Dense & Attention Convolutional Neural Networks for Toe Walking Recognition

Junde Chen, Rahul Soangra, Marybeth Grant-Beuttler, Y. A. Nanehkaran, and Yuxin Wen

***Abstract*— Idiopathic toe walking (ITW) is a gait disorder where children's initial contacts show limited or no heel touch during the gait cycle. Toe walking can lead to poor balance, increased risk of falling or tripping, leg pain, and stunted growth in children. Early detection and identification can facilitate targeted interventions for children diagnosed with ITW. This study proposes a new one-dimensional (1D) Dense & Attention convolutional network architecture, which is termed as the DANet, to detect idiopathic toe walking. The dense block is integrated into the network to maximize information transfer and avoid missed features. Further, the attention modules are incorporated into the network to highlight useful features while suppressing unwanted noises. Also, the Focal Loss function is enhanced to alleviate the imbalance sample issue. The proposed approach outperforms other methods and obtains a superior performance. It achieves a test recall of 88.91% for recognizing idiopathic toe walking on the local dataset collected from real-world experimental scenarios. To ensure the scalability and generalizability of the proposed approach, the algorithm is further validated through the publicly available datasets, and the proposed approach achieves an average precision, recall, and F1-Score of 89.34%, 91.50%, and 92.04%, respectively. Experimental results present a competitive performance and demonstrate the validity and feasibility of the proposed approach.**

***Index Terms*— Idiopathic toe walking, dense & attention network, data mining, machine learning.**

I. Introduction

IDIOPATHIC toe walking (ITW) is an exclusion diagnosis granted when a child walks on the toes without a medical reason [1]. The severity of ITW can vary from landing on the middle foot during the standing phase to loading only on the metatarsal head [2], and the prevalence of toe walking is evaluated to affect up to 5% of normal children [3]. Persistent ITW without treatment may cause an increased risk of falling or tripping [4], leg pain [5], injured muscles and motor coordination [6], and organic anomalies [7]. Early identifying toe walking in clinical diagnosis can facilitate timely intervention and treatment. The conventional identification of toe walking primarily relies on the visual observation of clinical specialists. This depends entirely on the experience of the specialist and has a certain degree of subjectivity. Besides, rigorous laboratory-based gait analysis protocol requires special laboratory apparatus like an instrumented walkway, infrared camera-based motion capture system, or treadmill with an integrated force plate [4]. Such laboratory equipment is costly and restrictive, requiring specialized personnel to manipulate and analyze gait data. This is undoubtedly inefficient, labor-intensive, expensive, and cannot be promoted widely. Hence, there is a great need and important realistic significance to seek new systems for automatically detecting toe walking in natural settings.

With the recent advancements in embedded intelligence and sensing technologies, novel approaches for detecting toe walking have emerged. The integration of machine learning (ML) and embedded sensor devices such as inertial measurement units (IMUs) has made it easier to collect various gait data and diagnoses for patients, which significantly enhances the prediction accuracy and operational efficiency. ML strategies for human activity identification, including support vector machines (SVM), multi-layer perceptron (MLP), and random forest (RF), have been widely investigated in healthcare, owing to their promising ability to address multiple-dimensional and nonlinear data patterns. The most popular applications include fall detection [8], gait pattern classification in post-stroke patients [9], walking versus running [10], Parkinson's disease diagnosis [11], [12], among others [13]. For example, Ilias et al. [14] combined the artificial neural network (ANN) and SVM methods to classify the gait patterns of autistic children from normal gait. Their experimental results reveal that the fusion of kinematic and temporal-spatial contributes the highest accuracy of 95% for the ANN classifier. Chakraborty et al. [15] detected pathological gait using several ML models. They found that the multiple adaptive regression splines (MARS) algorithm outperformed the SVM

Manuscript received 5 January 2023; revised 20 March 2023; accepted 27 April 2023. Date of publication 2 May 2023; date of current version 10 May 2023. *(Corresponding author: Yuxin Wen.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Children's Hospital of Orange County under Application No. 170870.

Junde Chen and Yuxin Wen are with the Dale E. and Sarah Ann Fowler School of Engineering, Chapman University, Orange, CA 92866 USA (e-mail: yuwen@chapman.edu).

Rahul Soangra is with the Department of Physical Therapy, Crean College of Health and Behavioral Sciences, Chapman University, Irvine, CA 92618 USA.

Marybeth Grant-Beuttler is with the College of Health, Arts and Sciences, Oregon Institute of Technology, Klamath Falls, OR 97601 USA.

Y. A. Nanehkaran is with the School of Information Engineering, Yancheng Teachers University, Yancheng, Jiangsu 224000, China.

Digital Object Identifier 10.1109/TNSRE.2023.3272362

and logistic regression models with the best accuracy of 88.3%. Based on the SVM classifier, Pendharkar et al. [16] proposed a technology to identify ITW gait patterns on the heel accelerometry data. Using a feature selection algorithm, they realized a maximum accuracy of 87.5% for the SVM classifier. Despite impressive performance obtained in the literature, the conventional ML methods encounter some bottlenecks, such as the dependence on hand-crafted features, lack of robustness, overfitting risks, and low accuracy. More recently, deep learning (DL), especially convolutional neural network (CNN), has achieved great success to address most technical challenges relevant to object recognition and classification. After using IMU sensors to capture the spatial data, Bijalwan et al. [17] employed four different DL models to implement the classification of human gait data. Their proposed CNN model achieved the best accuracy of 90%. Martinez-Hernandez et al. [18] introduced a CNN to recognize walking activities and predict gait periods using the data captured by wearable sensors. Their proposed CNN model attained the best average accuracies of 98.32% and 100% for recognizing gait periods and walking activities, respectively. Depending upon the data captured through the sensor-enabled insoles, Mei et al. [19] exploited a one-dimensional convolutional neural network (1D CNN) to build an exhaustive wearable gait analysis framework for gait classification. Their proposed method reached the highest accuracy of 99.26%. Combining the spatial transformer and temporal convolutional networks, Zhang et al. [20] developed a novel network architecture named Gait-TR for skeleton-based gait recognition and they got around 90% accuracy rate in walking with coats cases. Though reasonably good findings have been reported in the literature, deep neural networks require a large amount of data to train models, which is undoubtedly a challenging task. Besides, numerous training samples are prone to introduce noise data, increasing DL models' overfitting risks. Despite the limitation, previous research has confirmed the efficacy of CNN models for identifying toe walking. In this study, we propose a novel convolutional network architecture to address the above concern. As such, a 1D-based Dense & Attention Convolutional Neural Network, which we termed DANet, is proposed to identify toe walking. The dense block is integrated into the network to maximize information transfer and avoid missed features for the classification. Also, the attention mechanism is incorporated into the network to infer a more powerful hidden representation while suppressing unwanted noises. In addition, the existing Focal loss function is enhanced to make it suitable for multi-classification tasks and alleviate unbalanced samples. Overall, the key contributions of this study are recapitulated as follows.

- A large toe-walking dataset of 593,880 sample data is captured from real-world experimental scenarios via wearable sensors. This dataset is expected further to facilitate research on the recognition of toe walking.
- The study proposes a novel 1D Dense & Attention convolutional network architecture, DANet, that borrows the idea of DenseNet and introduces a 1D dense block to maximize information transfer while reducing redundancy.
- The attention mechanism is incorporated into the network to infer a more powerful hidden representation and realize dynamic recalibration, highlighting useful features while suppressing unwanted noises. Besides, the enhanced Focal-Loss (EFL) function is utilized for alleviating the unbalanced sample issue.

The remaining writing is organized as follows. Section II describes the materials and discusses the methodology. Notably, the proposed approach is primarily discussed in this section. Section III elaborates on experimental analysis. Extensive experiments are implemented in this section along with comparative analysis. Section IV concludes this paper with a summary and points out the direction of future work.

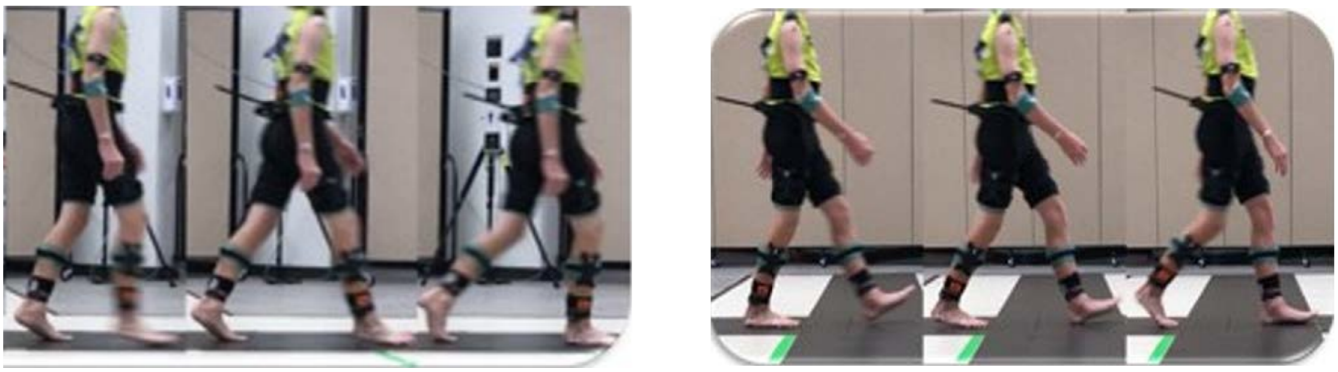
II. Materials and Methods

A. Materials

The collected data came from the idiopathic toe walking (ITW) experiment conducted on a total of 36 children diagnosed as ITW with the age of 9.4 $\pm$ 2.8 years old (The children are 53.8$\pm$6.6cm tall and body weight 75.0$\pm$27.2lbs). All participants signed a written consent form before participation that Chapman Institutional Review Board (IRB) approved (Children's Hospital of Orange County #170870). Four wireless sensor modules of Xsens MTw sensors, which include 3D rate gyroscopes and 3D accelerometers for measuring the angular velocities and acceleration, are utilized in the experiment. The sensors are affixed on the subject's sacrum, posterior torso at T4, and the right and left lateral calves adjacent to the lateral malleolus. These participants were instructed to walk more than 15 meters barefoot, and each participant walked multiple trials. 10 trials of 10m walk were recorded for each walk type of Best Heel Strikes (BHS) and Toe Walk (TW). Each 10m walk trial was comprised of 3 to 4 gait cycles. Data was collected using the Xsens MT Manager Software suite. The sampling frequency was set to 75Hz, which has been proven to be sufficient for human movement analysis in daily activities [21]. A total of 593,880 sample data points, corresponding to 492 complete trials, were acquired in the experiment, where 119,988 samples were used as the test set and 473,892 samples were used for training. The sample data are classified into two categories, BHS and TW, which are determined by 24 attribute features including Sacrum_Acc_X, and Trunk_Acc_X, among others. Fig. 1 shows an example of BHS and TW gait trials. The representative sample data collected is presented in Table I.

B. Methodology

1) Related Work: Deep learning methods have proven to be quite promising and show a strong ability to address both large-scale and small-scale problems for human activity recognition. One of the core advantages of deep learning is the essential ability of end-to-end learning. Among various ML models, CNN is a favorable deep learning architecture due to its powerful adaptive learning capability, and has been widely studied for human activity recognition [22], [23]. Jiang and Yin [24] assembled signal sequences of accelerometers and gyroscopes into images. Then, a deep CNN was utilized to



(a) Typical or Toe Walking (b) Best Heel Strikes (BHS)

Fig. 1. Data collection process from a Child diagnosed with ITW a) Typical or toe walking (TW), b) Best Heel Strikes (BHS).

TABLE I
THE REPRESENTATIVE SAMPLE DATA COLLECTED DURING GAIT TRIALS

FileID	Label	Sacrum_Acc_X	Sacrum_Acc_Y	Sacrum_Acc_Z	...	Trunk_Gyr_Y	Trunk_Gyr_Z	code
TW40MM_BHS006	BHS	-0.14234	9.4322	0.950012	...	0.518371	-0.04338	0
TW40MM_BHS006	BHS	-0.15055	9.344354	0.970887	...	0.599973	0.002638	0
TW40MM_BHS006	BHS	-0.12717	9.368429	0.928864	...	0.66572	0.052088	0
TW40MM_BHS006	BHS	-0.05669	9.382129	0.920413	...	0.66572	0.052088	0
TW26BR_NW006	TW	1.475965	9.559959	-0.59324	...	-0.07272	0.04183	1
TW26BR_NW006	TW	1.463219	9.587491	-0.54691	...	-0.05691	0.047288	1
TW26BR_NW006	TW	1.500867	9.620604	-0.56865	...	-0.06403	0.04644	1
TW26BR_NW006	TW	1.473951	9.619473	-0.58866	...	-0.09164	0.041641	1

learn optimal features from images for the activity recognition task. Ignatov [25] exploited CNNs for local feature extraction coupling with manually crafted statistical features, which can preserve contextual information pertaining to the overall structure of time series. Considering the features obtained by fixed convolution kernel sizes are insufficient, Han et al. [26] proposed a heterogeneous two-stream CNN architecture to encode contextual information of sensor signals from different receptive field sizes, which is capable of generating more discriminative features at different time scales compared with regular CNN. Xi et al. [27] adopted a dilated CNN method to expand the receptive field to solve the information loss issue, and their results achieved a satisfactory performance. Recently, Transformers [28] have gained increasing popularity as they are usually believed to own higher modeling capacity and representation flexibility than classical CNN methods. Transformers implement an attention-based encoder-decoder architecture for sequence analysis. By stacking attentional layers that scan the sequence, Transformers are capable of producing position and context aware representations. Inspired by Transformers, a few attempts have been made to introduce transformer-like architectures to vision tasks [29], [30], one of which, called vision transformer (ViT) [31], has been successfully applied for image recognition and shows competitive performance [32], [33]. Hussain et al. [34] explored a pretrained Vision Transformer to extract frame-level features and then passed the features to a long short-term memory to recognize human activities. Nevertheless, it is questionable whether such potential has been fully unleashed in practice, as the learned transformer networks often suffer from limited access to higher-level representations, over-smoothing, yielding likely redundant models, and higher computational overhead, etc [35]. Moreover, most research focuses on human activities recognition, where the data are collected from healthy individuals. Only a limited number of research investigate ITW gait recognition, i.e., distinguish the toe walking gait from normal gait pattern in ITW children. The ITW pattern detection is quite challenging. The reason is that the frequency of toe walking in ITW children varies. Some ITW children walk on their toes 100% of the time, while for others, the frequency keeps the change. The recognition of ITW has not been investigated to its full potential. To fill this gap, we develop a 1D Dense & Attention convolutional network architecture, which we termed DANet, to perform the recognition and diagnosis of idiopathic toe walking.

2) DenseNet: DenseNet is a popular deep convolutional neural network (DCNN) architecture initally proposed by Huang et al. [36] to alleviate the gradient vanishing issue with increased network depth. As a densely connected network, DenseNet is comprised of dense blocks and transition blocks to improve the propagation of features, thereby reducing information loss during transmission. The output of all the layers is input into the subsequent layers in the dense block, which is a repetition of Batch Normalization (BN), 1×1 convolution, BN, Rectified Linear Unit (ReLU) function, and 3×3 convolution for a certain number of times. The formula of dense connection is expressed as

$$f_i = C_i([f_0, f_1, \cdots, f_{i-1}]) \tag{1}$$

where i indexes the layer number, $[f_0, f_1, \cdots, f_{i-1}]$ denotes the cascading feature maps from 0 to i-1 layers, and $C_i(\cdot)$ indicates a transformation operation such as BN, ReLU,

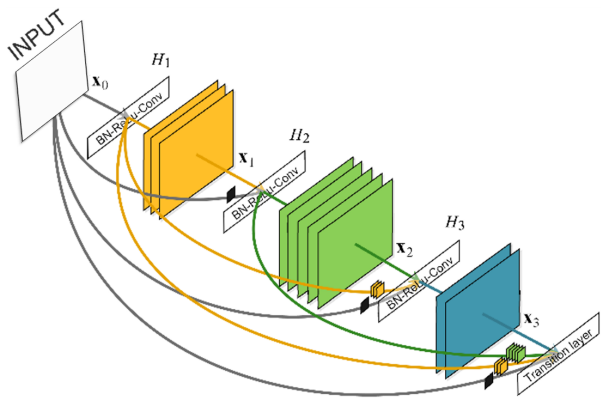


Fig. 2. A DenseNet comprised of a 3-layer dense block.

Convolution (Conv) or Pooling. Subsequently, the transition layer is dedicated to connecting two adjacent dense blocks to reduce the dimension using a 1×1 convolution pooling layer. The relatively narrow convolution kernels, such as 1×1 and 3×3 convolutions, are utilized in the network, limiting the number of learned feature maps, thereby reducing redundancy. Fig. 2 depicts a typical framework of DenseNet comprised of a 3-layer dense block. Borrowing the idea of DenseNet, we exploit the advantages of dense blocks and introduce this structure into our network to maximize information transfer while reducing redundancy.

3) Attention Mechanism: The attention mechanism in deep learning is very similar to human visual attention which can focus on meaningful information while ignoring needless noises [28]. It can efficiently highlight the features that are favorable for classification tasks and has been utilized widely in practical business scenarios. Referring to the work of [37], the intuition of using the attention layers is to infer more powerful hidden representations by weighting the context vectors (Cvec), which efficiently aggregates inputs from different sources. The weighted vector can be written using Eq. (2).

$$C_{vec}^{t} = \sum_{i=1}^{T} w_{ti} h_i^{enc} \tag{2}$$

where w_t indicates the weight of each annotation h_i^{enc} calculated by soft-maxing the corresponding attention score. The formula of the weight is expressed as

$$w_i = \frac{\exp(s_i)}{\sum_{k=1}^{T} \exp(s_k)} \tag{3}$$

In Eq. (3), s_i denotes the alignment score of h_i^{enc} at each step t. It can be computed using Eq. (4).

$$s_i = f(h_i^{enc}), i = 1, \ldots, T_x \tag{4}$$

where the function f symbolizes a feed-forward neural network operation with the $tanh$ function. The intuition of the scoring function is to make the model learn alignment weights together with translations while inferring the entire model layers.

4) Proposed DANet Architecture: As mentioned previously, inspired by the promising performance in computer vision, we exploit the merits of DenseNet and introduce the structure of dense block into our network. A 1D-based Dense & Attention Convolutional Neural Network, which we termed DANet, is proposed to implement the diagnosis of idiopathic toe walking. To maximize the information transfer while reducing redundancy, the dense block is integrated into the network to avoid missed features, and also, the attention mechanism is incorporated into the network to infer a more powerful hidden representation through a weighted context vector. More specifically, the detailed descriptions of this procedure are presented below.

First, the dimension of input data is expanded so that the dataset could satisfy the requirement of the 1D convolution operation. There are n attribute features extracted for the original data in the empirical analysis., e.g., the indicator variable number of the original input data is 24 here. The dimension of the original input data is extended from 24 to (24,1). Subsequently, a set of convolution operations with the filter number of 32, 16, and 8 are implemented, respectively. The sizes of the convolution kernels are all assigned as 5. In this process, the BN, and max Pooling are also operated following the convolution operations. Further, two attention modules are sequentially embedded into the network to highlight useful features while suppressing unwanted noises. More than that, the optimized dense block, where the convolution operation is executed using the 1d convolution kernel and the existing filter size of 3×3 is replaced by 3, is connected for maximizing information transfer. The dropout operation is added to decrease the over-fitting risks, where the dropout rate is set to 0.2. Besides, the LeakyReLU function is used in our network instead of the ReLU function, for the traditional ReLU function makes neurons unable to learn negative input [38]. As such, a BN, a dropout, a LeakyReLU, and a 1 and 3 convolutions are repeated 12 times (growth rate =12) in the optimized dense block. Following the enhanced 1D dense block and a BN layer, another attention module is added to normalize the input to avoid vanishing gradient and highlight the favorable features for classification. At last, a max Pooling (MAP) layer along with a flatten layer are incorporated into the dense convolution module and followed by a fully-connected (FC) Softmax layer with the actual number of categories. Here, the Adam optimizer [39] is used to substitute the Stochastic Gradient Descent (SGD) one in our network since it is more suitable for problems with noise and sparse gradients.

Moreover, to alleviate the issue of unbalanced sample distribution, a Focal-Loss (FL) function was recommended by reference [40] to substitute the classical Cross-Entropy (CE) Loss function since it regards the classifying loss weights of positive and negative samples as the same. The formula of the FL function is defined by

$$L(p_k) = -(1 - p_k)^{\beta} \theta_k \log(p_k) \tag{5}$$

In Eq. (5), k indexes the class number, θ_k is a weighting factor, β is a hyper-parameter of modulating factor, and p_k represents the predicted distribution. However, the existing FL function is designed to handle binary issues in object detection and is not suitable for multi-classification problems. On account of this, we modified the traditional FL function and introduced an enhanced FL function to replace the CE function in our

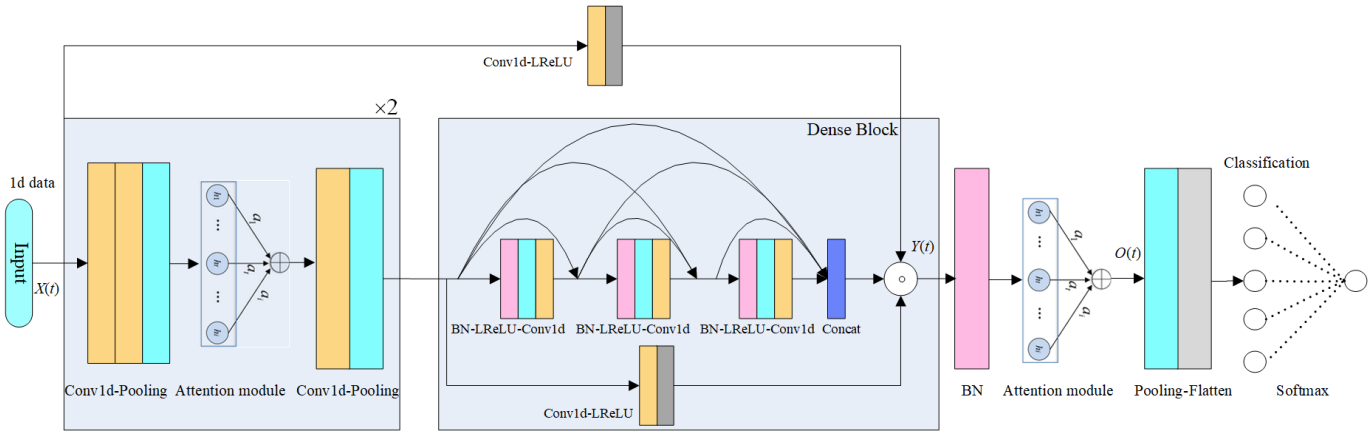


Fig. 3. The proposed DANet architecture.

TABLE II
THE MAIN PARAMETERS OF THE NETWORK

Layer (type)	Input shape	No. of filters	Kernel size	Output shape	Repeated times
Input layer	(None, 24,1)	-		(None, 24,1)	1
Conv1d	(None, 24,1)	32	5	(None, 24,32)	2
BatchNormalization	(None, 24,32)	-	-	(None, 24,32)	1
MaxPooling1d	(None, 24,32)	-	-	(None, 12,32)	1
Attention Module	(None, 12,32)	-	-	(None, 12,32)	1
Conv1d	(None,12,32)	16	5	(None, 12,16)	2
Attention Module	(None, 6,16)	-	-	(None, 6,16)	1
Conv1d	(None,6,16)	16	5	(None, 6,16)	2
Conv1d	(None,3,16)	8	5	(None,3,8)	2
MaxPooling1d	(None, 3,8)	-	-	(None, 2,8)	1
Dense Block	(None,2,8)	4	3, 1	(None, 2,20)	1
BatchNormalization	(None, 2,20)	-	-	(None, 2,20)	1
Attention Module	(None, 2,20)	-	-	(None, 2,20)	1
MaxPooling1d	(None, 2,20)	-	-	(None, 1,20)	1
Flatten	(None, 1,20)	-	-	(None, 20)	1
Softmax	(None, 20)	-	-	(None, 2)	1

network. The formulas of the enhanced FL function are written as follows.

$$L_{multi}(p_k) = -\sum_{k=1}^{C} w_k(1 - p(k|x))^{\beta}\delta_k \log(p(k)) \quad (6)$$

$$w_k = count(x)/count(x \in k) \quad (7)$$

$$\delta_k = \begin{cases} 1, & k = true\_label \\ 0, & k \neq true\_label \end{cases} \quad (8)$$

where x represents the sample. Fig. 3. portrays the architecture of the proposed DANet, and the main parameters are presented in Table II.

III. EXPERIMENTAL RESULTS AND ANALYSIS

A. Experiment Setup

Extensive experiments have been implemented to validate the effectiveness of the proposed approach. In addition to some graphical representations implemented by *R* tools, the main algorithms were accomplished using Python 3.6, where the frequently-used libraries like scikit-learn, OpenCV3, Tensorflow, and Keras were applied and accelerated by GPU. The experimental hardware environment contains RTX 3070 Graphics Card, 32GB memory, and AMD 3.30GHZ CPU, which are utilized for algorithm operation.

B. Experiments on Local Dataset

As mentioned in Section II *A*, the extracted toe-walking data are utilized in our experiments. There are 593,880 instances in this toe walking dataset, where the training set includes 473,892 samples while the test set is comprised of 119,988 records. The toe walking dataset is divided into two categories, BHS and TW, determined by 24 attribute features. It is essential to emphasize that each fileID (object) contains 1,212 records and the type of all the records for each sensor is the same. In other words, this is a $1,212 \times 24$ matrix that determines the type of each object. The training set has 391 objects and the test set consists of 99 objects. Considering a matrix data input, we first employed a convolutional neural network method (CNN-2D) to build the classification model on the two-dimensional sample data. However, satisfactory results are not achieved from this method. Therefore, we further extracted the eigenvalue and eigenvector for each object through the principal component analysis (PCA), and the extracted eigenvalue is used as the input of the models. Whilst, the median value of each object is extracted in our experiments. Here, the 1d CNN (CNN-1D) and the proposed DANet are all conducted in the experiments. More than that, the current popular transformer network, which is a multi-head attention based deep learning model is selected in our comparison

experiments. The key components of the model include the self-attention module and point-wise feed-forward network, where 4 parallel attention layers, or heads, are contained in the self-attention module. For the point-wise feed-forward network, two linear layers with ReLU activation function and a dropout of 0.1 are in it. The settings of these parameters ensure the optimum performance of the model in this work. Besides, the commonly-used effective ML methods, including random forest (RF), support vector machines (SVM), and multi-layer perceptron (MLP or ANN) are also used in our comparative analysis. The hyper-parameters of model training for these compared models are assigned as a learning rate of 1×10^{-3}, a mini-batch size of 64, 100 epochs of training, and an Adam optimizer.

Considering the measurement of model efficiency, we evaluate the performance using different metrics like *Accuracy* (*Acc.*), *Recall* (*Rec.*), and *F1-Score* (*F1*). Among them, *Accuracy* has been mastered by the superiority of True Positive (TP) and True Negative (TN) over the total number of samples. *Recall* has been mastered by the superiority of TP over TP and Fales Negative (FN). $F1-Score$ has been mastered by the superiority of $2 \times TP$ over Fales Positive (FP), $2 \times TP$, and FN. Mathematically, the formulas of these indicators are presented as follows.

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP} \tag{9}$$

$$Recall = \frac{TP}{FN + TP} \tag{10}$$

$$F1 - Score = \frac{2TP}{2TP + FP + FN} \tag{11}$$

where TP denotes the number of positive samples correctly detected. FN is in reverse, which implies the number of negative samples mistakenly detected. FP indicates the number of wrong-detected positive samples. TN is the number of properly-detected negative samples. Table III reports the training and validation performance at 30 and 100 epochs, respectively. Using validation accuracy as early stopping criterion, Table IV summarize the prediction performance on the training and test dataset. To evaluate the efficiency of the proposed method, the running time for training 100 epochs of each method is also provided in Table III. Fig. 4 and Fig. 5 portray the training performance and the tested confusion matrices, respectively.

From Table III, it can be visualized that the proposed approach has achieved the best training and validation performance when compared with other methods. After training for 100 epochs, the proposed DANet has attained a validation accuracy of 82.28%, which is superior to other methods. In terms of computation efficiency, we report the running time of training 100 epochs for each method. Our findings reveal that the proposed method is more efficient than CNN-2D and DANet-PCA, slightly worse than CNN-PCA and CNN-1D. The largest discrepancy in training time between our method and the faster benchmarks is no more than 35 seconds, which is not a significant challenge for current hardware capabilities. From Table IV it can be observed that, the proposed approach outperforms all other methods on the test set and reached the highest accuracy of 88.89%. However, the test accuracy of all other ML methods has a significant decrease relative to that on the training set, which means that there may be noise interference and over-fitting problems suffered by these methods. It is worth noting that transformer slightly behaves better than the proposed method on the training dataset. However, when predicting on the test dataset, there is a significant decrease in accuracy. The reason is that the transformer suffers over-fitting issue especially when the training data is not sufficient. In fact, similar phenomena have been observed in [31] and [41]. This can result in a significant decrease in accuracy when predicting on the test dataset. In contrast, the proposed methods demonstrate a comparable performance, leading to a more robust and generalizable performance on the test dataset. In addition, it can be seen from Fig. 4 (a-h) that the validation accuracy of these compared methods can not be further improved and fluctuates around a solid value after training for 100 epochs,

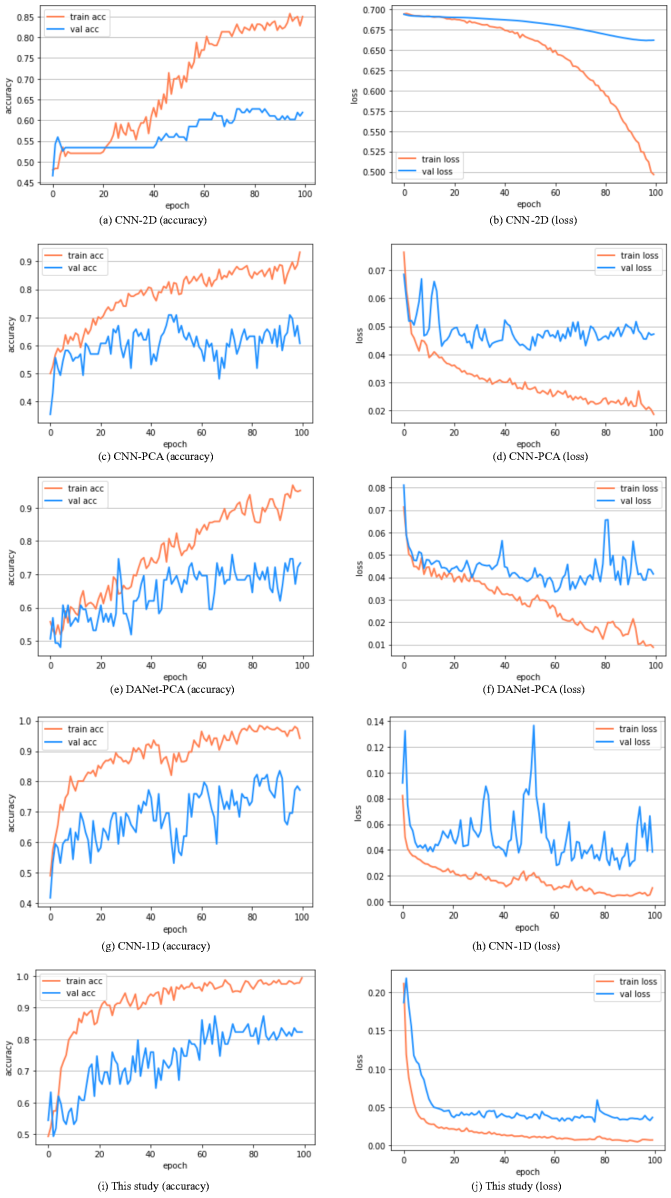


Fig. 4. The training performance of different methods.

TABLE III
THE TRAINING PERFORMANCE OF THE MODELS

Used Methods	Training for 30 epochs				Training for 100 epochs				
	Training Acc. %	Validation Acc. %	Training loss	Validation loss	Training Acc. %	Validation Acc. %	Training loss	Validation loss	Run time (min)
CNN-2D	56.41	50.63	0.6826	0.6888	84.98	61.86	0.4967	0.6618	01:00.97
CNN-PCA	74.04	55.71	0.0326	0.0505	88.87	60.76	0.0203	0.0469	00:23.16
DANet-PCA	66.67	58.23	0.0383	0.0457	95.19	73.24	0.0088	0.0415	01:27.31
CNN-1D	90.71	68.35	0.0151	0.0434	98.72	78.48	0.0031	0.0393	00:28.45
This study	92.95	70.89	0.0127	0.0500	99.36	82.28	0.0045	0.0370	00:57.29

TABLE IV
THE PREDICTION ACCURACY OF DIFFERENT METHODS

No.	Methods	Training set			Test set		
		Acc. (%)	Rec. (%)	F1-Score (%)	Acc. (%)	Rec. (%)	F1-Score (%)
0	CNN-2D	97.75	97.75	97.75	62.62	62.62	62.64
1	CNN-PCA	90.70	90.79	90.70	65.65	65.97	65.57
2	DANet-PCA	91.98	92.31	91.97	72.72	72.60	72.71
3	CNN-1D	91.98	91.96	91.98	85.86	85.92	85.87
4	RF	99.74	99.75	99.87	76.76	76.76	76.78
5	SVM	54.48	52.17	40.87	56.56	54.26	44.61
6	MLP	67.26	67.81	66.93	64.65	65.22	64.27
7	Transformer	94.37	94.38	94.37	81.82	81.98	81.83
8	This study	92.95	93.09	92.95	88.89	88.91	88.89

though the training accuracy increases steadily. The big differences between the training and validation accuracy indicate the data noises and the potential over-fitting problems of these benchmark methods. By contrast, the proposed approach shows a satisfactory result in the experiments of toe walking recognition. Furthermore, from the tested confusion matrices of Fig. 5, it can be observed that the sum of the numbers on the diagonal of the confusion matrix is the largest for the proposed approach, which outperforms other state-of-the-art methods. The crucial explanation for the substantial performance of the proposed approach is that the dense block coupled with the attention modules are incorporated into the network, which maximizes information transfer and highlights the favorable features for the classification, thereby improving the accuracy. The EFL function used in the network also alleviates the imbalanced sample problem. Additionally, this one-dimensional convolutional network architecture reduces the interference of noises and decreases the risk of over-fitting. By comparison, the other methods are commonly-used ML methods or single network structures. Though diverse transformations and feature extraction are implemented, these methods do not achieve satisfactory results. Consequently, the proposed approach achieved a competitive performance in the comparison experiments.

C. Ablation Study

We further implemented an ablation study on our model, in which we analyzed the efficacy of dense block and attention modules on the experimental dataset of idiopathic toe walking. In the first ablation study, we deleted the dense block in the network to probe the performance of the model training. We notice a declining accuracy for this ablated model, where the validation accuracy dropped to 75.95% after 100 epochs of training. The validation accuracy decreased by 6.33%. Likewise, we removed all three attention modules in the network to investigate the effectiveness of the model. After training for 100 epochs, the validation accuracy dropped to 70.89% and decreased by 11.39% for the ablated model. Table V presents the comparison results of ablation experiments. Also, it can be seen from this table that the time-consuming differences are not big among these ablation models after 100 epochs of training. As a consequence, this ablation study results reveal that both the dense block and attention modules significantly contribute to the performance of the proposed method, and relatively, removing the three attention modules causes a notable impact on the accuracy of the model. In the second ablation experiment, we kept the network structure unchanged and evaluated the impact of different loss functions on the model accuracy. To do so, we substituted the existing CE loss function for the EFL function used in the network, and we notice a minor decrease in the accuracy of this ablated model. After 100 epochs of training, the validation accuracy of this ablation model drops to 81.01% (a decrease by 1.17%). The ablation experiment indicates that the EFL function has delivered slightly better results than that of the CE loss function used in our network for toe walking diagnosis.

D. Experiments on Public Datasets

To ensure the scalability and generalizability of the proposed method, the algorithm is further validated through multiple publicly available datasets. UCI library [44] is a universal database comprised of the global collection of data, which dedicates to supplying a series of benchmark datasets to investigate the performance of Machine Learning algorithms in knowledge discovery tasks. Human Activity Recognition dataset (UCI_HAR_data), or HAR in short, is a publicly

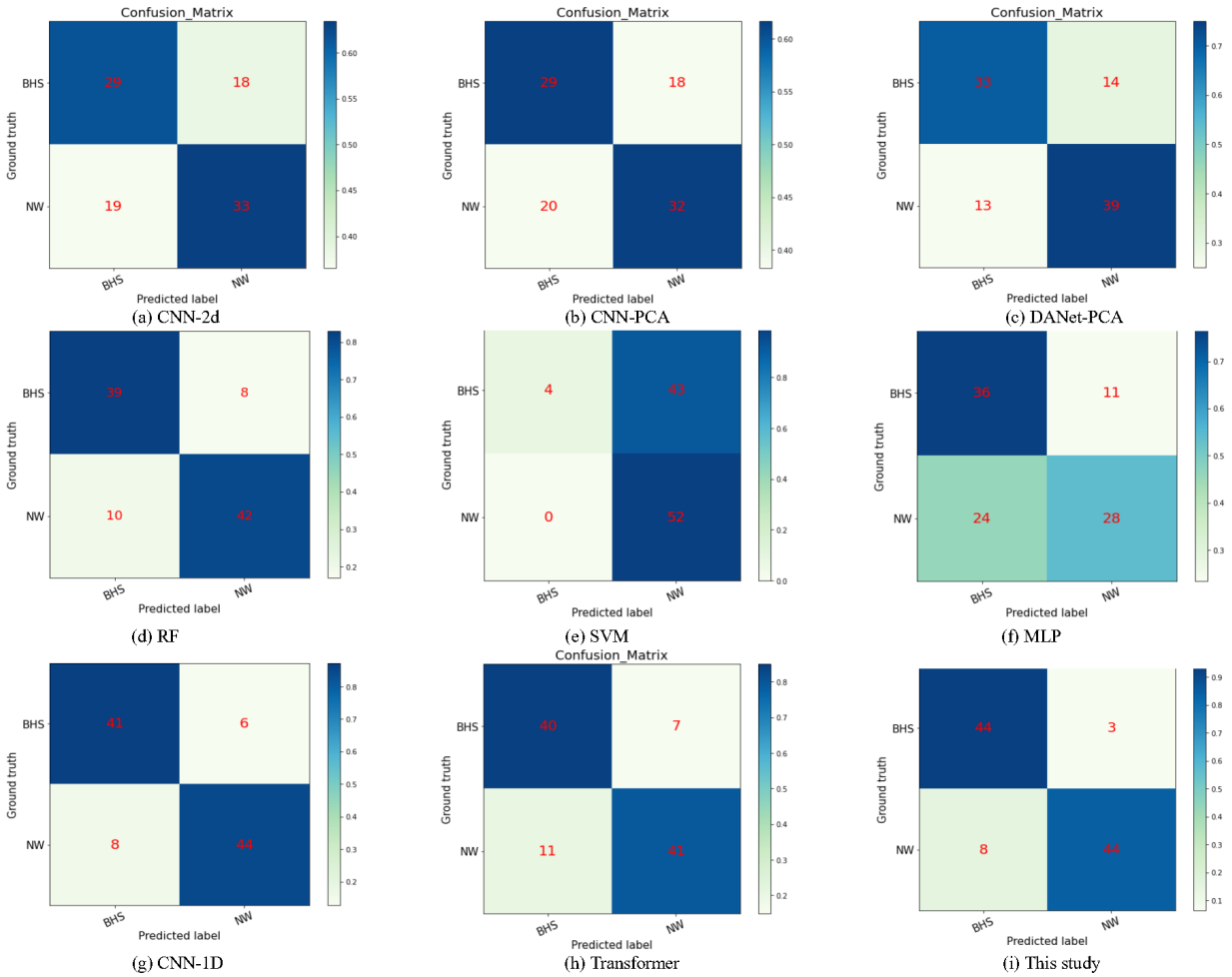


Fig. 5. The confusion matrix of different methods on test dataset.

TABLE V
THE COMPARATIVE FINDINGS OF ABLATION EXPERIMENTS

Ablation approach	Training for 30 epochs				Training for 100 epochs				
	Training Acc. %	Validation Acc. %	Training loss	Validation loss	Training Acc. %	Validation Acc. %	Training loss	Validation loss	Run time (min)
Delete dense block	84.29	55.70	0.0235	0.0468	94.55	75.95	0.0138	0.0373	00:48.10
Delete attention	91.03	64.56	0.0171	0.0998	94.23	70.89	0.0094	0.0657	00:46.52
CE loss function	83.65	67.09	0.0224	0.0503	97.44	81.01	0.0080	0.0439	00:47:88
This study	92.95	70.89	0.0127	0.0500	99.36	82.28	0.0045	0.0370	00:57.29

TABLE VI
DETAILED INFORMATION OF THE UCI DATASETS

Datasets	Number of samples	Number of features	Number of classes	Number of training samples	Number of testing samples
HAR	10299	561	6	7352	2947
Balance	625	4	3	437	188
Musk	6598	166	2	4619	1979
Diabetes	768	8	2	537	231
Skin Seg.	245057	4	2	178539	73518
Pendigits	10992	16	10	7694	3298

available repository established from recordings of 30 subjects implementing daily living activities, which is conducted on 30 volunteers carrying waist-mounted smartphones with built-in inertial sensors. The ages of these volunteers are located in 19-48 years old. Each volunteer performed 6 activities, such as walking, walking_downstairs, walking_upstairs, sitting, laying, and standing, and relevant record data are captured by sensors. A total of 10,299 instances are included in this dataset,

TABLE VII
TRAINING ACCURACY OF COMPARED METHODS (*Pre* : *Precision*, *Rec* : *Recall*, *F1: F1-Score*; %)

Datasets	SVM			MLP			RF			CNN-1D			DANet		
	Pre.	Rec.	F1	Pre.	Rec.	F1	Pre.	Rec.	F1	Pre.	Rec.	F1	Pre.	Rec.	F1
HAR	84.04	84.04	97.90	84.56	85.48	99.68	85.71	85.71	100.00	100.00	100.00	100.00	100.00	100.00	100.00
Balance	94.09	91.07	87.49	48.35	72.31	69.24	99.83	99.77	99.76	84.42	92.22	91.98	94.28	94.28	94.92
Musk	100.00	100.00	100.00	42.41	84.82	77.86	99.96	99.93	99.93	95.94	92.51	91.44	89.56	95.99	96.18
Diabetes	100.00	100.00	100.00	67.22	68.72	63.74	99.57	99.44	99.44	74.40	76.53	76.13	73.29	74.12	74.68
Skin Seg.	99.99	99.99	99.99	99.52	99.78	99.78	99.99	99.99	99.99	99.52	99.54	99.54	99.71	99.88	99.88
Pendigits	100.00	100.00	100.00	75.38	83.01	78.97	100.00	100.00	100.00	99.07	99.49	99.49	97.98	99.19	99.19
Average	96.35	95.85	97.56	69.72	82.35	81.54	97.51	97.47	99.85	92.22	93.38	93.09	92.47	93.91	94.14

TABLE VIII
TEST ACCURACY OF COMPARED METHODS

Datasets	SVM			MLP			RF			CNN-1D			DANet		
	Pre.	Rec.	F1	Pre.	Rec.	F1	Pre.	Rec.	F1	Pre.	Rec.	F1	Pre.	Rec.	F1
HAR	95.19	94.88	95.03	95.01	94.77	94.86	80.49	75.40	90.99	94.32	94.02	94.27	95.14	95.72	95.43
Balance	61.32	92.02	88.73	50.68	76.06	73.35	59.39	81.91	82.38	74.35	87.23	87.53	82.98	82.98	85.88
Musk	94.32	89.24	86.83	42.02	84.03	76.74	97.32	97.22	97.14	94.99	90.65	88.96	87.98	98.28	98.29
Diabetes	32.90	65.80	52.23	56.49	64.94	58.26	80.52	80.95	80.16	68.58	73.16	70.83	74.00	74.46	75.11
Skin Seg.	99.62	99.41	99.41	99.56	99.80	99.80	99.88	99.94	99.94	99.48	99.77	99.77	99.67	99.86	99.86
Pendigits	31.02	10.92	3.36	72.38	78.68	75.06	96.93	96.85	96.84	96.65	97.15	97.17	96.30	97.70	97.68
Average	69.06	75.37	70.93	69.35	83.04	79.68	85.75	88.71	91.24	88.06	90.33	89.75	89.34	91.50	92.04

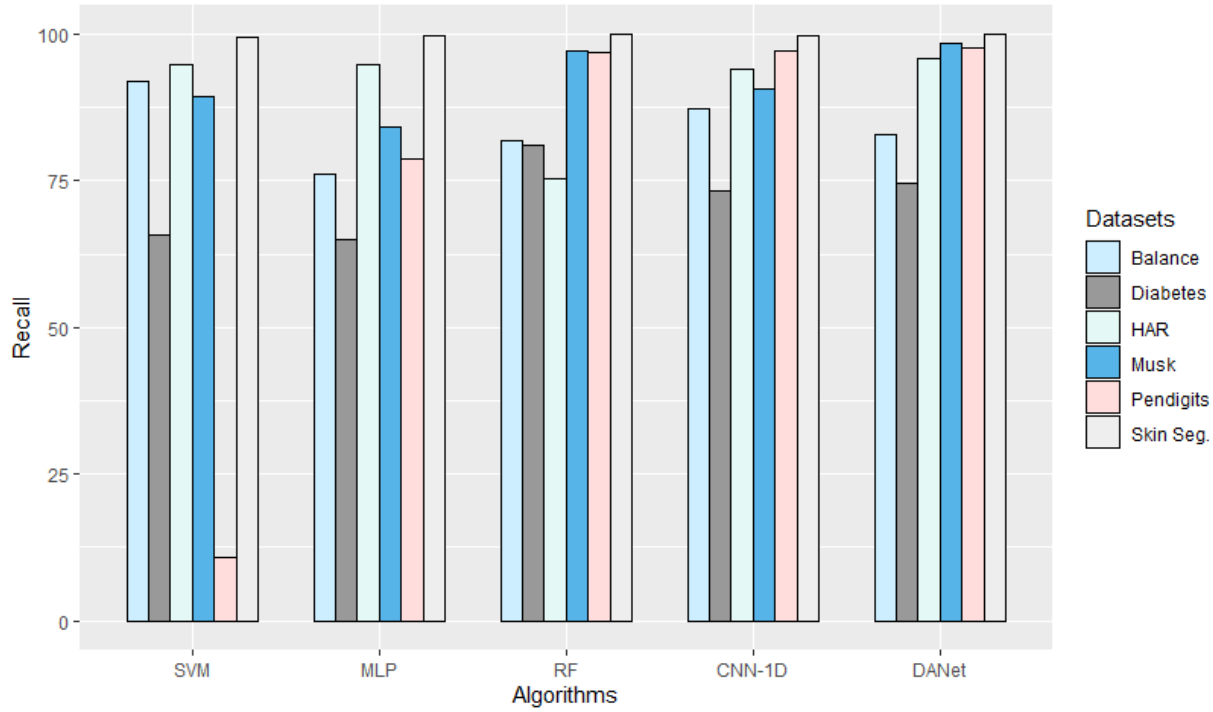


Fig. 6. The test recall rate of different methods.

where 7,352 samples are divided into the training set and 2,947 samples are into the test set. As mentioned above, the dataset is classified into 6 categories, which are determined by 561 frequency domain features. Besides that, to further probe the efficacy of the proposed approach, we also chose the other five UCI datasets including Balance, Musk, Diabetes, Skin segmentation (seg.), and Pegdigits for comparative experiments. Among them, the balance dataset with 625 samples is divided into 3 categories determined by 4 attribute features. There are 6,598 samples in the musk dataset, which is classified into 2 categories determined by 168 attributes. A total of 768 samples are included in the diabetes dataset, which consists of 8 attribute features determining 2 classes, such as positive class (diabetes) and negative class (non-diabetes). Skin seg. dataset is comprised of 245,057 samples which contain 4 attribute features determining 2 classes, such as skin and non-skin. Representing the digits collected from 44 writers, the Pendigits dataset has 10,992 samples which are categorized

TABLE IX
THE COMPARATIVE ANALYSIS WITH RECENT WORK

ID	References	Year	Dataset	Methods	Recall rate (%)
1	Anguita et al. [42]	2012	UCI_HAR_data	Multiclass Hardware-Friendly Support Vector Machine (MC-HF-SVM)	89.00
2	Anguita et al. [43]	2013	UCI_HAR_data	MultiClass SVM (MC-SVM)	89.30
3	Soangra et al. [8]	2022	Self-collected dataset	Wearable Sensors and Machine Learning Algorithms	87.05
3	This study	2022	UCI_HAR_data	1-D Dense & Attention network (DANet)	91.50

into 10 classes determined by 16 attributes. Precision $Pre = (TP/(TP + FP))$, which is mastered by the superiority of true positive overall positive outcome, is used as a measure indicator in the analysis. Table VI summarizes the detailed information of these datasets. The training and testing accuracy of different methods is listed in Tables VII-VIII, and Fig. 6 portrays the classifying recall rate of different methods on the testing datasets.

It can be seen from Tables VII-VIII that the proposed approach has gained an increased performance compared with other well-known algorithms, even though the optimum classifiers are utilized. On the training set, the proposed approach separately reaches the average $Precision$, $Recall$, and $F1 - Score$ of 92.47%, 93.91%, and 94.14%, which are higher than those of most comparison methods except for the RF and SVM. However, the test results of the SVM method are poor, which indicates the overfitting problems of SVM. Similar issues also exist in the RF, and especially, the RF is an ensemble learning (EL) method, which consists of multiple decision tree (DT) algorithms (Here is 20). By contrast, the proposed DANet is an independent network method and it achieves a competitive performance in the experiments of publicly available datasets. On the test dataset, the average accuracy of the proposed approach is superior to that of all other compared methods, which can also be observed from Fig. 6. The overall bars in all regions are higher than that of other methods, which reveals the stability and validity of the proposed approach.

Moreover, we have also accomplished a performance investigation of our method compared to the results reported in existing literature, as presented in Table IX. The comparative findings indicate that the proposed approach has delivered outperformance results compared with other advanced methods. Consequently, based on the experimental analysis, it can be concluded that the proposed approach is successful in detecting idiopathic toe walking, and can also be generalized to other related fields.

IV. CONCLUSION

Persistent toe walking among children affects the development of the foot and ankle muscles, leading to instability and pain, impaired muscle and movement coordination and thereby increasing the risk of falling or tripping. Additionally, toe walking can negatively impact a child's life, leading to teasing, bullying, and self-consciousness. Early identification and intervention can help prevent complications in children with idiopathic toe walking, such as shortening the Achilles tendon. By identifying and treating toe walking early, children can avoid these negative impacts and improve their quality of life. Therefore, looking for an efficient, reliable, and low-cost method to detect toe walking is of great realistic importance. Deep learning techniques, notably diverse convolutional neural networks, have presented an impressive performance in overcoming most technical challenges associated with recognition and classification tasks. In this study, we have proposed a new one-dimensional (1D) Dense & Attention convolutional network architecture, which we termed the DANet, to detect idiopathic toe walking. The dense block is integrated into the network to maximize information transfer and avoid missed features. Also, the attention modules are incorporated into the network to infer a more powerful hidden representation while suppressing unwanted noises. In addition, the EFL function is used to alleviate the unbalanced sample issue. Based on experimental analysis, it can be concluded that the proposed approach has a significant capability for identifying idiopathic toe walking and can also be extended to broad fields. The experimental findings along with the physiotherapists' verification in clinical practice, will enable generation of real-time toe walking monitoring systems based on machine learning. These real-time systems embedded in shoe insoles will quantify the number of heel-strike and toe-strike events during walking in children diagnosed with toe walking. This study will lay the foundation for automated toe-walking detection and demonstrate the effectiveness and feasibility of using ML models for toe-walking detection among children with ITW.

In our experiments, the proposed DANet has proven to be quite promising. However, it does have some limitations. The model has high identification accuracy, but consumes slightly more computational time. Model pruning algorithms can be added to simplify the model in future work. Moreover, we plan to deploy the model in the information system to automatically implement the toe walking recognition. Also, in the future, we would like to perform model inferencing on more practical applications, such as virtual defect assessment, cancer cell recognition, online fault detection, and so on.

REFERENCES

[1] K. Gray, V. Pacey, A. Caserta, D. Polt, and C. Williams, "Development of the idiopathic toe walking outcome (iTWO) proforma: A modified Delphi study and online parent survey for measurement consensus," *Gait Posture*, vol. 99, pp. 111–118, Jan. 2023.

[2] C. Alvarez, M. De Vera, R. Beauchamp, V. Ward, and A. Black, "Classification of idiopathic toe walking based on gait analysis: Development and application of the ITW severity classification," *Gait Posture*, vol. 26, no. 3, pp. 428–435, Sep. 2007.

[3] P. Engström and K. Tedroff, "Idiopathic toe-walking: Prevalence and natural history from birth to ten years of age," *J. Bone Joint Surgery*, vol. 100, no. 8, pp. 640–647, 2018.

[4] R. Soangra, Y. Wen, H. Yang, and M. Grant-Beuttler, "Classifying toe walking gait patterns among children diagnosed with idiopathic toe walking using wearable sensors and machine learning algorithms," *IEEE Access*, vol. 10, pp. 77054–77067, 2022.

[5] J. J. Ruzbarsky, D. Scher, and E. Dodwell, "Toe walking: Causes, epidemiology, assessment, and treatment," *Current Opinion Pediatrics*, vol. 28, no. 1, pp. 40–46, 2016.

[6] V. De Oliveira, L. Arrebola, P. De Oliveira, and L. Yi, "Investigation of muscle strength, motor coordination and balance in children with idiopathic toe walking: A case-control study," *Develop. Neurorehabilitation*, vol. 24, no. 8, pp. 540–546, Nov. 2021.

[7] C. M. Williams, P. Tinley, and M. Curtin, "The toe walking tool: A novel method for assessing idiopathic toe walking children," *Gait Posture*, vol. 32, no. 4, pp. 508–511, Oct. 2010.

[8] M. Musci, D. De Martini, N. Blago, T. Facchinetti, and M. Piastra, "Online fall detection using recurrent neural networks on smart wearable devices," *IEEE Trans. Emerg. Topics Comput.*, vol. 9, no. 3, pp. 1276–1289, Jul. 2021.

[9] K. Kaczmarczyk, A. Wit, M. Krawczyk, and J. Zaborski, "Gait classification in post-stroke patients using artificial neural networks," *Gait Posture*, vol. 30, no. 2, pp. 207–210, 2009.

[10] H. Zhang, Y. Guo, and D. Zanotto, "Accurate ambulatory gait analysis in walking and running using machine learning models," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 28, no. 1, pp. 191–202, Dec. 2020.

[11] E. Balaji, D. Brindha, and R. Balakrishnan, "Supervised machine learning based gait classification system for early detection and stage classification of Parkinson's disease," *Appl. Soft Comput.*, vol. 94, Sep. 2020, Art. no. 106494.

[12] F. Wahid, R. K. Begg, C. J. Hass, S. Halgamuge, and D. C. Ackland, "Classification of parkinson's disease gait using spatial–temporal gait features," *IEEE J. Biomed. Health Inform.*, vol. 19, no. 6, pp. 1794–1802, Nov. 2015.

[13] J. Figueiredo, C. P. Santos, and J. C. Moreno, "Automatic recognition of gait patterns in human motor disorders using machine learning: A review," *Med. Eng. Phys.*, vol. 53, pp. 1–12, Mar. 2018.

[14] S. Ilias, N. M. Tahir, R. Jailani, and C. Z. C. Hasan, "Classification of autism children gait patterns using neural network and support vector machine," in *Proc. IEEE Symp. Comput. Appl. Ind. Electron. (ISCAIE)*, May 2016, pp. 52–56.

[15] S. Chakraborty, S. Jain, A. Nandy, and G. Venture, "Pathological gait detection based on multiple regression models using unobtrusive sensing technology," *J. Signal Process. Syst.*, vol. 93, no. 1, pp. 1–10, Jan. 2021.

[16] G. Pendharkar, D. T. H. Lai, and R. K. Begg, "Detecting idiopathic toe-walking gait pattern from normal gait pattern using heel accelerometry data and support vector machines," in *Proc. 30th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Aug. 2008, pp. 4920–4923.

[17] V. Bijalwan, V. B. Semwal, and V. Gupta, "Wearable sensor-based pattern mining for human activity recognition: Deep learning approach," *Ind. Robot, Int. J. Robot. Res. Appl.*, vol. 49, no. 1, pp. 21–33, Jan. 2022.

[18] U. Martinez-Hernandez, A. Rubio-Solis, and A. A. Dehghani-Sanij, "Recognition of walking activity and prediction of gait periods with a CNN and first-order MC strategy," in *Proc. 7th IEEE Int. Conf. Biomed. Robot. Biomechatronics (Biorob)*, Aug. 2018, pp. 897–902.

[19] Z. Mei, K. Ivanov, G. Zhao, Y. Wu, M. Liu, and L. Wang, "Foot type classification using sensor-enabled footwear and 1D-CNN," *Measurement*, vol. 165, Dec. 2020, Art. no. 108184.

[20] C. Zhang, X.-P. Chen, G.-Q. Han, and X.-J. Liu, "Spatial transformer network on skeleton-based gait recognition," 2022, *arXiv:2204.03873*.

[21] C. V. C. Bouten, K. T. M. Koekkoek, M. Verduin, R. Kodde, and J. D. Janssen, "A triaxial accelerometer and portable data processing unit for the assessment of daily physical activity," *IEEE Trans. Biomed. Eng.*, vol. 44, no. 3, pp. 136–147, Mar. 1997.

[22] V. Pandiyan, J. Prost, G. Vorlaufer, M. Varga, and K. Wasmer, "Identification of abnormal tribological regimes using a microphone and semi-supervised machine-learning algorithm," *Friction*, vol. 10, no. 4, pp. 583–596, Apr. 2022.

[23] K. Chen, D. Zhang, L. Yao, B. Guo, Z. Yu, and Y. Liu, "Deep learning for sensor-based human activity recognition: Overview, challenges, and opportunities," *ACM Comput. Surv.*, vol. 54, no. 4, pp. 1–40, 2021.

[24] W. Jiang and Z. Yin, "Human activity recognition using wearable sensors by deep convolutional neural networks," in *Proc. 23rd ACM Int. Conf. Multimedia*, Oct. 2015, pp. 1307–1310.

[25] I. Andrey, "Real-time human activity recognition from accelerometer data using convolutional neural networks," *Appl. Soft Comput.*, vol. 62, pp. 915–922, Jan. 2017.

[26] C. Han, L. Zhang, Y. Tang, W. Huang, F. Min, and J. He, "Human activity recognition using wearable sensors by heterogeneous convolutional neural networks," *Exp. Syst. Appl.*, vol. 198, Jul. 2022, Art. no. 116764.

[27] R. Xi, M. Hou, M. Fu, H. Qu, and D. Liu, "Deep dilated convolution on multimodality time series for human activity recognition," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2018, pp. 1–8.

[28] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.

[29] C. Xu, Y. Makihara, X. Li, Y. Yagi, and J. Lu, "Cross-view gait recognition using pairwise spatial transformer networks," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 1, pp. 260–274, Jan. 2021.

[30] J. N. Mogan, C. P. Lee, K. M. Lim, and K. S. Muthu, "Gait-ViT: Gait recognition with vision transformer," *Sensors*, vol. 22, no. 19, p. 7362, Sep. 2022.

[31] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[32] Y. Cui and Y. Kang, "GaitTransformer: Multiple-temporal-scale transformer for cross-view gait recognition," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2022, pp. 1–6.

[33] D. Pinčić, D. Sušanj, and K. Lenac, "Gait recognition with self-supervised learning of gait features based on vision transformers," *Sensors*, vol. 22, no. 19, p. 7140, Sep. 2022.

[34] A. Hussain, T. Hussain, W. Ullah, and S. W. Baik, "Vision transformer and deep sequence learning for human activity recognition in surveillance videos," *Comput. Intell. Neurosci.*, vol. 2022, pp. 1–10, Apr. 2022.

[35] A. Fan, T. Lavril, E. Grave, A. Joulin, and S. Sukhbaatar, "Addressing some limitations of transformers with feedback memory," 2020, *arXiv:2002.09402*.

[36] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4700–4708.

[37] B. Belay, T. Habtegebrial, M. Liwicki, G. Belay, and D. Stricker, "A blended attention-CTC network architecture for amharic text-image recognition," in *Proc. 10th Int. Conf. Pattern Recognit. Appl. Methods*, 2021, pp. 435–441.

[38] A. L. Maas et al., "Rectifier nonlinearities improve neural network acoustic models," in *Proc. ICML*, vol. 30, no. 1, Atlanta, GA, USA, 2013, p. 3.

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

[40] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.

[41] J.-N. Chen, S. Sun, J. He, P. Torr, A. Yuille, and S. Bai, "TransMix: Attend to mix for vision transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 12135–12144.

[42] D. Anguita, A. Ghio, L. Oneto, X. Parra, and J. L. Reyes-Ortiz, "Human activity recognition on smartphones using a multiclass hardware-friendly support vector machine," in *Proc. Ambient Assist. Living Home Care, 4th Int. Workshop*. Vitoria-Gasteiz, Spain: Springer, Dec. 2012, pp. 216–223.

[43] D. Anguita, A. Ghio, L. Oneto, X. Parra, and J. R. Reyes-Ortiz, "Energy efficient smartphone-based activity recognition using fixed-point arithmetic," *J. Universal Comput. Sci.*, vol. 19, no. 9, pp. 1295–1314, Jan. 2013.

[44] C. Blake. (1998). *UCI Repository of Machine Learning Databases*. [Online]. Available: http://www.ics.uci.edu/mlearn/MLRepository.html