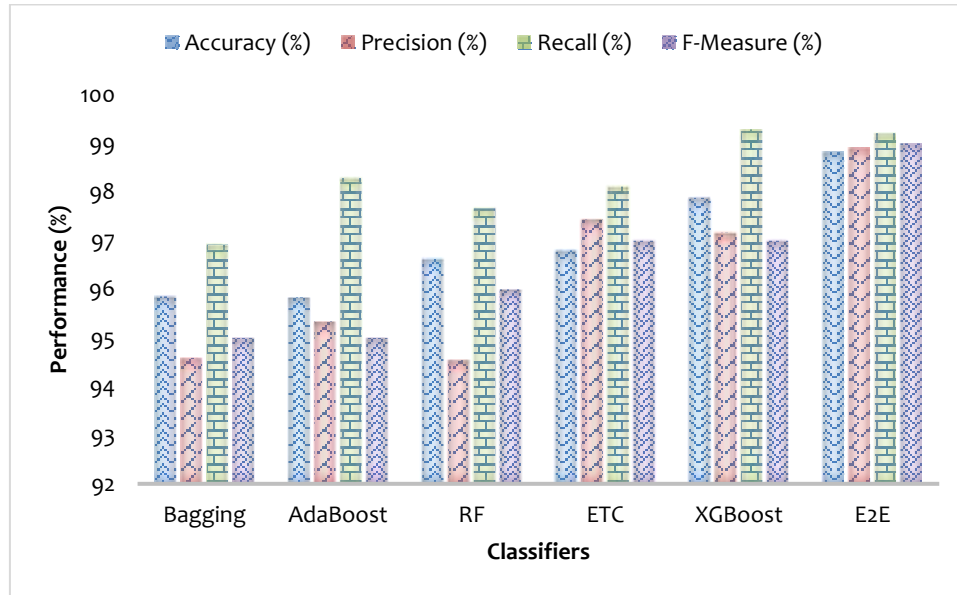
	ETC	97.61	96.33	98.21	0.97
	XGBoost	98.29	99.30	99.20	0.99
	E2E	98.84	98.91	99.21	0.99

Fig. 5: Feature model vs classifier's performance

The above discussions confirm that the HDSTF multimodal feature with the E2E ensemble classifier can yield superior person identification and verification performance. However, to quantify in details whether the proposed E2E ensemble classifier performs effectively, average performances were measured with the different feature models (i.e., GLCM, DenseNet121 and HDSTF). The average performance by the different machine learning classifiers is given in the Table IV. As depicted in Table IV, the bagging classifier exhibited the average prediction accuracy of 95.85%, while AdaBoost, RF, ETC and XGBoost classifiers exhibited the PAS prediction accuracies of 95.83%, 96.62%, 96.80%, 97.88%, respectively. The average performance by the proposed E2E learner was measured as 98.82%, signifying its superiority over other standalone ensemble classifiers.

Table IV: Average classifier performance

Classifier	Accuracy (%)	Precision (%)	Recall (%)	F-Measure
Bagging	95.85	94.59	96.92	0.95
AdaBoost	95.83	95.34	98.29	0.95
RF	96.62	94.54	97.66	0.96
ETC	96.80	97.43	98.11	0.97
XGBoost	97.88	97.16	99.29	0.97
E2E	98.82	98.91	99.21	0.99

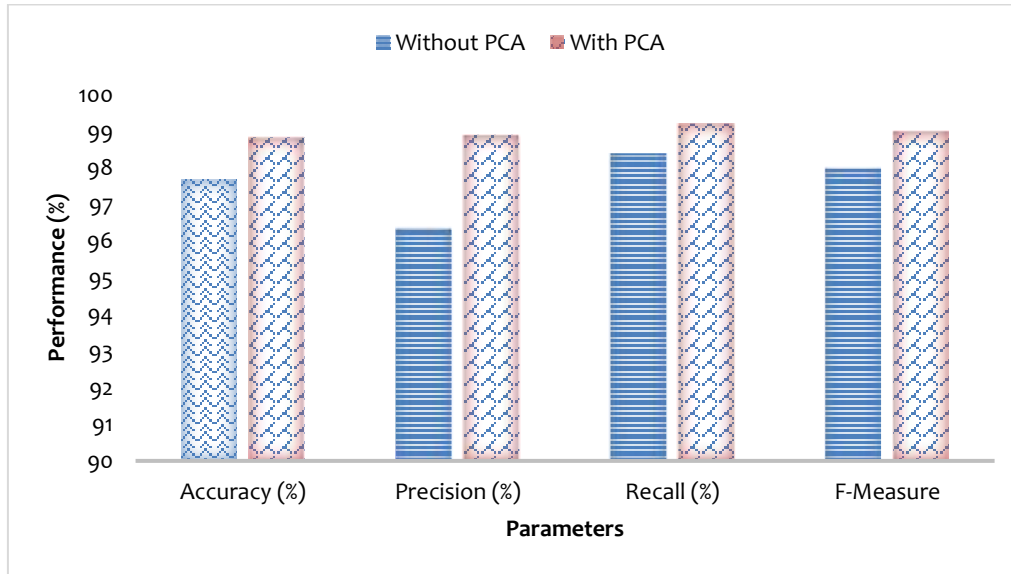
Fig. 6: Average performance by the different classifiers

To assess whether the PCA feature selection method over the proposed HDSTF feature model helped to improve PAS prediction accuracy and hence reliability, the simulations were done with and without PCA selected features. For simulation the proposed method used overall pre-processing tasks, segmentation (as proposed with iris modality), feature normalization and E2E ensemble classification. The HDSTF features with and without feature selection were trained with the proposed E2E ensemble classifier. The results obtained are given in Table V. The results (Table V) shows that without PCA feature selection, the proposed HDSTF feature with E2E ensemble classifier shows the average PAS prediction accuracy of 97.68%. On the contrary, without PCA feature selection it exhibited the average accuracy of 97.68% accuracy. The similar pattern of performances were measured in terms of precision, recall and F-Measure as well. These results clearly indicate that the PCA feature selection method helps to improve the accuracy, though, not much significantly (1.12%). However, its significance to improve model's reliability in run-time environment can't be ruled out.

Table V: Average performance with PCA feature selection

Feature Selection	Accuracy (%)	Precision (%)	Recall (%)	F-Measure
Without PCA	97.68	96.35	98.40	0.98
With PCA	98.84	98.91	99.21	0.99

The overall results confirm that the proposed multimodal HDSTF feature with the allied feature engineering tasks such as the pre-processing, segmentation, PCA feature selection and normalization followed by E2E ensemble classification can exhibit superior performance towards PAS prediction and classification. These results also confirm acceptance of the research questions (RQ1-RQ4). To examine relative performance by the proposed multimodal biometrics driven PAS solution, some of the state-of-art solutions were compared.

Fig. 7: Performance with and without feature selection

The discussion of the results obtained and allied inferences are given in the subsequent section.

B. Inter-Model Assessment

To assess whether the proposed solution performs superior over the existing PAS solutions, the performance of the proposed model was compared with other state-of-arts. Noticeably, in the past almost major biometrics-based PAS solutions are designed with single or standalone. A very few efforts have been made towards multimodal biometrics, especially by amalgamating fingerprint and iris [44][45]. However, to assess relative performance, we have tried to assess whether the use of Fingerprint-Iris modalities altogether can achieve superior results towards PAS verification or not. In this reference, the optimal performance by the proposed model is compared with other state-of-arts. Here onwards, we define efficiency of the proposed model as accuracy 98.84%, precision 98.91%, recall 99.21% and F-Measure 0.99%. To assess whether the use of fingerprint and iris together can yield superior performance we compared the performance with other state-of-arts employing fingerprint and iris biometrics for PAS design. Table VI presents the relative performance outputs by the proposed model and other existing methods.

Table VI Inter-Model Performance Characterization

Ref.	Modality	Performance (%)			
		Accuracy	Precision	Recall	F-Measure
[16]	Fingerprint	96.17	-	-	-
[19]	Fingerprint	99.00	-	-	-
[20]	Fingerprint	99.67	-	-	-
[24]	Fingerprint	96.92	-	-	-
[23]	Fingerprint	98.97	-	-	-
[39]	Fingerprint	86.67	-	-	-
[40]	Fingerprint	96.67	-	-	-

[41]	Fingerprint	96.14	-	-	-
[42]	Fingerprint	94.75	-	-	-
[18]	Iris	99.95	-	-	-
[20]	Iris	99.10	-	-	-
[22]	Iris	94.74	-	-	-
[25][28]	Iris	98.00	-	-	-
[29]	Iris	98.20	-	-	-
[32]	Iris	88.94	-	-	-
[34]	Iris	97.70	-	-	-
[36]	Iris+ Fingerprint +Face	92.00	-	-	-
[43]	Iris	96	-	-	-
[44]	Face-finger	70.00	0.25	1.00	
	Face-Iris	80.00	0.33	1.00	
	Finger-Iris	90.00	0.50	1.00	
[55]	Iris+Palmprint+face	99.80	-	-	-
Proposed	Finger-Iris	98.84	98.91	99.21	0.99

Table VI presents some of the recent methods made towards (physiological) biometrics-based PAS solution. Observing the results, it can clearly be found that the despite different feature methods and allied machine learning optimizations, the state-of-arts [16][19][20][23][24] [46-49], Iris [26][28][30][33][36][37][40][42] still lags behind the proposed ROI-specific DST feature driven multimodal biometric based PAS verification system. To cope up with enterprise demands, though, a few approaches including [46-49] improved data reconstruction, three-dimensional mapping and multi-view feature learning to achieve high accuracy, especially in touchless operating conditions. However, the proposed model performs superior (accuracy 98.84%) than them. Here, the role of ROI-specific feature learning and more importantly deep spatio-textural feature learning 9DST encompassing GLCM and ResNet101) can't be ruled out in achieving such significant performance improvement. The PAS solutions applying Iris data as biometric input too could exhibit highest accuracy of 96%, which is significantly lower than the proposed method (98.84%). This is the matter of fact that there exists lack in sufficient researches, especially towards multimodal biometrics. However, a few recent efforts were identified which used visual cues like iris, fingerprint and face as input for PAS [50][51]. Interestingly, the authors in [51] assessed the different biometric combinations including face and iris, finger and iris and face and finger to perform multimodal biometric or verification tasks. Noticeably, the amalgamation of face and finger [51] resulted verification accuracy of 70%, precision and recall of 25% and 100%, respectively. On the contrary, for face and iris modalities, it exhibited the verification accuracy of 80%, while for finger and iris based multimodal PAS solution it showed the accuracy of 90%, precision 50% and recall 100%. Irrespective of their algorithmic contributions and novelties, it can be inferred that the combination of fingerprint and iris can result superior over other multimodal biometrics such as face and finger and face and iris. In this reference, considering it as secondary reveal the hypothesis H01 stating that the combination of fingerprint and iris biometrics can yield superior PAS efficiency is found justified or confirmed [52]. Quantitatively, the simulation results (Table V) confirm that the use of fingerprint and iris can yield superior accuracy or performance (i.e., accuracy 98.84%) other methods [51]. Unlike [51], we introduced numerous computational as well as feature optimization measures including the FFCM driven ROI segmentation, ROI-specific deep spatio-textural feature extraction (GLCM + ResNet101) and RF enabled

multi-class classification that consequently has improved performance greatly and confirms superiority of the proposed model over other state-of-arts (Table VI). To generalize efficacy of the proposed multimodal PAS solution, it was compared with a recent state-of-art method [55]. Rachapalli et al. [55] proposed a multimodal cancellable biometric template protection and person verification solution in transformed domain. Realizing that the inappropriate subject data and compromised keys can make PAS vulnerable, they proposed a multimodal PAS that extracts biometric features from person's face, iris, and palmprint. Amalgamating biometric features involved a colorization mechanism that generated three distinct colors Quick Response (QR) codes from a user-defined Red, Green, Blue (RGB) tuple random seed, which was later fused to yield a one-way cancellable template. Though, the authors claimed to have achieved accuracy of 99.84%. Their method yielded the AUC of 0.97. Noticeably, the proposed method showed the average personal authentication accuracy of 99.74%, which is almost same as that of [55]. However, unlike state-of-art [55], which made use of the three biometric modalities including face, iris and palmprint, our proposed method applied merely Iris and touchless fingerprint whose computational cost is lower and hence can be more computationally efficient.

VI. CONCLUSION

This paper proposed a robust deep segmentation assisted residual spatio-textural ensemble feature learning framework for multi-modal person identification and verification. Unlike traditional touch-based biometrics, this research proposes a robust touchless multimodal person authentication system embodying Iris and touchless fingerprint images. Here, the prime motive was to contribute a robust touchless PAS solution with the highest level of imitation resilience and reliability. In this reference, the proposed method applied the different optimization measures including pre-processing, feature extraction, feature selection and classification. To achieve it, the proposed method at first applied a robust information sensitive pre-processing tasks including image normalization, local orientation estimation, region masking and Gabor-filtering based smoothening on Iris images, which was later processed for GLCM and DenseNet121 feature extraction. Here, GLCM method served information rich feature (i.e., energy, entropy, correlation, homogeneity, contrast, mean, standard deviation, and variance), while DenseNet121 provided residual feature set. The extracted GLCM and DenseNet121 features are fused together to yield a composite feature set for further feature level fusion with the Iris features. Unlike touchless fingerprint modality, the Iris images used to be more complex and can have the non-region-of-interest (non-ROI) details as well based on the camera acquisitions. Realizing that extracting features and learning them from the non-ROI regions can impact overall prediction results (ex. False positive/false negative), the proposed method at first executes semantic segmentation that retains ROI-specific images for further feature extraction and learning. To achieve it, the proposed method employed UNet deep segmentation method that retained ROI-specific (iris) details from each user or subject, which were subsequently processed for GLCM and DenseNet121 feature extraction. Once extracting GLCM and DenseNet121 features, horizontal concatenation was done to yield deep spatio-textural feature (DSTF) set. The DSTF features from both fingerprint and iris were fused to yield composite feature set, which was later processed for PCA feature selection and z-score normalization that helped to alleviate local minima and convergence and over-fitting issues. Later, the normalized features were processed for ensemble-of-ensemble (E2E) learner that classified each user as identified or non-identified. Additionally, it also classified an input subject or image with the respective accuracy class (say, name) identifying the exact subject to meet run-time verification. Unlike traditional standalone machine learning driven solutions, the use

of E2E MVE learner embodying bagging, AdaBoost, Random Forest, Extra Tree Classifier and XGBoost ensemble classifiers as base classifier performed person identification and classification accuracy of 98.84%, precision 98.91%, recall 99.21% and F-Measure 0.99, which is higher than the other state-of-art solutions. The results confirmed superiority, robustness and run-time efficacy of the proposed PAS solution to meet real-world application demands.

REFERENCE

- [1] M. Gomez-Barrero et al., "Biometrics in the Era of COVID-19: Challenges and Opportunities," in *IEEE Transactions on Technology and Society*, vol. 3, no. 4, pp. 307-322, Dec. 2022.
- [2] A. K. Jain, D. Deb and J. J. Engelsma, "Biometrics: Trust, But Verify," in *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 4, no. 3, pp. 303-323, July 2022.
- [3] S. Soleymani, A. Dabouei, F. Taherkhani, S. M. Iranmanesh, J. Dawson and N. M. Nasrabadi, "Quality-Aware Multimodal Biometric Recognition," in *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 4, no. 1, pp. 97-116, Jan. 2022.
- [4] A. F. Abate, L. Cimmino, I. Cuomo, M. D. Nardo and T. Murino, "On the Impact of Multimodal and Multisensor Biometrics in Smart Factories," in *IEEE Transactions on Industrial Informatics*, vol. 18, no. 12, pp. 9092-9100, Dec. 2022.
- [5] S. Li, B. Zhang, L. Fei, S. Zhao and Y. Zhou, "Learning Sparse and Discriminative Multimodal Feature Codes for Finger Recognition," in *IEEE Transactions on Multimedia*, vol. 25, pp. 805-815, 2023.
- [6] T. Ahmed, S. Samima, M. Zuhair, H. Ghayvat, M. A. Khan and N. Kumar, "FIMBISAE: A Multimodal Biometric Secured Data Access Framework for Internet of Medical Things Ecosystem," in *IEEE Internet of Things Journal*, vol. 10, no. 7, pp. 6259-6270, 1 April, 2023.
- [7] R. Ryu, S. Yeom, S. -H. Kim and D. Herbert, "Continuous Multimodal Biometric Authentication Schemes: A Systematic Review," in *IEEE Access*, vol. 9, pp. 34541-34557, 2021.
- [8] C. Yuan, S. Jiao, X. Sun and Q. M. J. Wu, "MFFFLD: A Multimodal-Feature-Fusion-Based Fingerprint Liveness Detection," in *IEEE Trans. on Cognitive and Developmental Sys.* vol. 14, no. 2, pp. 648-661, June 2022.
- [9] I. Goel, N. B. Puhan and B. Mandal, "Deep Convolutional Neural Network for Double-Identity Fingerprint Detection," in *IEEE Sensors Letters*, vol. 4, no. 5, pp. 1-4, May 2020, Art no. 7001404.
- [10] V. Anand and V. Kanhangad, "Cross-Sensor Pore Detection in High-Resolution Fingerprint Images," in *IEEE Sensors Journal*, vol. 22, no. 1, pp. 555-564, 1 Jan.1, 2022.
- [11] Y. Feng and A. Kumar, "Detecting Locally, Patching Globally: An End-to-End Framework for High Speed and Accurate Detection of Fingerprint Minutiae," in *IEEE Transactions on Information Forensics and Security*, vol. 18, pp. 1720-1733, 2023.
- [12] Y. Zhang, S. Pan, X. Zhan, Z. Li, M. Gao and C. Gao, "FLDNet: Light Dense CNN for Fingerprint Liveness Detection," in *IEEE Access*, vol. 8, pp. 84141-84152, 2020.
- [13] M. Ali, C. Wang and M. O. Ahmad, "An Efficient Convolutional Neural Network for Fingerprint Pore Detection," in *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 3(3), pp. 332-46, 2021.
- [14] C. Yuan, S. Jiao, X. Sun and Q. M. J. Wu, "MFFFLD: A Multimodal-Feature-Fusion-Based Fingerprint Liveness Detection," in *IEEE Trans. on Cognitive and Developmental Systems*, vol. 14(2), pp. 648-661, June 2022.

- [15] G. Bhattacharya and N. B. Puan, "Stand-Alone Multi-Attention Fusion Network for Double-Identity Fingerprint Detection," in *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 4, no. 4, pp. 596-602, Oct. 2022.
- [16] L. J. González-Soler, M. Gomez-Barrero, L. Chang, A. Pérez-Suárez and C. Busch, "Fingerprint Presentation Attack Detection Based on Local Features Encoding for Unknown Attacks," in *IEEE Access*, vol. 9, pp. 5806-20, 2021.
- [17] L. Spinoulas, H. Mirzaalian, M. E. Hussein and W. AbdAlmageed, "Multi-Modal Fingerprint Presentation Attack Detection: Evaluation on a New Dataset," in *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 3, no. 3, pp. 347-364, July 2021.
- [18] C. Yuan, Z. Xia, X. Sun and Q. M. J. Wu, "Deep Residual Network with Adaptive Learning Framework for Fingerprint Liveness Detection," in *IEEE Transactions on Cognitive and Developmental Systems*, vol. 12, no. 3, pp. 461-473, Sept. 2020.
- [19] P. Ji, J. Feng, F. Ma, X. Wang and C. Li, "Fingertip Detection Algorithm Based on Maximum Discrimination HOG Feature in Complex Background," in *IEEE Access*, vol. 11, pp. 3160-3173, 2023.
- [20] C. -W. Hung, J. -R. Wu and C. -H. Lee, "Device Light Fingerprints Identification Using MCU-Based Deep Learning Approach," in *IEEE Access*, vol. 9, pp. 168134-168140, 2021.
- [21] W. Lei and Y. Lin, "A Novel Dynamic Fingerprint Segmentation Method Based on Fuzzy C-Means and Genetic Algorithm," in *IEEE Access*, vol. 8, pp. 132694-132702, 2020.
- [22] Y. Xun, J. Liu, N. Kato, Y. Fang and Y. Zhang, "Automobile Driver Fingerprinting: A New Machine Learning Based Authentication Scheme," in *IEEE Trans. on Indus. Informatics*, vol. 16(2), pp. 1417-1426, Feb. 2020.
- [23] J. Galbally, L. Beslay and G. Böstrom, "3D-FLARE: A Touchless Full-3D Fingerprint Recognition System Based on Laser Sensing," in *IEEE Access*, vol. 8, pp. 145513-145534, 2020.
- [24] A. Bera and D. Bhattacharjee, "Human Identification Using Selected Features from Finger Geometric Profiles," in *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 50, no. 3, pp. 747-761, March 2020.
- [25] Z. Fang, A. Czajka and K. W. Bowyer, "Robust Iris Presentation Attack Detection Fusing 2D and 3D Information," in *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 510-520, 2021.
- [26] S. Khade, S. Gite, S. D. Thepade, B. Pradhan and A. Alamri, "Detection of Iris Presentation Attacks Using Hybridization of Discrete Cosine Transform and Haar Transform with Machine Learning Classifiers and Ensembles," in *IEEE Access*, vol. 9, pp. 169231-169249, 2021.
- [27] M. Choudhary, V. Tiwari and U. Venkanna, "Iris Liveness Detection Using Fusion of Domain-Specific Multiple BSIF and DenseNet Features," in *IEEE Trans. on Cybernetics*, vol. 52, no. 4, pp. 2370-2381, April 2022.
- [28] J. E. Tapia, S. Gonzalez and C. Busch, "Iris Liveness Detection Using a Cascade of Dedicated Deep Learning Networks," in *IEEE Transactions on Information Forensics and Security*, vol. 17, pp. 42-52, 2022.
- [29] A. Agarwal, A. Noore, M. Vatsa and R. Singh, "Generalized Contact Lens Iris Presentation Attack Detection," in *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 4, no. 3, pp. 373-385, July 2022.
- [30] M. Yildiz and M. Yorulmaz, "A Novel Gaze Input System Based on Iris Tracking with Webcam Mounted Eyeglasses," in *Interacting with Computers*, vol. 33, no. 2, pp. 211-222, 2021.
- [31] D. M. Daniel and B. Monica, "Person authentication technique using human iris recognition," 2010 9th International Symposium on Electronics and Telecommunications, Timisoara, Romania, 2010, pp. 265-268.

- [32] M. Choudhary, V. Tiwari and V. U, "Ensuring Secured Iris Authentication for Mobile Devices," 2021 IEEE International Conference on Consumer Electronics (ICCE), Las Vegas, NV, USA, 2021, pp. 1-5.
- [33] C. -S. Hsiao, C. -P. Fan and Y. -T. Hwang, "Iris Location and Recognition by Deep-Learning Networks Based Design for Biometric Authorization," 2021 IEEE 3rd Global Conference on Life Sciences and Technologies (LifeTech), Nara, Japan, 2021, pp. 144-145.
- [34] T. Sueishi, A. Jingu, S. Yachida, M. Inoue, Y. Ogino and M. Ishikawa, "Dynamic Iris Authentication by High-speed Gaze and Focus Control," 2021 IEEE/SICE International Symposium on System Integration (SII), Iwaki, Fukushima, Japan, 2021, pp. 813-814.
- [35] N. Prasanth, C. U. Sai Kiran, S. Nethi, T. Ketineni, N. Srinivasu and G. Pradeepini, "Fusion of Iris and Periocular Biometrics Authentication using CNN," 2023 7th International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2023, pp. 378-382.
- [36] C. -S. Hsiao, C. -P. Fan and Y. -T. Hwang, "Design and Analysis of Deep-Learning Based Iris Recognition Technologies by Combination of U-Net and EfficientNet," 2021 9th International Conference on Information and Education Technology (ICIET), Okayama, Japan, 2021, pp. 433-437.
- [37] C. -S. Hsiao and C. -P. Fan, "EfficientNet Based Iris Biometric Recognition Methods with Pupil Positioning by U-Net," 2021 3rd International Conference on Computer Communication and the Internet (ICCCI), Nagoya, Japan, 2021, pp. 1-5.
- [38] E. A. W. Hachim, M. T. Gaata and T. Abbas, "Iris-based Authentication Model in Cloud Environment (IAMCE)," 2022 Intl. Conf. on Elec, Comp. and Energy Tech. (ICECET), Prague, Czech Republic, 2022, pp. 1-6.
- [39] O. Ouda, "On the Practicality of Local Ranking-Based Cancelable Iris Recognition," in IEEE Access, vol. 9, pp. 86392-86403, 2021.
- [40] A. Jardine, T. D. Ramotsoela and G. P. Hancke, "Biometric Authentication System for Industrial Applications using Iris Recognition," 2021 IEEE 30th Intl. Symp. on Industrial Electronics (ISIE), Kyoto, Japan, 2021, pp. 01-06.
- [41] J. Chen, Y. Cui, F. Shen, J. Shen and T. Wei, "DBANet: A Dual Branch Attention-Based Deep Neural Network for Biological Iris Recognition," 2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Las Vegas, NV, USA, 2022, pp. 1497-1500.
- [42] R. W. Jalal and M. F. Ghanim, "Enhancement of Iris Recognition System using Deep learning," 2022 IEEE Symposium on Industrial Electronics & Applications (ISIEA), Langkawi Island, Malaysia, 2022, pp. 1-7.
- [43] S. M. R. Bagwan, G. Gupta and S. B.Thigale, "Robust Multi-Bio-Metric Authentication Framework in Face and Iris recognition," 2023 2nd Int. Conf. for Inno. in Tech. (INOCON), Bangalore, India, 2023, pp. 1-10.
- [44] B. E, S. M, V. P, V. R and V. K. A, "Fusion of Iris, Face, Fingerprint using Score Level Mechanism for Biometric Application," 2023 International Conference on Innovative Data Communication Technologies and Application (ICIDCA), Uttarakhand, India, 2023, pp. 265-270.
- [45] Z. Chang, X. Zhang and J. Zhang, "PCA Based Cancelable Biometric Protection Using Feature-Level Fusion of Iris and Fingerprint," 2022 4th Intl. Conf. on Natural Language Processing, Xi'an, China, 2022, pp. 74-80.
- [46] R. Vyas and A. Kumar, "A Collaborative Approach using Ridge-Valley Minutiae for More Accurate Contactless Fingerprint Identification", Technical Report No.: COMP-K-25, 2018, pp. 1-15.
- [47] A. Kumar and C. Kwong, "Towards Contactless, Low-Cost and Accurate 3D Fingerprint Identification", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 37, No. 3, March 2015, pp. 681-696.

- [48] C. Lin and A. Kumar, "Contactless and Partial 3D Fingerprint Recognition using multi-view Deep Representation", *Pattern Recognition* (2018), DOI: 10.1016/j.patcog.2018.05.004.
- [49] Q. Zheng, A. Kumar, and G. Pan, "Contactless 3D Fingerprint Identification Without 3D Reconstruction", Department of Comp. Sci. and Tech. Zhejiang Univ, Hangzhou, Zhejiang, China, [Accessed on 21 April 2022].
- [50] M. H. Hamd and S. K. Ahmed, "Biometric System Design for Iris Recognition Using Intelligent Algorithms", *IJ. Modern Education and Computer Science*, 2018, 3, 9-16.
- [51] N. Damodaram and D. V. S. Reddy, "Multi-Modal Biometric System using Iris, Face and Fingerprint Images for High-Security Application", *International Journal of Engineering Research & Technology (IJERT)*, Volume 7, Issue 11, pp. 1-6.
- [52] O. Dere, S. Gaikwad, "Biometric Accreditation Adoption using Iris and Fingerprint: A Review", *International Research Journal of Engineering and Technology (IRJET)*, Volume: 08 Issue: 07, 2021, pp. 4551-56.
- [53] <https://www.kaggle.com/datasets/naureenmohammad/mmu-iris-dataset>.
- [54] <https://www.kaggle.com/datasets/peace1019/fingerprint-dataset-for-fvc2000-db4-b>.
- [55] D. Reddy Rachapalli, V. Dondeti and H. K. Kalluri, "Multimodal Cancellable Biometric Template Protection and Person Verification in Transformed Domain," in *IEEE Access*, vol. 12, pp. 173557-173582, 2024.
- [56] https://www4.comp.polyu.edu.hk/~csajaykr/IITD/Database_Iris.htm