Wearable-based Classification of Running Styles with Deep Learning

Setareh Rahimi Taghanaki[1], Michael Rainbow[2], Ali Etemad[3]

Abstract— Automatic classification of running styles can enable runners to obtain feedback with the aim of optimizing performance in terms of minimizing energy expenditure, fatigue, and risk of injury. To develop a system capable of classifying running styles using wearables, we collect a dataset from 10 healthy runners performing 8 different pre-defined running styles. Five wearable devices are used to record accelerometer data from different parts of the lower body, namely left and right foot, left and right medial tibia, and lower back. Using the collected dataset, we develop a deep learning solution which consists of a Convolutional Neural Network and Long Short-Term Memory network to first automatically extract effective features, followed by learning temporal relationships. Score-level fusion is used to aggregate the classification results from the different sensors. Experiments show that the proposed model is capable of automatically classifying different running styles in a subject-dependant manner, outperforming several classical machine learning methods (following manual feature extraction) and a convolutional neural network baseline. Moreover, our study finds that subject-independent classification of running styles is considerably more challenging than a subject-dependant scheme, indicating a high level of personalization in such running styles. Finally, we demonstrate that by fine-tuning the model with as few as 5% subject-specific samples, considerable performance boost is obtained.

I. INTRODUCTION

Running is a common aerobic exercise which many people worldwide partake in as a sport. It is often considered one of the most historic types of competitive sport, dating back centuries [1]. Despite its pervasiveness, running can be performed in a wide variety of different '*styles*' [2], [3], often as a result of different physiological/biomechanical factors [4], training [5], and even the intention of the runner. These variations can play major roles in energy expenditure [6], speed/endurance [7], and even injury [2], [8], [9].

Based on the above, we believe it is highly valuable and beneficial to develop systems capable of automatically detecting different running styles. Such a system can be used for injury analysis, coaching and training, and other applications. Moreover, we believe the use of wearable sensors for such a system, as opposed to cameras for instance, will allow for more detailed, exact, and targeted monitoring as well as lack of occlusions of key body parts.

The notion of using wearables and machine learning to classify different activities has been widely studied in the literature. For instance, in [10], hand-crafted features were used along with an ensemble of Support Vector Machine (SVM) learners to classify daily activities using wearable data. Similar approaches have been used in [11], [12], and [13], where different classical machine learning solutions have been used. In [14], deep learning, in the form of a Convolutional Neural Network (CNN) was used to perform end-to-end classification of different activities. In [15], the learned representations from a CNN were fed to a Long Short-Term Memory (LSTM) network to classify actions. More recently, self-supervised solutions have been proposed in [16], [17], [18], and [19] to reduce reliance on labels and learn more generalized representations by means of pretext training with a number of transformations applied to wearable data. However, despite the rich available literature on classification of different activity classes, the notion of detecting specific running styles with machine learning, which may be manifested as a result of very minor and subtle changes in running patterns, has not yet been explored.

To tackle the problem of automatic wearable-based running style classification, we design a system for automatic wearable-based classification of different running styles with deep learning. We first record a rich dataset of different subjects running in a number of different pre-defined running styles. The data are recorded using several wearable sensors placed strategically on different parts of the body including each foot, each medial tibia, and on the lower back. We then use this dataset to train deep neural networks for classification of the different styles. Our results show that our proposed CNN can accurately classify the different running styles and outperforms some classical machine learning models, namely support vector machine (SVM), naive Baysed (NB), decision tree (DT), random forest (RF). Interestingly, however, we find that the performance considerably drops when the experiments are performed in a leave-one-subject-out manner, even when different CNN architectures are explored. This finding points to the fact that running styles, even in pre-defined categories, are highly subject-specific and vary across different runners.

Lastly, we demonstrate that by fine-tuning the model with small portions of unseen subjects' samples, a more robust performance is achieved.

II. METHODOLOGY

A. Dataset

10 healthy and unimpaired runners (5 women and 5 men), ranging in age between 20 to 26 years, were recruited for this study [20]. One of the inclusion criteria was for participants to have been running around 10 km per week over the last three months leading to the data collection session. For each

[1,3]Setareh Rahimi Taghanaki and Ali Etemad are with the Department of Electrical and Computer Engineering, Queen's University, Kingston, Canada, {19srt3, ali.etemad}@queensu.ca

[2]Michael Rainbow is with the the Department of Mechanical Engineering, Queen's University, Kingston, Canada, michael.rainbow@queensu.ca

subject, a set of 5 IMeasureU [21] Inertial Measurement Units (IMUs) sensors were placed on each foot (top of the shoe, attached through laces), each medial tibia (attached with elastic Velcro straps), and one at the centre of mass on the lower back (attached with Velcro to a torso wrap) to record the accelerations and angular changes of their body segments when running. The accelerometer data are sampled at 500 Hz and collected by a personal computer via Bluetooth. For each subject, their height, weight, age, and sex were also recorded. For all the experiments, an instrumented treadmill was used. First, the participants were asked to start running on the treadmill at 1.5 m/s and then gradually increase their speed. This helped the runners to warm up and also was used to determine the comfortable speed for each runner to perform the experiments. Eight different running styles were considered for this experiment: (1) Egg-Beater Gait (where instead of only moving forward, the leg also travels laterally to the side); (2) Bouncing Gait (where the vertical bounce after each step is higher than normal); (3) Heel Strike (where landing occurs on the heels as opposed to mid-foot landing); (4) Toe Strike (where landing occurs on the fore-foot); (5) Lengthened Strides (where the strides are abnormally longer than normal); (6) Shortened Strides (where the strides are abnormally shorter than normal); (7) Wide Stance (where the horizontal space between the two feet is wider than normal); and (8) Narrow Stance (where the horizontal space between the two feet is shorter than normal). These particular classes were selected as they encapsulate exaggerated common variations in running styles among people. Each subject was asked to perform these running styles for approximately 5 minutes after each style was verbally described to them. This approach allowed for subjects to perform slight variations of each style given their own comfort and physiology. Ethics approval was secured from Queen's General Research Ethics Board.

B. Proposed Model

In this study, we propose a CNN-LSTM architecture to classify different running styles based on input multi-sensor accelerometer data. The first part of our model is a CNN which learns to automatically extract representations from the 3-D accelerometer signals. This network consists of 3 blocks with 2 1-D convolutional layers in each block. Figure 1(a) demonstrates the CNN component of our model. The kernel size is set to 3 and the numbers of filters for the convolutional layers in each block are 64, 128, and 256, respectively. These blocks are then followed by a 1D maxpooling layer with a pool size of 2 and a flatten layer to generate the output features from the input signals.

Following the CNN, a Recurrent Neural Network is used in the form of a 2-layer Bidirectional LSTM (BLSTM), each consisting of 256 hidden units. The BLSTM aims to learn the temporal traits and relationships between the representations extracted by the CNN from input raw accelerometer data in both forward and backward directions. The output of the final bidirectional layer is then fed to 2 fully connected layers (where the first one is followed by ReLU activation) of sizes

200 and 8 (number of running classes) and a softmax layer. A similar CNN-LSTM architecture is used for the data from each sensor, and the final outputs of the 5 CNN-LSTM networks are fused through score-level fusion (averaging the output probabilities). Figure 1(b) presents the overall architecture of our proposed solution. The dashed green rectangle in the figure illustrates the details of the network used for each sensor.

To feed each CNN-LSTM model, we divide the running data into segments of 10 seconds in duration, with 50% overlap. As shown in Figure 1, the input of the network is a 10-second 3-D acceleration signal collected from each of the five IMUs placed on different parts of the body. Each 10-second input signal is segmented into 8 successive sub-segments of 1.25 seconds in duration with no overlap. The 1.25 second sub-segments are fed to the CNN network to generate vector representations, to subsequently feed 8 LSTM cells organized in 2 bidirectional layers.

C. Implementation and Evaluation

We implement our proposed model using Keras with TensorFlow backend on an NVIDIA GeForce RTX 2080 Ti GPU. The models are trained over 300 epochs with a batch size 64 and Adam optimizer with a learning rate of 0.0002.

In order to evaluate our method, we use 2 schemes: k-folds cross-validation and Leave-One-Subject-Out (LOSO). For k-folds, we randomly picked 20% of the segments for testing, while using the rest for training. We carried out this approach for 5 trials and presented the means and standard deviations of accuracy and F1 scores among all 5 folds. In each trial, 10% of the training set was allocated for validation. In LOSO, for each trial, 20% of the subjects were selected for testing and the remaining subjects were set to train the model. Similar to the previous approach, 10% of the training data was set aside for validation.

D. Baselines

In order to better evaluate the robustness of our proposed method, several common classification methods are applied for comparison. Specifically, 4 classical machine learning classifiers are used. These classifiers are: Naive Bayes, Decision Tree, SVM, and Bagged Tree Ensemble (BTE). We extract 8 common manual features (mean, standard deviation, minimum, maximum, root mean square, skewness, kurtosis, and peak-to-peak time) from each segment to train these classical models.

The hyperparameters for each model are tuned with the help of Bayesian optimization. For SVM, cubic kernel and one-vs-one scheme were the best fits for all the sensors, with the exception of the LFoot sensor where a Gaussian kernel function and one-vs-all scheme showed the best results. For the Decision Tree, maximum deviance reduction was set as the Split criterion for all sensors except for the COM sensor which used the Towing rule.

In addition to the 4 classical machine learning methods, a CNN was designed to automatically extract the features from each input signals and classify the running styles
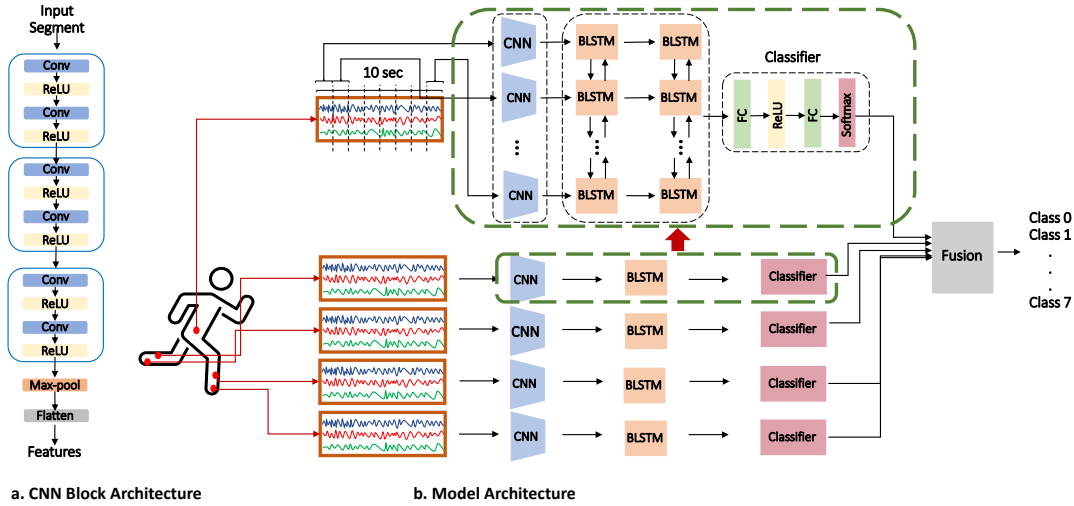
Fig. 1: The architecture of the proposed CNN-LSTM model.

TABLE I: Results for the k-fold cross-validation scheme.

Models	COM		LFoot		LShank		RFoot		RShank		Fusion	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
Naive Bayes	0.309±0.012	0.294±0.012	0.306±0.016	0.289±0.013	0.323±0.015	0.298±0.013	0.316±0.019	0.291±0.015	0.332±0.014	0.302±0.010	0.317±0.018	0.295±0.014
Decision Tree	0.560±0.020	0.560±0.019	0.478±0.016	0.477±0.016	0.603±0.012	0.603±0.013	0.472±0.010	0.471±0.010	0.610±0.012	0.610±0.011	0.545±0.061	0.544±0.061
SVM	0.689±0.011	0.688±0.012	0.658±0.013	0.657±0.012	0.715±0.013	0.714±0.013	0.614±0.008	0.613±0.007	0.745±0.011	0.745±0.011	0.685±0.046	0.683±0.046
BTE	0.763±0.015	0.763±0.015	0.676±0.003	0.675±0.003	0.750±0.007	0.750±0.008	0.669±0.013	0.668±0.011	0.765±0.004	0.765±0.005	0.724±0.044	0.724±0.044
CNN	0.812±0.004	0.814±0.004	0.823±0.012	0.826±0.013	0.817±0.010	0.820±0.007	0.834±0.011	0.839±0.012	0.813±0.015	0.815±0.015	0.863±0.010	0.865±0.010
CNN-LSTM	0.857±0.010	0.859±0.009	0.865±0.005	0.866±0.004	0.860±0.007	0.861±0.007	0.870±0.014	0.871±0.014	0.870±0.010	0.873±0.010	0.931±0.008	0.932±0.007

TABLE II: Results for the LOSO scheme. FT indicated fine-tuning with 5% of unseen subjects' data.

Models	COM		LFoot		LShank		RFoot		RShank		Fusion	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
Naive Bayes	0.144±0.035	0.103±0.032	0.172±0.029	0.118±0.038	0.183±0.043	0.129±0.032	0.259±0.0.118	0.214±0.134	0.287±0.169	0.262±0.191	0.209±0.027	0.165±0.043
Decision Tree	0.140±0.033	0.126±0.031	0.150±0.031	0.139±0.030	0.168±0.047	0.148±0.042	0.423±0.324	0.414±0.331	0.452±0.352	0.440±0.362	0.267±0.093	0.253±0.096
SVM	0.187±0.028	0.168±0.031	0.196±0.029	0.174±0.027	0.207±0.084	0.193±0.077	0.491±0.358	0.475±0.371	0.440±0.283	0.421±0.294	0.304±0.086	0.286±0.093
BTE	0.193±0.029	0.167±0.028	0.184±0.052	0.167±0.050	0.192±0.035	0.169±0.039	0.520±0.393	0.502±0.407	0.525±0.389	0.507±0.402	0.323±0.100	0.303±0.106
CNN	0.259±0.062	0.264±0.055	0.347±0.085	0.351±0.082	0.408±0.111	0.443±0.102	0.298±0.114	0.290±0.105	0.339±0.111	0.356±0.105	0.376±0.083	0.401±0.063
CNN (FT)	0.475±0.059	0.475±0.057	0.484±0.096	0.491±0.091	0.489±0.061	0.493±0.057	0.479±0.040	0.519±0.047	0.530±0.091	0.578±0.084	0.618±0.067	0.654±0.067
CNN-LSTM	0.265±0.044	0.276±0.054	0.365±0.084	0.372±0.091	0.374±0.095	0.378±0.105	0.293±0.029	0.301±0.031	0.337±0.094	0.360±0.091	0.385±0.045	0.404±0.038
CNN-LSTM (FT)	0.608±0.086	0.638±0.084	0.609±0.080	0.637±0.081	0.636±0.072	0.663±0.063	0.585±0.058	0.615±0.062	0.578±0.084	0.607±0.080	0.714±0.053	0.746±0.052

TABLE III: Results for fine-tuning the proposed CNN-LSTM model with LOSO.

Tuning	Acc.	F1
No Tuning	0.385±0.045	0.404±0.038
Tuning 2%	0.596±0.084	0.637±0.076
Tuning 5%	0.714±0.053	0.746±0.052
Tuning 10%	0.766±0.050	0.789±0.047
Tuning 20%	0.8067±0.041	0.825±0.040

based on the 5 wearable accelerometers. This also forms an ablated version of our model, where the BLSTM is removed. Successive to removal of the BLSTM network, the hyperparameters of the CNN were further optimized to yeild the best results. The CNN baseline consists of 4 convolution blocks with 128, 256, 384, and 512 filters, followed by 4 FC layers with 384, 200, 120, and 8 hidden neurons. For all the convolution layers, the kernel size was set to 3. The models were trained over 300 epochs with a batch size 64 and Adam optimizer with a learning rate of 0.0002.

III. RESULTS

In this section, we present the results of our experiments described in the previous section. Our proposed CNN-LSTM method outperforms all the baselines in both k-fold and LOSO schemes. Table I presents the accuracy and F1 scores for all the models with k-fold cross-validation. It can be observed that the proposed CNN-LSTM model obtains the best results achieving an accuracy and F1-score of 0.931±0.008 and 0.932±0.007 respectively. Among the baseline models, the CNN shows the best performance, outperforming the classical machine learning solutions. Exploring the performance of the model on different parts of the body (sensor locations) indicates that all the sensors provide effective information for classification of running styles. Moreover, we observe that by fusing the 5 sensors pipelines, a boost in performance is achieved, demonstrating that different parts of the body contain complimentary information that can be exploited for automatic classification of running styles. We also present the confusion matrix for this model in Figure 2, where we observe that different running styles have been classified consistently across the board. Meanwhile, the 'wide-stance' style proves the easiest to classify whereas the 'bouncing' style proves the most difficult.

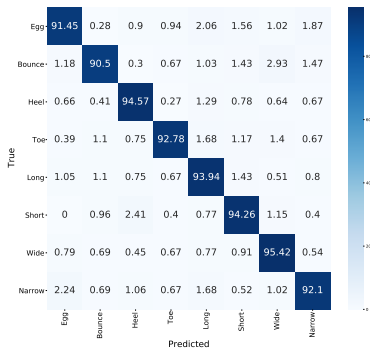The results of the models in the LOSO scheme are presented in Table II. It can be observed that while our

Fig. 2: Confusion matrix for the proposed method.

proposed CNN-LSTM model outperforms the baselines, the overall performance of all the models, both baselines and the proposed, significantly drop. It can therefore be inferred that variations in running styles are highly subject-specific, making it difficult for automatic subject-independent classification, highlighting the need for solutions capable of learning generalized representations via transfer learning or self-supervised learning. Moreover, the fact that subjects performed the running styles based on their comfort could be another contributing factor to this matter. To further evaluate this matter, we fine-tune both CNN and CNN-LSTM models with a small fraction of test data (5%). As observed in Table II, for both models, the performance is significantly improved, reaching an accuracy of 0.618 for CNN model and 0.714 for CNN-LSTM model. The results of fine-tuning the proposed CNN-LSTM model with various portions of test data are presented in Table III. As expected, increasing the amount of data used for fine-tuning improves the performance of the model, while we observe that utilizing even a very small amount of user-specific data (2%), results in a significant boost in performance.

IV. CONCLUSIONS AND FUTURE WORK

In this study, we implemented a system for classifying running style variations. For this purpose, we collected data using 5 wearable accelerometers from 10 healthy runners while performing 8 different running styles on a treadmill. We then proposed a deep CNN-LSTM architecture to detect these target classes by learning to automatically extract effective representations followed by learning temporal dependencies in the time-series. We compared the performance of our solution to a number of baseline techniques, where CNN-LSTM model showed the best results. Our findings suggest that even pre-defined running styles are personalized for different people and based on comfort and physiology, making it difficult for machine learning and deep learning methods to automatically classify them, however fine-tuning the model with a small portion of each subject's samples, significantly improved the LOSO performance.

For future work, we will explore approaches such as transfer learning, self-supervised learning, and contrastive learning, that have proven effective in enabling deep learning solutions to extract highly generalized features from input data. Moreover, domain transfer techniques may be explored to perform cross-subject domain adaptation.

REFERENCES

[1] K. Hausken. Evolutions in the physiology of skiing, skating and running in the olympics. *The Journal of Sports Medicine and Physical Fitness*, 59(7):1175–1194, 2018.
[2] D. L. Goss and M. T. Gross. A review of mechanics and injury trends among various running styles. 2012.
[3] D. L. Goss. A comparison of lower extremity joint work and initial loading rates among four different running styles. *Ph.D. Thesis, The University of North Carolina at Chapel Hill*, 2012.
[4] L. Ceyssens, R. Vanelderen, C. Barton, P. Malliaras, and B. Dingenen. Biomechanical risk factors associated with running-related injuries: a systematic review. *Sports Medicine*, 49(7):1095–1115, 2019.
[5] R. C. Blagrove, G. Howatson, and P. R. Hayes. Effects of strength training on the physiological determinants of middle-and long-distance running performance: a systematic review. *Sports Medicine*, 48(5):1117–1149, 2018.
[6] R. Pedersen, A. Vand Stokke and A. Mamen. Effects of extra load position on energy expenditure in treadmill running. *European Journal of Applied Physiology*, 102(1):27–31, 2007.
[7] J. Vorup, J. Tybirk, T. P. Gunnarsson, T. Ravnholt, S. Dalsgaard, and J. Bangsbo. Effect of speed endurance and strength training on performance, running economy and muscular adaptations in endurance-trained runners. *European Journal of Applied Physiology*, 116(7):1331–1341, 2016.
[8] Ž. Kozinc and N. Sarabon. Common running overuse injuries and prevention. *Montenegrin Journal of Sports Science and Medicine*, 6(2):67, 2017.
[9] S. P. Messier and K. A. Pittala. Etiologic factors associated with selected running injuries. *Medicine and Science in Sports and Exercise*, 20(5):501–505, 1988.
[10] B. Romera-Paredes, M. S. Aung, and N. Bianchi-Berthouze. A one-vs-one classifier ensemble with majority voting for activity recognition. In *Esann*, 2013.
[11] 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 *International Workshop on Ambient Assisted Living*, pages 216–223. Springer, 2012.
[12] E. Büber and A. M. Guvensan. Discriminative time-domain features for activity recognition on a mobile phone. In *IEEE Ninth International Conference on Intelligent Sensors, Sensor Networks and Information Processing*, pages 1–6, 2014.
[13] J. Yang. Toward physical activity diary: motion recognition using simple acceleration features with mobile phones. In *Proceedings of the 1st International Workshop on Interactive Multimedia for Consumer Electronics*, pages 1–10, 2009.
[14] J. Kulchyk and A. Etemad. Activity recognition with wearable accelerometers using deep convolutional neural network and the effect of sensor placement. In *IEEE Sensors*, pages 1–4, 2019.
[15] R. Mutegeki and D. S. Han. A cnn-lstm approach to human activity recognition. In *International Conference on Artificial Intelligence in Information and Communication*, pages 362–366, 2020.
[16] A. Saeed, T. Ozcelebi, and J. Lukkien. Multi-task self-supervised learning for human activity detection. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 3(2):1–30, 2019.
[17] H. Haresamudram, A. Beedu, V. Agrawal, P. L. Grady, I. Essa, J. Hoffman, and T. Plötz. Masked reconstruction based self-supervision for human activity recognition. In *Proceedings of the International Symposium on Wearable Computers*, pages 45–49, 2020.
[18] A. Saeed, F. D. Salim, T. Ozcelebi, and J. Lukkien. Federated self-supervised learning of multisensor representations for embedded intelligence. *IEEE Internet of Things Journal*, 8(2):1030–1040, 2020.
[19] S. Rahimi Taghanaki and A. Etemad. Self-supervised wearable-based activity recognition by learning to forecast motion. *arXiv e-prints*, pages arXiv–2010, 2020.
[20] L. Rodgers. Investigating measurement modalities for integrated use in overuse injury research. *MASc Thesis, Queen's University*, 2020.
[21] Imeasureu. https://imeasureu.com/. Accessed: 20210-03-25.