2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC) | 979-8-3503-2447-1/23/$31.00 ©2023 IEEE | DOI: 10.1109/EMBC40787.2023.10340596

Feature Learning Networks for Floor Sensor-based Gait Recognition

Ala Salehi, Alex Roberts, Angkoon Phinyomark and Erik Scheme

***Abstract*— Deep learning (DL) has become a powerful tool in many image classification applications but often requires large training sets to achieve high accuracy. For applications where the available data are limited, this can become a severely limiting factor in model performance. To address this limitation, feature learning network approaches that integrate traditional feature extraction methods with DL frameworks have been proposed. In this study, the performances of traditional methods: discrete wavelet transform (DWT), discrete cosine transform (DCT), independent component analysis (ICA), and principal component analysis (PCA); and their corresponding feature networks based on a convolutional neural network (CNN) framework: ScatNet (wavelet scattering network), DCTNet, ICANet, and PCANet, were investigated for use in pressure-based footstep recognition when the limited sample size is available for person authentication. The results show that the feature learning networks (90.6% accuracy) achieved significantly better performance on average than the conventional feature extraction methods (79.7% accuracy) ($p < 0.05$). Among the different feature networks, PCANet provided the best verification performance, with an accuracy of 92.2%. Feature learning networks are simple and effective approaches that can be a promising solution for applications like floor-based gait recognition in a security access scenario (such as workspace environment and border control) when small amounts of data are available for training models to differentiate between a larger group of users.**

I. INTRODUCTION

Floor sensor-based gait recognition (or footstep recognition) employs sensors to measure pressure patterns (distribution and amplitude) of a person's feet during walking. The collected data are normally processed through different feature extraction methods that employ either machine learning (ML) or deep learning (DL) algorithms. These features are used to identify the unique characteristics of walking patterns (or "gait signature") for each person. The success of gait recognition is thus dependent on the selection of features that best represent foot pressure patterns.

The convolutional neural network (CNN/ConvNet) has become one of the most popular DL algorithms due to its strong classification performance across a range of tasks. Several CNN variations have been developed to analyze gait data and identify gait characteristics in biometric authentication systems based on underfoot pressure [1], [2]. In a study by Costilla-Reyes *et al.* [1], however, performance was found to degrade substantially (i.e., the equal error rate (EER) increased from 2.1% to 10.7%) when the number of training samples (stride footsteps) per user was reduced from 500 to 40 (excluding additional validation samples). Even with the largest training data size, comparable results were found between the deep residual neural network (ResNet) and conventional ML methods like a support vector machine (SVM) (EER of 2.1% vs. 2.6%). This may be partly due to there being insufficient training data for CNNs. [1]. Importantly, the collection of even 500 stride footsteps per user may not be feasible for real-world scenarios (e.g., access control at the airports or in workplace environments).

To improve the performance of recognition systems when the training sample size is limited, feature learning networks have been proposed that integrate traditional feature extraction methods within the CNN framework [3]–[6]. By leveraging more structured approaches, these feature learning networks have less parameters to optimize, and may thus require less training samples. Surprisingly, these lightweight networks have achieved state-of-the-art results and even yielded better performance than carefully learned networks for many classification tasks.

Some widely-known feature learning networks include wavelet scattering networks (ScatNet) [3], discrete cosine transform (DCT) network (DCTNet) [4], independent component analysis (ICA) network (ICANet) [5], and principal component analysis (PCA) network (PCANet) [6]. The first two networks: ScatNet and DCTNet are learning-free (data-independent) approaches as the convolutional filters are prefixed. ICANet and PCANet are feature learning-based approaches wherein ICA and PCA are used, respectively, to create a data-informed convolution filter bank in each stage. Given the potential of feature extraction methods like the discrete wavelet transform (DWT), ICA, and PCA having been demonstrated for person authentication in previous studies [7], [8], feature network approaches that leverage the benefits of cascaded feature learning and extraction architectures [6] should be able to generate better discriminative features from foot pressure images. To the best of our knowledge, an investigation of these feature networks for floor-based gait recognition has not previously been reported.

In this study, two aims are explored: (1) to evaluate the performance of feature extraction methods (DWT, DCT, ICA, and PCA) and feature learning networks based on the CNN framework (ScatNet, DCTNet, ICANet, and PCANet) for person verification using foot pressure-based gait; and (2) to investigate further the effectiveness of feature learning networks with a limited sample size. The ultimate goal of

*This project was generously supported by the New Brunswick Innovation Foundation's Strategic Opportunities Fund, the Atlantic Canada Opportunities Agency's Regional Innovation Ecosystem program, and the Natural Sciences and Engineering Research Council of Canada's Alliance Grants program [ALLRP 558340-20].

Ala Salehi, Alex Roberts, Angkoon Phinyomark, and Erik Scheme are with the Institute of Biomedical Engineering, University of New Brunswick, Fredericton, NB, Canada. ala.salehi@unb.ca, arobert6@unb.ca, aphinyom@unb.ca, escheme@unb.ca

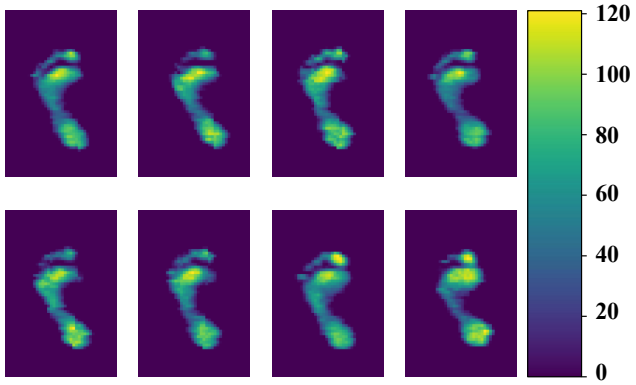


Fig. 1: Samples of the 2D pre-feature peak pressure images (blue-to-yellow: low-to-high pressure levels/values (N/cm^2)) with a width and height of 40 and 60 pixels, respectively.

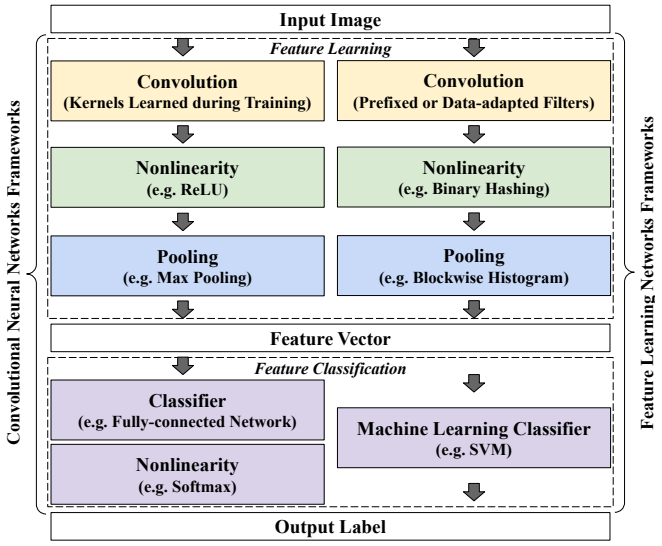


Fig. 2: Comparison of CNN and feature learning network frameworks.

the present study is to determine a potential DL framework for floor sensor-based gait recognition in a security access scenario, i.e., when a small amount of data is available per user for model training for a larger group of users.

II. Methods

A. Foot Pressure Data and Pre-feature Images

The performance of person verification models using different feature extraction approaches was investigated using the CASIA-D dataset [9]. These foot pressure data were recorded using the RSScan Footscan device with 255-by-64 sensors. Each footstep sample is a three-dimensional (3D) matrix consisting of 100 two-dimensional (2D) matrices of a foot pressure map. Each $60 \times 40 \times 100$ footstep sample was converted to a 60×40 pre-feature image by computing peak or maximum pressure from each pixel time series [10] (Fig. 1). Ten trials of barefoot walking were collected for each subject, with three footsteps per trial. For a total of 97 subjects, there are 2,658 2D pre-feature images (the left footsteps were flipped to make them more similar to the right footsteps) for use in the next operation.

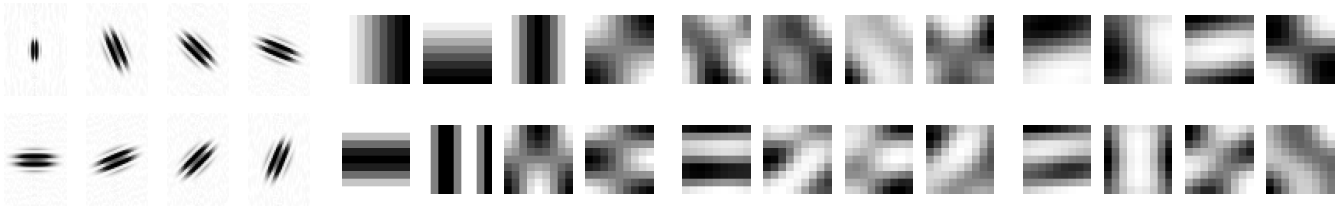
B. Feature Learning Networks

Convolutional neural network (CNN) is a network architecture for DL that consists of three types of layers: convolutional layers, pooling layers, and fully-connected layers. Comparing the CNN architecture to the common framework of pattern recognition systems, feature extraction is performed using the convolutional layers. Dimensionality reduction is achieved via pooling layers which are implemented to reduce the dimensions of the learned feature set (or feature maps). A fully connected layer then maps the extracted features to the final output, i.e., classification.

The first, and maybe the most important, building block of the CNN architecture is the convolutional layers which consist of a set of kernels (or filters). In traditional CNNs, the parameters of these kernels are carefully learned throughout the training using, for instance, back-propagation and gradient descent. Because CNNs must learn millions of parameters, properly learning kernel parameters can be difficult for small datasets. Feature learning networks, on the contrary, replace the deep learned kernels in the convolutional layers with conventional feature extraction methods (Fig. 2). These methods can use either pre-fixed (learning-free) filters like wavelet transform for ScatNet [3] and 2D DCT for DCTNet [4] or the data-informed (learning-based) filters like ICA and PCA for ICANet [5] and PCANet [6].

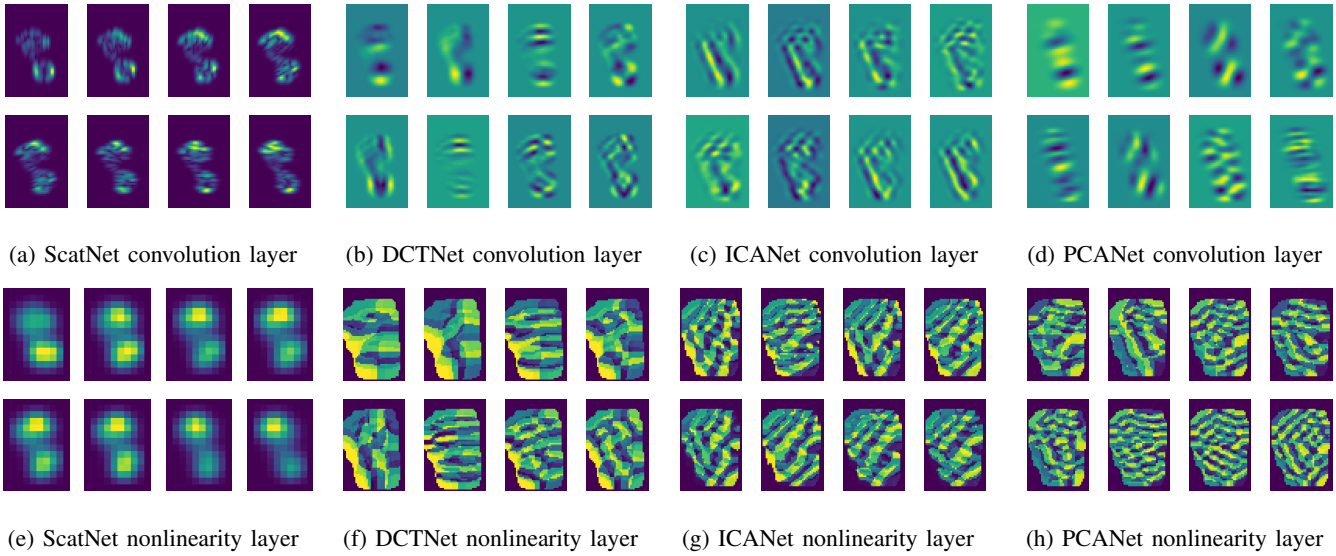
These conventional methods can each extract different characteristics of the foot pressure patterns. Specifically, DWT and DCT are well-known and commonly used for image compression. DWT decomposes data in terms of functions that are localized both in time and frequency, whereas DCT converts data into sets of spatial frequencies. One-dimensional (1D) DWT and 2D DCT were used as implemented in the corresponding feature networks. The Haar wavelet function with a decomposition level of 2 was chosen for 1D DWT, and the 2D DCT was computed by applying 1D DCT for each of the individual rows of the 2D image and then for each column of the image. On the other hand, ICA and PCA are well-known and widely used for dimensionality reduction. While both are linear transformation and unsupervised learning techniques, ICA decomposes data into distinct non-Gaussian (independent) components by optimizing higher-order statistics, and PCA decomposes data into uncorrelated components by optimizing the covariance matrix of the data (second-order statistics).

Besides the kernels/filters, there are two other key components of the convolution layers: hyperparameters and nonlinear operations. The hyperparameters are commonly set before the training process begins (e.g. the size and number of kernels, padding, and stride), and were optimized



(a) DWT filters (b) DCT filters (c) ICA filters (d) PCA filters

Fig. 3: Visualization of different convolution kernels.



(a) ScatNet convolution layer (b) DCTNet convolution layer (c) ICANet convolution layer (d) PCANet convolution layer

(e) ScatNet nonlinearity layer (f) DCTNet nonlinearity layer (g) ICANet nonlinearity layer (h) PCANet nonlinearity layer

Fig. 4: Visualization of features extracted from (a-d) convolutional layers as well as (e-h) nonlinearity layers.

empirically in this study. For nonlinear operation, the outputs of convolution are passed through a nonlinear activation function (e.g. a rectified linear unit (ReLU)) for each layer in the case of CNNs. For feature learning networks, however, each convolutional layer often includes only the convolution operation. Then the nonlinear activation functions are attached after the last convolution layer to reduce the complexity of the architecture. DCTNet, ICANet, and PCANet [4]–[6] use binary hashing methods for nonlinear operation while some methods, such as ScatNet [3], introduce nonlinearity through the feature learning process.

The second building block, the pooling layer, is used to extract the final features for feature learning networks. ScatNet [3] utilizes the common pooling method, i.e., average pooling, whereas the histogram is used for PCANet, ICANet, and DCTNet [4]–[6] (Fig. 2). Unlike the traditional pooling layer, the histogram approach not only reduces the in-plane dimensionality but also provides a nonlinear mapping of the features, leading to increased robustness to small shifts and distortions in the data.

The output of the final convolutional or pooling layer is typically flattened, i.e., transformed into a 1D feature vector. Instead of using the fully-connected layers to transform the 1D array of feature maps into a compact representation suitable for the final decision-making process, the transformed features from the pooling layer are used as the input to conventional ML classifiers. In this study, a linear SVM classifier was employed.

C. Person Verification Models

Biometric systems can work in two modes: verification and identification. The verification mode, i.e., verifying the identity of a claimed user by comparing the enrolled sample with the stored one (1:1), was used to evaluate the performance of feature learning networks in the current study. For each user model, subject labels were binarized to 1 and 0, which indicate the target user and other users (known imposters), respectively. On average, 30 footstep samples were provided for each target user and 2,628 for the other class (the same number of footstep samples were randomly selected for each model to represent imposters). The hold-out (train/test split) method, a more robust algorithm performance estimate when the sample size is small [11], was used: 80% for the training set and 20% for the test set. Biometric performance was measured using the balanced accuracy (BACC), which is the arithmetic mean of sensitivity and specificity, and equivalent to the half total error rate (HTER): 1 - BACC/100. Statistical analysis was conducted using a t-test to compare the BACC results between the two groups. A p-value of less than 0.05 was deemed a significant

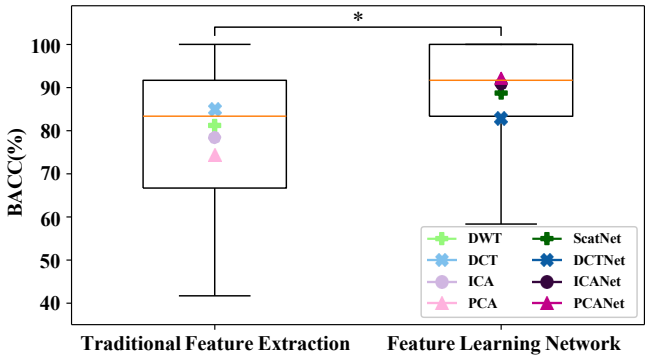


Fig. 5: Balanced accuracies of traditional feature extraction methods (DWT, DCT, ICA, and PCA) and feature learning network methods (ScatNet, DCTNet, ICANet, and PCANet) for person verification of 97 subjects (* $p \leq 0.05$).

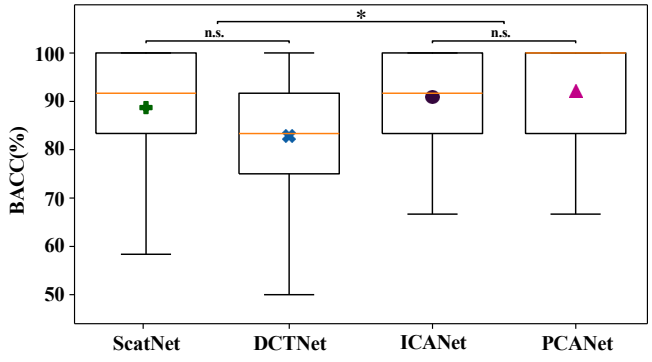


Fig. 6: Balanced accuracies of feature learning-free approaches (ScatNet and DCTNet) and feature learning-based approaches (ICANet and PCANet) for person verification of 97 subjects (n.s. $p > 0.05$; * $p \leq 0.05$).

difference between the two groups.

III. Results

Examples of the calculated filter banks derived from each feature learning method and their corresponding outputs from the convolution and nonlinearity operations are shown respectively in Fig. 3 and Fig. 4. On average, a BACC of 90.62% was found for the four feature learning network approaches, and the BACC of 79.69% was found for the four conventional ML counterparts (Fig. 5). Among four feature learning networks, PCANet yielded the highest accuracy (92.17%), followed by ICANet (91.83%), ScatNet (88.72%), and DCTNet (82.81%) (Fig. 6).

IV. Discussion

The first purpose of this study was to evaluate the performance of conventional feature extraction methods (DWT, DCT, ICA, and PCA) and their corresponding feature learning networks based on the CNN framework (ScatNet, DCTNet, ICANet, and PCANet) for person verification using foot pressure-based data during gait. The results showed that, on average, the feature learning networks (90.6% accuracy) achieved significantly better performance than the conventional feature extraction methods (79.7% accuracy) ($p < 0.05$) (Fig. 5).

When each pair of conventional and deep methods was considered separately, DCTNet is the only feature learning network approach that did not provide an improvement over its counterpart (DCT). Interestingly, the comparison of results with 2D DCT was not provided in the original DCTNet research study [4]. In that study [4], however, DCTNet was found to perform poorly when the input image does not follow the assumption of high local correlation (such as in texture images). Unfortunately, for foot pressure images, fine details are one of the most important pieces of information for classification (Fig. 1). With the exception of DCT, however, the performance of other traditional feature extraction methods (DWT, ICA, and PCA) was improved significantly using the CNN framework (Fig. 5). It should be noted that these conventional methods have shown the potential for person authentication using pressure footsteps in previous studies [7], [8]. Future works should thus consider investigating other successful conventional feature extraction methods such as convolution filter banks for floor-based gait recognition. For instance, locally linear embedding (LLE) was previously shown to outperform three other dimensionality reduction techniques (kernel PCA, Laplacian eigenmaps, and normalized spectral clustering with symmetric Laplacian) for foot pressure-based identification of 104 subjects [10].

Among the four feature learning networks in the present investigation, PCANet provided the best verification performance, with a balanced accuracy of 92.2% (Fig. 6). PCANet [6] should therefore be considered as a simple but highly competitive baseline for floor-based gait recognition when a small amount of data is made available for model training.

As PCANet was originally proposed to be a very simple DL network, it lacks several key components compared to normal CNNs, such as a nonlinearity between two successive convolutional layers and the histogram for a pooling method. Several variations of the two-layer PCANet have been proposed in the past years, such as PCANet+ (i.e., utilizing mean pooling between two adjacent convolution layers for nonlinearity and expanding the network depths more than 2) [12] and PCANet-II (i.e., utilizing second order statistical pooling methods) [13]. The performance of different PCANet alternatives should be further investigated. Based on the preliminary results of our ongoing research work, PCANet+ yielded an additional improvement in the person verification by 2% over PCANet (using the same dataset and pre-feature images as the current investigation). Moreover, as many feature learning networks are inspired and developed based on the PCANet architecture (i.e., using binary hashing for nonlinearity and histogram for pooling), the effectiveness and robustness of previously proposed feature learning networks may also be improved by applying state-of-the-art nonlinearity, and pooling methods for deep CNNs [14].

In the current study, the two feature learning-based approaches (PCANet and ICANet) had better recognition performance than the two approaches based on learning-free (ScatNet and DCTNet) ($p < 0.05$) (Fig. 6). The learning-based methods achieved 100% accuracy for 43 individual subject models while there are only three individual models that the learning-free methods yielded 100% accuracy. Based on this finding, future works may consider other state-of-the-art methods that can automatically learn and extract features from the data, such as UMAP (uniform manifold approximation and projection) [15], or even a fusion of the learning-based and learning-free approaches [16]. Moreover, as the present list of feature learning network approaches is incomplete, future works may consider other potential methods such as self-organizing map network (SOMNet) [17], local manifold discriminant analysis projection network (LMDAPNet) [18], and randomized nonlinear PCANet (RN-PCANet) [19].

The second purpose of this study was to further investigate the effectiveness of feature learning networks with a limited sample size. In this study, 80% of approximately 30 footsteps per user were used to train the verification models. This is used to simulate applications (such as gate access for secure buildings or border control at airports) where quick enrollment sessions may only permit a collection of several strides (i.e., walking for a few tens of seconds). In this study, feature learning network approaches that leveraged the architecture of CNNs in DL showed better performance than traditional ML frameworks for footstep recognition with a limited training sample size ($\approx$ 24 footsteps per user). Conversely, in the study of Costilla-Reyes *et al.* [1], when the number of available footsteps was around 40 per user ($\approx 1.6\times$ the number of training samples in this work), traditional ML algorithms outperformed (or were comparable) to end-to-end deep ResNet using the same feature inputs. With this context, this suggests that the feature learning approaches described here were better able to capture relevant discriminative information than their end-to-end network counterparts, despite having less training data. Nevertheless, a future comprehensive comparison of conventional, feature learning network, and end-to-end approaches in the context of data needs is warranted.

Although the 92.2% verification accuracy obtained by the feature learning networks is encouraging, and the number of subjects employed ($n = 97$) is similar to previous studies, it is worth noting that some applications may necessitate the verification of several hundreds/thousands of individuals or more. Thus, the development of new large population datasets using floor sensors is needed, with which the performance of such feature learning networks could be re-examined.

In conclusion, these findings suggest that feature learning networks are a simple and effective solution for floor sensor-based gait recognition for use in security access scenarios when only a small amount of training data are made available during user enrollment. Feature learning networks may be considered for other pattern recognition problems when only a small number of training samples per observation is available.

REFERENCES

[1] O. Costilla-Reyes, R. Vera-Rodriguez, P. Scully, and K. B. Ozanyan, "Analysis of Spatio-Temporal Representations for Robust Footstep Recognition with Deep Residual Neural Networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 285–296, 2019.

[2] P. Terrier, "Gait Recognition via Deep Learning of the Center-of-Pressure Trajectory," *Applied Sciences*, vol. 10, no. 3, 2020.

[3] J. Bruna and S. Mallat, "Invariant Scattering Convolution Networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1872–1886, 2013.

[4] C. J. Ng and A. Beng Jin Teoh, "DCTNet: A Simple Learning-free Approach for Face Recognition," in *2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, pp. 761–768, 2015.

[5] Y. Zhang, T. Geng, X. Wu, J. Zhou, and D. Gao, "ICANet: A Simple Cascade Linear Convolution Network for Face Recognition," *EURASIP Journal on Image and Video Processing*, vol. 2018, no. 1, p. 51, 2018.

[6] T.-H. Chan, K. Jia, S. Gao, J. Lu, Z. Zeng, and Y. Ma, "PCANet: A Simple Deep Learning Baseline for Image Classification?," *IEEE Transactions on Image Processing*, vol. 24, no. 12, pp. 5017–5032, 2015.

[7] P. C. Connor, "Comparing and Combining Underfoot Pressure Features for Shod and Unshod Gait Biometrics," in *2015 IEEE International Symposium on Technologies for Homeland Security (HST)*, pp. 1–7, 2015.

[8] R. Khokher, R. C. Singh, and R. Kumar, "Footprint Recognition with Principal Component Analysis and Independent Component Analysis," *Macromolecular Symposia*, vol. 347, no. 1, pp. 16–26, 2015.

[9] S. Zheng, K. Huang, T. Tan, and D. Tao, "A Cascade Fusion Scheme for Gait and Cumulative Foot Pressure Image Recognition," *Pattern Recognition*, vol. 45, no. 10, pp. 3603–3610, 2012.

[10] T. C. Pataky, T. Mu, K. Bosch, D. Rosenbaum, and J. Y. Goulermas, "Gait Recognition: Highly Unique Dynamic Plantar Pressure Patterns Among 104 Individuals," *Journal of The Royal Society Interface*, vol. 9, no. 69, pp. 790–800, 2012.

[11] R. Larracy, A. Phinyomark, and E. Scheme, "Machine Learning Model Validation for Early Stage Studies with Small Sample Sizes," in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, 2021.

[12] C.-Y. Low, A. B.-J. Teoh, and K.-A. Toh, "Stacking PCANet +: An Overly Simplified ConvNets Baseline for Face Recognition," *IEEE Signal Processing Letters*, vol. 24, no. 11, pp. 1581–1585, 2017.

[13] C. Fan, X. Hong, L. Tian, Y. Ming, M. Pietikainen, and G. Zhao, "PCANet-II: When PCANet Meets the Second Order Pooling," *IEICE Transactions on Information and Systems*, vol. E101.D, no. 8, pp. 2159–2162, 2018.

[14] A. Zafar, M. Aamir, N. Mohd Nawi, A. Arshad, S. Riaz, A. Alruban, A. K. Dutta, and S. Almotairi, "A Comparison of Pooling Methods for Convolutional Neural Networks," *Applied Sciences*, vol. 12, no. 17, 2022.

[15] L. McInnes, J. Healy, N. Saul, and L. Großberger, "UMAP: Uniform Manifold Approximation and Projection," *Journal of Open Source Software*, vol. 3, no. 29, p. 861, 2018.

[16] C.-Y. Low, A. B.-J. Teoh, and C.-J. Ng, "Multi-Fold Gabor, PCA, and ICA Filter Convolution Descriptor for Face Recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 1, pp. 115–129, 2019.

[17] R. Hankins, Y. Peng, and H. Yin, "SOMNet: Unsupervised Feature Learning Networks for Image Classification," in *2018 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, 2018.

[18] Y. Li, G. Cao, and W. Cao, "LMDAPNet: A Novel Manifold-Based Deep Learning Network," *IEEE Access*, vol. 8, pp. 65938–65946, 2020.

[19] M. Qaraei, S. Abbaasi, and K. Ghiasi-Shirazi, "Randomized Nonlinear PCA Networks," *Information Sciences*, vol. 545, pp. 241–253, 2021.