The Classification of Minor Gait Alterations Using Wearable Sensors and Deep Learning

Alexander Turner and Stephen Hayes

Abstract—Objective: This paper describes how non-invasive wearable sensors can be used in combination with deep learning to classify artificially induced gait alterations without the requirement for a medical professional or gait analyst to be present. This approach is motivated by the goal of diagnosing gait abnormalities on a symptom by symptom basis, irrespective of other neuromuscular movement disorders patients may be affected by. This could lead to improvements in treatment and offer a greater insight into movement disorders. Methods: In-shoe pressure was measured for 12 able-bodied participants, each subject to 8 artificially induced gait alterations, achieved by modifying the underside of the shoe. The data were recorded at 100 Hz over 2520 data channels and was analysed using the deep learning architecture, long term short term memory networks. Additionally, the rationale for the decision-making process of these networks was investigated. Conclusion: Long term short term memory networks are applicable to the classification of gait function. The classifications can be made using only 2 seconds of sparse data (82.0% accuracy over 96,000 instances of test data) from participants who were not part of the training set. Significance: This work provides potential for gait function to be accurately classified using non-invasive techniques, and at more regular intervals, outside of a clinical setting without the need for healthcare professionals to be present.

Index Terms—Gait abnormalities, gait diagnostics, gait alterations, deep learning, LSTM, high performance computing

I. INTRODUCTION

HUMAN gait is the process of locomotion achieved through coordinated limb movement and the controlled displacement of the individuals centre of mass. Gait is a complex dynamic process consisting of multiple interacting elements over varying time scales. Abnormalities in gait are a phenotype prevalent to multiple disorders with causes ranging from neurological disease, brain damage, physical disabilities or combinations thereof. The loss of gait function and its effect on mobility can be of significant detriment to a person's quality of life. The diagnosis and treatment of such disorders is essential to preserve or improve an individual's mobility. Abnormal gait function is often diagnosed by specialist clinicians using a combination of previous diagnoses, gait function observation, genetic data, MRI, CT and overall health.

The numerous and varying symptoms associated with different abnormalities make clinical diagnosis difficult. This can be further exacerbated as movement quality can change and symptoms may not be present at any given time. Three prominent systems used in the analysis and diagnosis of gait disorders are the Gross Motor Function Classification System

Fig. 1. The F-scan (Tekscan, Boston, USA). A Wireless telemetry system enabling freedom of movement, that uses soft, flexible sensors worn inside the shoe, connected to a data acquisition unit worn around the waist via data cables running up the legs. Examples of the data produced from this system can be seen in Figures 3 and 6.

(GMFCS) ([43], [34], [44]), the Edinburgh Visual Gait Score ([42], [32], [33]) and the Gait Profile Score ([2], [38]). Each of these diagnostic criteria have both overlapping aspects in which they correlate well, and distinct aspects where they don't ([44], [5]). Technology used to provide data for each method includes force plates, which provide ground reaction forces and centre-of gravity data, pressure sensing mats or insoles, which provide underfoot pressure profiles and somewhat invasive 3D motion capture. This typically uses the tracking of reflective or active markers adhered to the individual's body to build a model of the person's gait. However the data is required to be interpreted by trained professionals in a medical or laboratory setting and only a snap-shot of the individuals movement can be analysed.

Specific movement disorders each have their own diagnostic criteria, which to varying degrees cover abnormalities in gait and movement. Sufferers of multiple sclerosis are commonly diagnosed and treated according to the Expanded Disability Status Scale (EDSS) developed in 1983 which is currently still used ([19], [25], [27], [1], [37], [25]). Suffers of cerebral palsy are often diagnosed and treated using the GMFCS ([43], [34], [44]) and individuals with Parkinson's disease are typically treated based on clinical evaluation. There has been a recent move towards the MDS Clinical Diagnostic Criteria for Parkinson's disease (MDSPD) ([40], [39], [20]). Each of these frameworks has a different weighting and perspective on gait function and specific abnormal gait patters have their own descriptive criteria. Some of the most common of these

are hemiplegic gait, diplegic gait, neuropathic gait, myopathic gait, ataxic gait and Parkinsonian gait. ([35], [50], [31]).

Movement disorders are often diagnosed and evaluated according to the diagnostic grading criteria associated with the primary diagnosis of the individual ([44], [5]). This work aims to provide a proof of concept demonstrating that gait alterations and abnormalities can be diagnosed on a symptom by symptom basis, non-invasively, without being in the presence of a clinician or the requirement to be in a healthcare setting. To achieve this we used a combination of two technologies. The first is the Fscan (Tekscan, Boston, USA) in-shoe pressure sensing system (Figure 1) The second technology was deep learning and more specifically, long term short term memory networks (LSTMs). Deep learning techniques have demonstrated excellent results in pattern matching in image classification, signal processing and feature extraction. LSTMs are a deep learning architecture inspired by recurrent neural networks which have proved to be particularly adept at sequence to sequence learning and pattern recognition. LSTMs are capable of learning representations of information over both short and long periods of time, which are well suited to the classification of complex dynamic data sets spanning differing time scales. By using the raw unprocessed data, the LSTM will generate its own understanding of gait function without pre-processing or data reduction.

A. Related Work

There has been a recent interest in wearable sensors to promote diagnostics and treatment in healthcare settings with a focus on understanding gait. However, limited work has incorporated machine learning to facilitate the interpretation of the data collected. A recent systematic review focused on gait analysis across multiple conditions ranging from ankle fractures to Parkinson's disease and cerebral palsy using different wearable sensors [7]. They identified accelerometers and/or gyroscopic measures as the most common wearable technology. Machine learning was rarely used. When machine learning was used, its purpose was to discern a characteristic about gait such as speed rather than diagnostics. Patton et. al., [36] used the F-scan and its software to process the raw data and identify four key characteristics. These four criteria were then used to make a diagnosis pertaining to the risk of ulceration of diabetic patients. Zequera and Solomonidis [54] used a similar method , however the analysis of the data was much more sparse, only focusing on the median pressure over ten arbitrarily created areas of the foot. Strohrmann et. al., [48] used a custom built sensing insole to analyse the movements of children with cerebral palsy, machine learning was implemented in the form of support vector machines to classify centre of pressure (COP) trajectories according to the Edinburgh Visual Gait Score.

Deep neural networks are becoming more relevant in heath informatics [41]. Inertial measurement units have been used to quantify and classify gait temporal-spatial parameters [29]. Deep learning has been successfully used in the detection of freezing gait in people with Parkinson's disease [6], and deep neural networks have been used to differentiate between individuals based on gait pattern using both video footage [23] and data from body mounted sensors [10].

Smartphone technology combined with deep learning has been used to provide a pre-clinical Parkinson's disease diagnosis [51]. More importantly, it has been demonstrated that the use of deep learning generated the highest level of accuracy when compared to a range of classification algorithms. These findings were echoed in Staamate et. al., [46] where it was reported that the use of deep learning generated the highest level of accuracy when compared to a range of classification algorithms. With an ever-aging population who require increasing medical assistance, this work demonstrates the significant potential for diagnostics and elements of treatment to move from clinical settings to the home, reducing costs and the burden on both practitioners and patients.

This work differs from previous research in the following ways. Firstly, its focus was to detect minor gait abnormalities, not specific to a particular condition and are not tied to a specific gait rating system as previously discussed. Secondly, the use of LSTMs using raw, high volume, high throughput channel data to identify differences in gait with no pre-processing, allowing the LSTMs to generate their own representations of the data. This data was acquired using non-invasive insole sensors which can be worn for long periods of time. These sensors do not need to be accurately placed and can provide data at the user's convenience as long as ambulating on a flat surface. This work proposes a method of gait analysis which is accurate, robust and can be used to analyse gait function 'on the fly', at a user's convenience, outside a gait assessment center and without the need for data processing.

We compared the results of the LSTMs with another deep learning architecture, convolutional neural networks (CNNs) which are adept at object recognition and feature extraction ([28], [21]). This was primarily as an objective comparison between the two technologies, however it also provides evidence as to whether there were artifacts or features within the data which could be exploited by CNNs to produce high classification results at the expense of clinical relevance. There are three objectives for this work:

- To provide a proof of principle that non-invasive wearable technologies in combination with deep learning can be used to reliably detect gait alterations without the need for the patient to be in a clinical setting.
- To understand if increasing the amount of data available to the deep learning architectures from single frame data to multiple data frames (Figure 2) improves classification accuracy.
- Analyse the networks to provide a rationale for their decision making process, and to evidence the clinical relevance of the classifiers and data in this work.

II. METHODS

Twelve able-bodied participants (21-34 years, 69-90 kg, 6-11 UK shoe size) were asked to complete eight walking trials around a figure of 8 walk-way (40 m in length) for the duration of 60 seconds per trial. Upon arrival to the
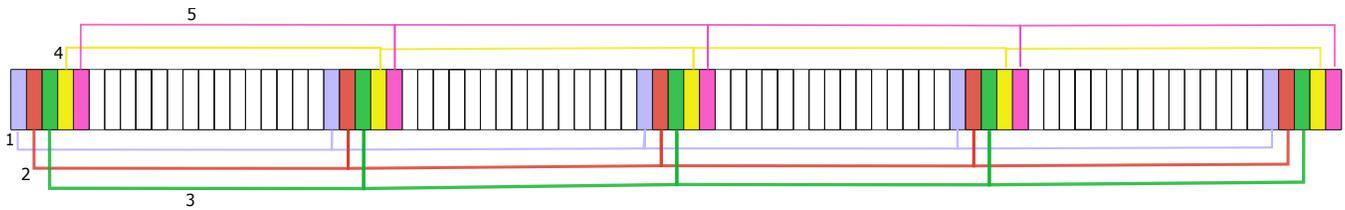
Fig. 2. A representation of 1 second of data sampled at 100 Hz and how this data was used to create an instance of training data. Five frames of data used with a span of 20 frames, thus sparsely capturing 1 second of data. In this image, five successive instances of data are shown. This is the pattern in which the networks from Figure 9 and 10 use.

laboratory participants provided written informed consent prior to any testing. Participant mass was recorded to be used in the calibration process of the F-Scan system (Tekscan, Boston, USA) All participants wore standardised trainers, provided by the authors for all trials (size specific) to ensure consistency. Pressure sensing insoles were placed inside the shoes and were connected to the waist worn data acquisition unit via cables running to the sensor connectors affixed to the participants legs just superior to the ankles. A simple calibration process was carried out consisting of the participant standing on one foot for a pre-defined time whilst the system sampled the pressure under the foot and normalised it to the participants body weight to ensure appropriate scaling. This insured that the LSTMs were less able to learn traits of specific participants and thus limited the potential for over fitting. The F-Scan is a wireless telemetry device and was set to capture data at 100 Hz over 1260 channels per insole resulting in 252,000 data samples per second across both feet (Figure 3). The data was transmitted in real time, wirelessly to a laptop computer.

The eight perturbation conditions (PCs) that participants were asked to complete were designed to try and alter the gait characteristics by changing the movement and location of the COP under the foot. To achieve this the base of the shoe was fitted with a series of soft rubberised pads. The reasons for this were twofold. Firstly, the pad material allowed for a shift in the COP without introducing a significant hard artefact

which could be easily detected within the shoe resulting in a non-clinically relevant classification. Secondly the ease in which the pads could be applied to the shoe in a fixed location enabled a quick change between conditions without causing any discomfort to the participants. Each pad had dimensions of 3.5 cm ^{2} and a depth of 1.5 cm. Under a mass of 20 kg, an individual pad's depth reduced by 0.4 cm. The sole of the shoe was divided into three sections and a pad was affixed to a section or not via an adhesive, the patterns of which are presented in Figure 4.
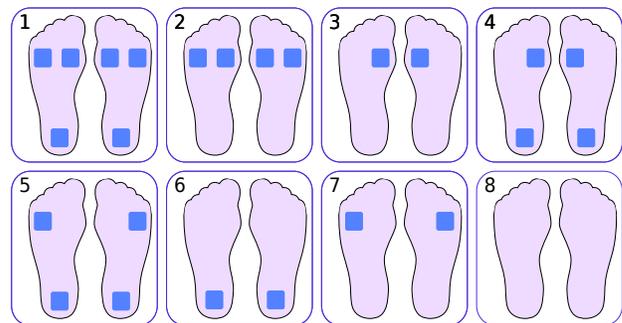

Fig. 4. Under shoe rubberised pad configurations for perturbation conditions 1-8. Each pad was a 3.5 cm^2, with a depth of 1.5cm.

A. Deep Learning

Long term short term memory networks (LSTMs) are specifically designed to extract features from time-series data and are particularly well suited to the data in this work. The LSTMs were compared to convolutional neural networks (CNNs). CNNs are particularly adept at hierarchical feature extraction in images which typically represent static data (that is all the information is held within a single image).

We use two deep learning architectures for two reasons. Firstly, as an objective comparison to determine which is more suitable for the type of data available in this work. Secondly, we wanted to avoid the perturbations producing a specific pressure profile which could be detected by the insole sensing system. This would result in a classifier being produced with little clinical relevance. With CNNs being particularly adept at feature extraction in images, if this was present, CNNs would have a high likelihood of detecting it and exploiting this to achieve high classification results. This provided evidence for evaluation of the integrity of the data as well as the comparative performance of the different architectures.
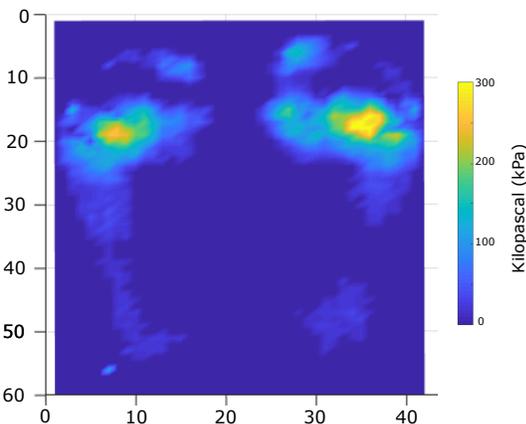

Fig. 3. A representation of the data from two individual footprints from the same participant using the F-scan. Each footprint contains 2520 (42*60) individual pressure points, data was sampled at 100 Hz.
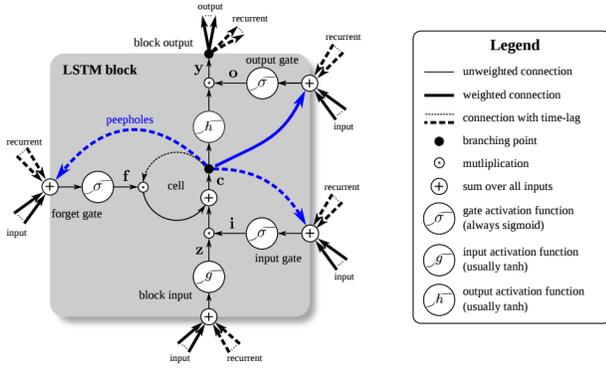
Fig. 5. Representation of the LSTM cell (adapted from Greff et al., [12]). The LSTM cell takes an input and stores it for an amount of time. The input gate controls how new values flow into the cell, the forget gate controls how a value remains in a cell and the output gate controls to what extent a value in the cell is used to compute the output activation of the LSTM unit. Each of these gaits has an activation function associated with it. The weight of each cell is then optimised during the training process.

1) Long Term Short Term Memory Networks: LSTMs are a deep learning architecture based upon a RNNs and were first proposed in [11]. LSTM networks differ from RNNs as they contain LSTM units. LSTM units typically comprise of a cell, an input gate, an output gate and a forget gate (Figure 5). They were specifically developed to improve the classification and modelling of complex time series data. One of the main features of LSTMs and their ability to work with complex time series data is that they are not affected by problems more typically attributed to RNNs, such as gradient vanishing. As such they have been able to push the state of the art in a range of problems such as natural language modelling ([55], [52], [18]) and time series forecasting ([17], [24]). These problems all contain complex dynamics, many exist over a range of time scales and all contain high volume, high throughput multi-channel data.

LSTMs have been specifically applied to a range of time-series devised, diagnostic classification tasks ([15], [47], [22]). They are well suited for multi-label classification ([53], [22]) which will be used in this work. We focus on an LSTM network which contains a bi-directional layer allowing for the input sequence to be given to the network both in its original format and in a reversed format. This has generally resulted in improvements when classifying time series data over unidirectional LSTMs ([11], [9]). The architecture of the network used in this work can be seen in Table 1.

2) Convolutional neural networks: CNNs are a deep learning architecture which were specifically built for object recognition and can learn a hierarchy of features. CNNs work by detecting low level features in images, building them up into higher level representations by taking inspiration from the visual structures and processes within the human brain in particular receptive fields. CNNs are currently the superior technology in object recognition and are widely used in the medical domain where they have demonstrated strong results ([28], [21]).

TABLE I
THE ARCHITECTURE OF THE LSTM USED IN THIS WORK. THE bi-LSTM LAYER CHANGES BETWEEN EXPERIMENTS, AND CAN CONTAIN BETWEEN 40 AND 400 NODES WITH 40 NODE INTERVALS. ALL OTHER LAYERS WILL REMAIN THE SAME DURING EXPERIMENTATION.

Number	Type	Description
1	Sequence input	Input length 42
2	Dropout layer	50% probability of dropout
3	Bi-LSTM layer	Variable number of units (40-400)
4	Fully connected layer	Fully connected layer
5	Softmax layer	Softmax layer
6	Classification output layer	Classification output layer

B. Experimental Design

Due to the nature of this study there was a strong potential for over fitting, where a machine learning architecture learns a data set too specifically and the rules it creates cannot be generalised to a new scenario. Overfitting could occur if the neural networks learnt the gait of a specific individual, as a result sensor drift over time could loosely correlate with a particular PC. Another possibility would be for the neural networks to focus on the pressure readings from a very small section of the insole or a sensor artifact. These sections could differ between participants, allowing the neural networks to produce rules which could not be generalised. To combat this, we used a test data set to evaluate the performance of the neural networks containing data from participants who were not involved in training. The training set contained 60 seconds of walking data for each PC from ten of the twelve participants, and the test set contained 60 seconds of data for each PC from the remaining two participants.

Deep neural networks have different levels of complexity where more complex networks are generally more capable of complex decision making, however they are more costly to train in terms of time. There are two general parameters which are a measure of any given neural networks architectural complexity. The first is the number of layers in the network. In this work, we will focus on a single layered network (layers in this sense only refer to the LSTM / biLSTM layers). The second measure of complexity is the number of nodes in a layer. To explore this concept further, and to ascertain the type of architecture which best suits the data collected, we conducted multiple runs of each experiment, with varying levels of nodes in the LSTM layer (Table 1). The numbers of nodes varied between 40 and 400, at 40 node intervals, meaning that 10 different runs were conducted for each experiment.

The CNN contained two convolutional layers with a [20,50] node layer in the first and a [5,10] layer in the second. There were two ReLU layers, one between the convolutional layers and the other between the second convolutional layer and the softmax layer. The softmax layer connects directly to the classification layer. All of the networks contained a single dropout layer in-between the input layer and the first layer. The purpose of this was to randomly remove input data during training to ensure that classification could not be dependent on a single data channel, and to encourage data representations which could be well generalised and not overfit. A mini batch
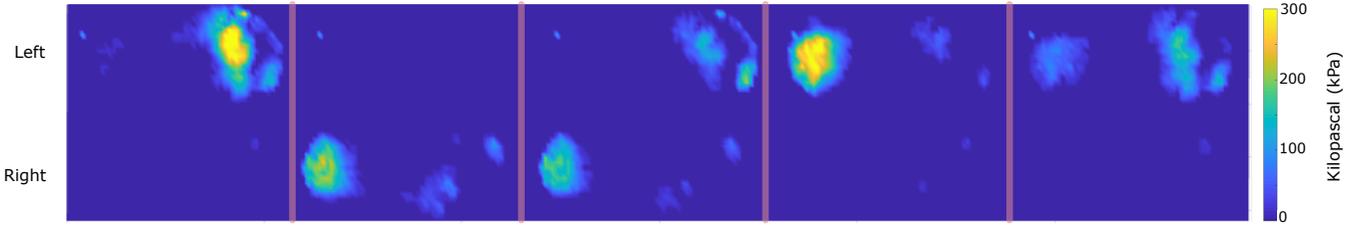
Fig. 6. A graphical representation of one instance of training data spanning 1 second with intervals of 200 milliseconds used in the second experiment.This provides a sparse representation of gait over 1 second, and does not include sequentially successive data which would be more homogeneous. This both captures a wide range of data, without introducing similar data (data which in very close in sequence is likely to more similar). This is the data which will be used to understand if the LSTM architectures can classify gait alterations using dynamic data. An individual instance of data can be seen in Figure 3.

size of 64 samples was used alongside L2Regularization of 0.001 and a learning rate of 0.01 for both architectures.

In total, 3 different experiments were conducted :

- The first experiment used data comprised of single time frames (Figure 3) for training and evaluating the networks. The primary benefit of this was that training of the network was relatively fast. It was likely that this approach would not provide enough information to accurately classify the PCs as it did not capture dynamic data.
- The second experiment used data consisting of five time frames spliced together (Figures 2 and 6) for both training and evaluating the networks. The data had 200 millisecond gaps between each frame, thus sparsely capturing one second worth of data. This data captured the dynamic nature of gait by looking at the data over multiple successive frames.
- The third experiment used the same networks developed in the five frame experiment however, the networks are evaluated with ten frames of data with 200 millisecond gaps between each frame, sparsely capturing two seconds of data.

III. RESULTS AND ANALYSIS

A. Single Frame Data

The number of nodes in the network had minimal effect on how it learnt to classify data. The highest level of accuracy was achieved with the network containing 400 nodes (40.9%). This was only marginally greater than when the network consisted of 200 nodes (40.7%) (Figure 7). The confusion matrix (Figure 8) presents data for the LSTM with the highest accuracy rating (40.9%) representing how the network classified data and more specifically how the accuracy related to specific PCs. Condition 7 is the most accurately classified PC (6373 correct classifications) this equates to 6.6% of the total data. PC 2 has the lowest classification accuracy (2488 correct classifications), amounting to only 2.6% of the total data. The optimum classification rate is 12.5% as there are eight different perturbation conditions. Only gait condition 7 was correctly classified over 50% of the time using this LSTM, all other conditions were classified incorrectly more often than correctly.
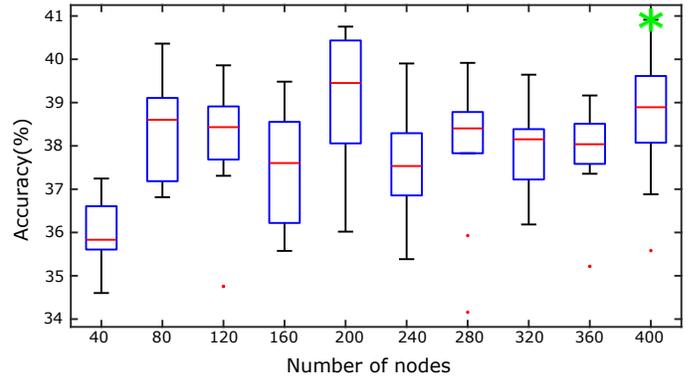

Fig. 7. The results of how well the LSTM can classify the gait alterations when trained using a single frame of data (Figure 3) over a range of LSTM topologies ranging from 40 - 400 nodes. This data shows what percentage of the test set was classified accurately. The best result for the unseen data is 40.9% and is achieved with an LSTM containing 400 nodes. This score is highlighted, and the confusion matrix for this particular network can be seen in Figure 8. When using a single instance of data the variation of nodes in the LSTM networks makes very little consistent difference to their performance. The results at each epoch are plotted for each number of nodes

B. Five Frame Data

The results presented in Figure 9 show the accuracy of data classification which substantially improved relative to the experiment using single frame data. The highest classification accuracy using five frames of data was 76.9% and was achieved by the LSTM containing 240 nodes. This represents an increase in classification accuracy of 36% relative to the highest level of accuracy achieved in the single frame data.

Figure 10 illustrates that the correct classification was predicted by the network for every PC more frequently than any other classification. The lowest classification accuracy was 55.9% (PC 6) and the highest was 93.4% (PC 5). In total PC 5 was only misclassified 783 times of 11800 instances. This suggests that by using the full 60 seconds of the test data set, the correct classification would be found for each PC. The most mistakes in classification made by this LSTM were made when PC 7 was misclassified as PC 5. This occurred 2906 times. Certain PCs were never misclassified as other PCs, for example, PC 1 was never misclassified as PC 3 or 7.

Fig. 8. The confusion matrix for the best network when using a single frame of foot data. The exact network this matrix is drawn from is highlighted in Figure 7 using the test data. The overall best classification accuracy for this network is 40.9%, and a network which made random guesses would achieve 12.5%, which indicates that some valid representations have been made within the LSTMs. The green diagonal squares indicate guesses which have been made correctly. The correct class is most frequently predicted by the network for all of the conditions. That is, the most selected classification for each perturbation by the LSTM was the correct one overall. The most accurately guessed perturbation was number 7, which was correctly guessed 53.1% of the time.
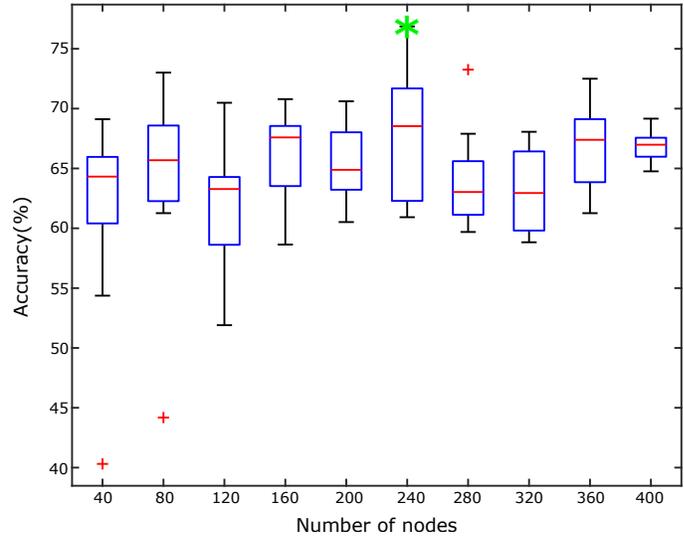

Fig. 9. The results of how well the LSTMs can classify the gait alterations when trained using 1 second with 200 millisecond intervals (Figure 2 - five instances of data stitched together). This was conducted over range of LSTM topologies with between 40 - 400 nodes. Each of the accuracy figures are taken at each epoch and are the accuracy of the LSTM when classifying the test data set. The best accuracy achieved is 76.9% and is achieved with an LSTM containing 240 nodes (highlighted by the green star), suggesting that accurate representations have been made by the LSTMs which can classify the test data. The confusion matrix for this particular network can be seen in Figure 10. The changes in the number of nodes have a slightly more pronounced effect on its performance when compared to the single frame data from Figure 7. Additionally, the results when training the networks on one second of data are significantly better than using a single instance (Figure 7) of data (p = 3.57 e-14).

C. Ten Frame Data

The networks used for the ten frames of data experiment were the same as those trained in the five frames of data tests (Figure 9). The increase in data length from five to ten frames led to a uniform increase in classification accuracy of 4.7% to 7.5%. The highest level of classification accuracy achieved using two seconds of test data was 82.0% and was produced by the same LSTM which achieved the highest accuracy level in the five frame experiment (Figure 9).

The confusion matrix for this network (Figure 11) shows that the correct classification rates for each PC improved, except for PC 6 when compared to the results presented in Figure 10. Additionally the number of categories containing misclassification percentages of < 0.1% improved; 33/64 when using two seconds of data compared to 26/64 when using a single second of data. Overall, by increasing the length of the input data it can be seen that for almost all metrics, the classification performance of the LSTM improved to what is the maximum achieved in this study of 82.0% (Figure 11).

D. Convolutional Neural Networks

The objective performance of CNNs for single frame data was comparatively poor compared to that of LSTM's, achieving a best result of 24.0% (Figure 12) this was lower than any of the results presented by the LSTM architectures, the lowest being 34.6% (Figure 7). For the five frame data experiment,

Fig. 10. The confusion matrix for the best network when using a 5 frames of data over a full second. The exact network this matrix is drawn from is highlighted in Figure 9 using the test data. The overall classification accuracy for this network is 76.9% as seen in the bottom right hand corner, which is significantly better than the networks which only use a single frame of data. When looking at the green diagonal, showing the correct guesses, it can be seen that for every perturbation, the correct one is guessed correctly over the majority of the test data. The bottom of the x-axis shows that for the majority of perturbations, all are correctly guessed over 50% of the time.

Fig. 11. The confusion matrix for the best network which was trained using 5 frames of data over a full second, but evaluated here using 10 frames of test data over 2 seconds. The overall classification accuracy for this network is 82.0% as seen in the bottom right hand corner of the matrix, and represents the best classification accuracy achieved in this work.

Fig. 12. The confusion matrix for the best CNN when trained using a single frame of data. The overall classification accuracy for this network is 24.0% as seen in the bottom right hand corner of the matrix, and is markedly worse than any of the results achieved by the LSTMs for this task (Figure 7).

the CNNs were unable to discriminate between any of the PCs and were unable to achieve a performance better than a random classifier at 12.5%. It is possible that with more computational power and more complex networks this could improve, however compared to an LSTM which took similar overall resources to train, the CNN performed poorly.

IV. DICSUSSION

This paper investigated if high throughput, high volume data could be used to detect artificially induced perturbations of gait using LSTM networks. The primary aim of this research was to establish a non-invasive methodology that could enable the classification of gait abnormalities on a symptom by symptom basis, using unprocessed dynamic gait data. This concept was motivated by the desire to use data collected in real life situations rather than snapshot data identified in limited clinic time and to alleviate the inconsistencies with how gait abnormalities are diagnosed and treated in patients with movement disorders. The results of this work show that LSTMs are well suited to classifying artificially induced gait alterations suggesting that these methods could be used to classify real gait abnormalities.

A. Patterns of classification

By looking at the confusion matrices presented throughout this work, it is apparent that the best LSTMs from each experiment formed patterns pertaining to how each PC was classified. The confusion matrices seen in the five and ten frame data experiments (Figures 10 and 11) clearly show that some categories produce no misclassifications. If we look at the highest performing LSTM (Figure 11), it can be seen that PC 7 is very rarely misclassified. PC 7 was never misclassified as PC 1, 2, 3, 4 or 8 and rarely as PC 6 ($<0.1\%$). However, the second most common misclassification this network produced was when PC 7 was misclassified as PC 5 (1799 times) suggesting that PCs 5 and 7 were similar. Further emphasising this, PC 5 was never misclassified as PC 2, 3 4, 6 or 8 and rarely as PC 1 ($<0.1\%$). Figure 4 shows that the PCs 5 and 7 are the only two conditions that were likely to cause over pronation based on perturbation pads being located under the lateral border of the forefoot.

The category most often misclassified in Figure 11 was PC 6 when it was misclassified as PC 2 (1956 times). By linking this information back to the perturbation pad locations (Figure 4), this misclassification initially appears to be surprising as the locations of the pads suggest that movement of the COP under the foot should react very differently. However, when considering the natural progression of the COP in typical walking it is evident that the location of these perturbations pads should not make a radical difference. Initial contact at the heel would be evident in both PCs followed by a transition of the COP along the lateral border of the foot until reaching the forefoot, where the COP shifts medially across the metatarsal heads until toe off (pronation) [13]. The perturbation pads in PC2 may delay the transition of force from mid to forefoot but would be unlikely to alter the spatial component of force transition. Finally it is evident that PCs 3 and 7, which should generate different movements, supination and (over) pronation respectively are very rarely misclassified as each other (5 times of a possible 23,200 classifications). All of this suggests that the PCs generating the greatest differences are, as expected, the easiest for the network to distinguish. These manifest in the PCs that generate medio-lateral underfoot pressure alterations rather than anterior-posterior changes as every step still requires a posterior to anterior transition of the COP.

This grouping of classifications by medio-lateral rather than anterior-posterior rather than similarity of location of pads

adds evidence that the alterations induced in this work were indeed modifying gait rather than introducing artifacts into gait which can be easily identified by the insole sensing system.

B. Rationale of the decision making process

Nguyen et, al,. [30] demonstrated that although deep neural networks objectively perform well, their underlying representations of the problem domain are often non-intuitive and can lead to problems in understanding the rationale behind the decision making process of the models. To provide a level of validation of the models created in this work other than the objective performance, we tried to understand what data the LSTMs were using to make their decisions. To achieve this we started with the five frame data (Figures 6 and 13) that was classified correctly as a specific PC (Figure 4) with 99.9% confidence. We iteratively set random elements of the five frame data to 0. If this modification did not lower the confidence value of the classification, the modification remained. If the confidence value reduced below 99.9%, the modification was removed. This process was repeated 20,000 times to produce a minimal representation of what the LSTM required to confidently assign the correct PC to the data. This was completed for 10 different, correctly classified data samples, and 10 incorrectly classified data examples. Figure 13 is a representative example of the results for a correctly classified data sample. The LSTM used to generate the classifications was the best performing LSTM found in this work (Figures 10 and 11).

The upper image in Figure 13 shows the pressure readings for each footstep, with high pressure readings denoting foot strike. There appear to be artifacts present for both feet, where small instances of high pressure were recorded but don't appear to be part of a foot strike. When the data removal algorithm was applied, a large amount of the data was removed (lower image in Figure 13) and both images have a 99.9% confidence rating according to the LSTM. The data which remains was in all cases very sparse, with 39 non-zero data points being present in the lower image compared to an average of 659 in the upper image (Figure 13). These findings were consistent throughout, and are representative when the data is incorrectly classified, suggesting that the underlying behavior of the LSTMs did not significantly change depending on which data it is provided with. Although it is not possible to demonstrate the exact rationale behind the LSTMs decision making process, this work provided evidence to its method. The majority of the data used in the images was not essential for the LSTMs to provide an accurate classification. It appears that the LSTMs look for relatively sparse points in the data from high pressure regions, and uses their location in reference to other points in order to make a decision. It is this representation of the data which has been able to prescribe the accuracy seen in this paper.

The first objective of this work was to provide a proof of concept as to whether wearable sensors and deep learning could be used to detect gait alterations. The highest level of classification accuracy achieved was 82.0% over 96,000 instances of test data containing participants which were not part of the training data (Figure 11). This provided robust evidence that rules generated by the LSTM could be generalised to new individuals. CNNs were objectively worse than LSTMs in this work and for the five frame data performance no better than random chance, highlighting that the LSTMs are better suited to this type of data.

A secondary aspect of this work focused on identifying the optimum type of data the LSTM could use to classify the data correctly. This issue was considered from two perspectives. Firstly how many time frames from the source data should be used to create a data instance for the LSTMs. Secondly whether using five or ten frames of data for evaluation affected the accuracy of classification. When five frames of data of data were used (Figure 2) the classification performance rose significantly compared to when using a single frame of data (Figure 3). The classification accuracy improved significantly again, when the LSTM was trained on five frame data but was evaluated on ten frame data. This suggests that in both instances, data of greater length is beneficial as it allows more accurate classification, however this must be balanced with computational training times, which increase significantly when increasing the length of the data.

The third objective was to analyse the decision making process and to support the clinical relevance of the data in this work. By analysing the data provided to the networks and removing elements which do not affect its classification, it was found that in all instances, very little of the available data was used by the LSTMs to provide a classification. It is likely that specific areas and positions within high pressure locations of the data and the distances between them were being used by the LSTMs to determine a classifications. The comparatively poor performance of CNNs overall suggests that there are no artifacts in the data which CNNs can exploit to produce non-clinically relevant classifications, and thus supports the suitability of the data collected and the LSTMs in this work, and provides a proof of principle for this work which can be translated to a clinical setting. It is therefore more likely that the pads are producing more complex disruptions in normal gait function which produce features are not easily discriminated.

V. CONCLUSION

The results presented in this paper show the potential of using non-invasive devices for the diagnosis of movement disorders without the need to visit a medical specialist. Deep learning architectures used to interpret the data were revealed to be accurate, capable of distinguishing between different artificially induced gait alterations and robust enough to cope with data from participants for which it had no prior experience. This work demonstrates the capacity for gait abnormalities to be diagnosed on a symptom by symptom basis. These methods could be further applied to help guide the treatment for sufferers of movement disorders.

REFERENCES

[1] Ricardo Alonso, Maria Barbara Eizaguirre, Berenice Silva, Cecilia Pita, Sandra Vanotti, Fernando Caceres, and Domingo Garcea. Usefulness of the unification of the criterion for the evaluation of the functional system
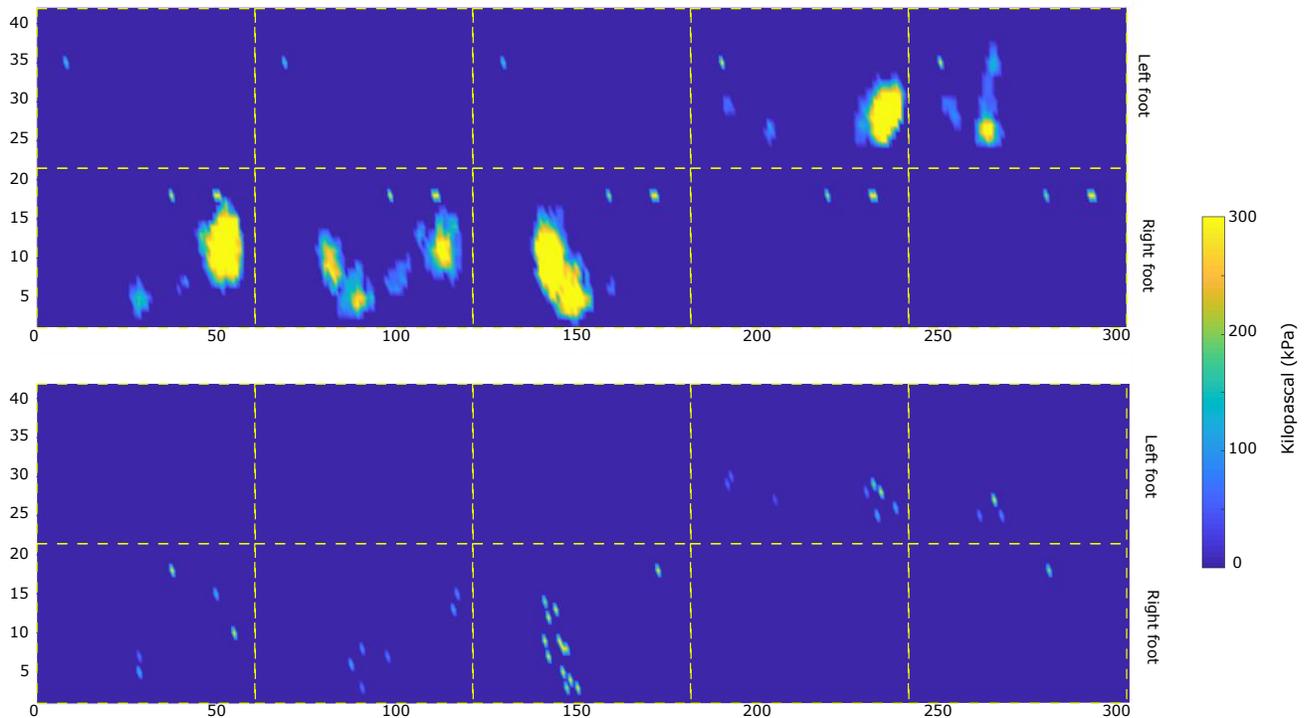
Fig. 13. This instance of data (upper image) produced a 99.9% confidence that it was condition 5 (Figure 4). Each individual footprint is highlighted with the yellow dotted lines. We iterated through this data and randomly set the values in this image to 0 on the condition that the 99.9% confidence figure didn't get lower. That is, we removed the data in the image iteratively as long as it made no significant difference to how it was classified. We did 20,000 times to the image, and removed set 620 values to 0 without changing the classification of the image. The lower image illustrates what representations of the data the LSTM used to make the classification, with the surplus data removed. The LSTM used to derive these classification was the one which achieved the highest accuracy in this work (Figure 10)

cerebral functions of the expanded disability status scale (edss)(p4. 389), 2018.

[2] Richard Baker, Jennifer L McGinley, Michael H Schwartz, Sarah Beynon, Adam Rozumalski, H Kerr Graham, and Oren Tirosh. The gait profile score and movement analysis profile. *Gait & posture*, 30(3):265–269, 2009.

[3] Richard Baker, Jennifer L McGinley, Mike Schwartz, Pam Thomason, Jill Rodda, and H Kerr Graham. The minimal clinically important difference for the gait profile score. *Gait & posture*, 35(4):612–615, 2012.

[4] Susan Barker, Rebecca Craik, William Freedman, Nira Herrmann, and Howard Hillstrom. Accuracy, reliability, and validity of a spatiotemporal gait analysis system. *Medical engineering & physics*, 28(5):460–467, 2006.

[5] Geruza P Bella, Nádia BB Rodrigues, Paola J Valenciano, Luciana MAE Silva, and Regina CT Souza. Correlation among the visual gait assessment scale, edinburgh visual gait scale and observational gait scale in children with spastic diplegic cerebral palsy. *Brazilian Journal of Physical Therapy*, 16(2):134–140, 2012.

[6] Julià Camps, Albert Samà, Mario Martín, Daniel Rodríguez-Martín, Carlos Pérez-López, Sheila Alcaine, Berta Mestre, Anna Prats, M Cruz Crespo, Joan Cabestany, et al. Deep learning for detecting freezing of gait episodes in parkinson's disease based on accelerometers. In *International Work-Conference on Artificial Neural Networks*, pages 344–355. Springer, 2017.

[7] Shanshan Chen, John Lach, Benny Lo, and Guang-Zhong Yang. Toward pervasive gait analysis with wearable sensors: A systematic review. *IEEE J. Biomedical and Health Informatics*, 20(6):1521–1537, 2016.

[8] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115, 2017.

[9] Xiaocheng Feng, Bing Qin, and Ting Liu. A language-independent neural network for event detection. *Science China Information Sciences*, 61(9):092106, 2018.

[10] Giacomo Giorgi, Fabio Martinelli, Andrea Saracino, and Mina Sheikhal-ishahi. Try walking in my shoes, if you can: Accurate gait recognition through deep learning. In *International Conference on Computer Safety, Reliability, and Security*, pages 384–395. Springer, 2017.

[11] Alex Graves and Jürgen Schmidhuber. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural Networks*, 18(5-6):602–610, 2005.

[12] Klaus Greff, Rupesh K Srivastava, Jan Koutník, Bas R Steunebrink, and Jürgen Schmidhuber. Lstm: A search space odyssey. *IEEE transactions on neural networks and learning systems*, 28(10):2222–2232, 2017.

[13] Michael Grundy, PA Tosh, RD McLeish, and L Smidt. An investigation of the centres of pressure under the foot while walking. *The Journal of bone and joint surgery. British volume*, 57(1):98–103, 1975.

[14] Mohammed Mehedi Hassan, Md Zia Uddin, Amr Mohamed, and Ahmad Almogren. A robust human activity recognition system using smartphone sensors and deep learning. *Future Generation Computer Systems*, 81:307–313, 2018.

[15] Katharine E Henry, David N Hager, Peter J Pronovost, and Suchi Saria. A targeted real-time early warning score (trewscore) for septic shock. *Science translational medicine*, 7(299):299ra122–299ra122, 2015.

[16] PJ Jino, Jomy John, and Kannan Balakrishnan. Offline handwritten malayalam character recognition using stacked lstm. In *Intelligent Computing, Instrumentation and Control Technologies (ICICICT), 2017 International Conference on*, pages 1587–1590. IEEE, 2017.

[17] Fazle Karim, Somshubra Majumdar, Houshang Darabi, and Shun Chen. Lstm fully convolutional networks for time series classification. *IEEE Access*, 6:1662–1669, 2018.

[18] Ankit Kumar, Ozan Irsoy, Peter Ondruska, Mohit Iyyer, James Bradbury, Ishaan Gulrajani, Victor Zhong, Romain Paulus, and Richard Socher. Ask me anything: Dynamic memory networks for natural language processing. In *International Conference on Machine Learning*, pages 1378–1387, 2016.

[19] John F Kurtzke. Rating neurologic impairment in multiple sclerosis: an expanded disability status scale (edss). *Neurology*, 33(11):1444–1444, 1983.

[20] Jun Li, Miao Jin, Li Wang, Bin Qin, and Kang Wang. Mds clinical diagnostic criteria for parkinson's disease in china. *Journal of neurology*, 264(3):476–481, 2017.

[21] Qing Li, Weidong Cai, Xiaogang Wang, Yun Zhou, David Dagan Feng, and Mei Chen. Medical image classification with convolutional neural network. In *Control Automation Robotics & Vision (ICARCV), 2014 13th International Conference on*, pages 844–848. IEEE, 2014.

[22] Zachary C Lipton, David C Kale, Charles Elkan, and Randall Wetzel. Learning to diagnose with lstm recurrent neural networks. *arXiv preprint arXiv:1511.03677*, 2015.

[23] Wu Liu, Cheng Zhang, Huadong Ma, and Shuangqun Li. Learning efficient spatial-temporal gait features with deep learning for human identification. *Neuroinformatics*, pages 1–15, 2018.

[24] Pankaj Malhotra, Lovekesh Vig, Gautam Shroff, and Puneet Agarwal. Long short term memory networks for anomaly detection in time series. In *Proceedings*, page 89. Presses universitaires de Louvain, 2015.

[25] Ryan S McGinnis, Nikhil Mahadevan, Yaejin Moon, Kirsten Seagers, Nirav Sheth, John A Wright Jr, Steven DiCristofaro, Ikaro Silva, Elise Jortberg, Melissa Ceruolo, et al. A machine learning approach for gait speed estimation using skin-mounted wearable sensors: From healthy controls to individuals with multiple sclerosis. *PloS one*, 12(6):e0178366, 2017.

[26] Ronaldo Messina and Jerome Louradour. Segmentation-free handwritten chinese text recognition with lstm-rnn. In *Document Analysis and Recognition (ICDAR), 2015 13th International Conference on*, pages 171–175. IEEE, 2015.

[27] Sandra Meyer-Moock, You-Shan Feng, Mathias Maeurer, Franz-Werner Dippel, and Thomas Kohlmann. Systematic literature review and validity evaluation of the expanded disability status scale (edss) and the multiple sclerosis functional composite (msfc) in patients with multiple sclerosis. *BMC neurology*, 14(1):58, 2014.

[28] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *3D Vision (3DV), 2016 Fourth International Conference on*, pages 565–571. IEEE, 2016.

[29] Kyung-Ryoul Mun, Gyuwon Song, Sungkuk Chun, and Jinwook Kim. Gait estimation from anatomical foot parameters measured by a foot feature measurement system using a deep neural network model. *Scientific reports*, 8(1):9879, 2018.

[30] Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 427–436, 2015.

[31] Jorik Nonnekes, Rianne JM Goselink, Ruzicka, Alfonso Fasano, John G Nutt, and Bastiaan R Bloem. Neurological disorders of gait, balance and posture: a sign-based approach. *Nature Reviews Neurology*, 14(3):183, 2018.

[32] AML Ong, SJ Hillman, and JE Robb. Reliability and validity of the edinburgh visual gait score for cerebral palsy when used by inexperienced observers. *Gait and posture*, 28(2):323–326, 2008.

[33] Maria del Pilar Duque Orozco, Oussama Abousamra, Chris Church, Nancy Lennon, John Henley, Kenneth J Rogers, Julieanne P Sees, Justin Connor, and Freeman Miller. Reliability and validity of edinburgh visual gait score as an evaluation tool for children with cerebral palsy. *Gait and posture*, 49:14–18, 2016.

[34] Robert Palisano, Peter Rosenbaum, Stephen Walter, Dianne Russell, Ellen Wood, and Barbara Galuppi. Development and reliability of a system to classify gross motor function in children with cerebral palsy. *Developmental Medicine & Child Neurology*, 39(4):214–223, 1997.

[35] Jyoti Patil, Deeksha Nandur, Meghana Mellikeri, Kritika Naik, and Pallavi Kulkarni. Integrated sensor system for gait analysis. In *Electrical, Electronics, and Optimization Techniques (ICEEOT), International Conference on*, pages 2298–2301. IEEE, 2016.

[36] Joanne S Paton, Elizabeth A Stenhouse, Graham Bruce, Daniel Zahra, and Ray B Jones. A comparison of customised and prefabricated insoles to reduce risk factors for neuropathic diabetic foot ulceration: a participant-blinded randomised controlled trial. *Journal of foot and ankle research*, 5(1):31, 2012.

[37] Massimiliano Pau, Silvia Caggiari, Alessandro Mura, Federica Corona, Bruno Leban, Giancarlo Coghe, Lorena Lorefice, Maria Giovanna Marrosu, and Eleonora Cocco. Clinical assessment of gait in individuals with multiple sclerosis using wearable inertial sensors: comparison with patient-based measure. *Multiple sclerosis and related disorders*, 10:187–191, 2016.

[38] Massimiliano Pau, Giancarlo Coghe, Claudia Atzeni, Federica Corona, Giuseppina Pilloni, Maria Giovanna Marrosu, Eleonora Cocco, and Manuela Galli. Novel characterization of gait impairments in people with multiple sclerosis by means of the gait profile score. *Journal of the neurological sciences*, 345(1-2):159–163, 2014.

[39] Ronald Postuma, Werner Poewe, Irene Litvan, Simon Lewis, Anthony Lang, Glenda Halliday, Christopher Goetz, Piu Chan, Elizabeth Slow, Klaus Seppi, et al. Validation of the mds clinical diagnostic criteria for parkinson's disease (s3. 001), 2018.

[40] Ronald B Postuma, Daniela Berg, Matthew Stern, Werner Poewe, C Warren Olanow, Wolfgang Oertel, José Obeso, Kenneth Marek, Irene Litvan, Anthony E Lang, et al. Mds clinical diagnostic criteria for parkinson's disease. *Movement Disorders*, 30(12):1591–1601, 2015.

[41] Daniele Ravì, Charence Wong, Fani Deligianni, Melissa Berthelot, Javier Andreu-Perez, Benny Lo, and Guang-Zhong Yang. Deep learning for health informatics. *IEEE journal of biomedical and health informatics*, 21(1):4–21, 2017.

[42] Heather S Read, M Elizabeth Hazlewood, Susan J Hillman, Robin J Prescott, and James E Robb. Edinburgh visual gait score for use in cerebral palsy. *Journal of pediatric orthopaedics*, 23(3):296–301, 2003.

[43] Susan A Rethlefsen, Gideon Blumstein, Robert M Kay, Frederick Dorey, and Tishya AL Wren. Prevalence of specific gait abnormalities in children with cerebral palsy revisited: influence of age, prior surgery, and gross motor function classification system level. *Developmental Medicine & Child Neurology*, 59(1):79–88, 2017.

[44] LW Robinson, N Clement, M Fullarton, A Richardson, J Herman, G Henderson, JE Robb, and MS Gaston. The relationship between the edinburgh visual gait score, the gait profile score and gmfcs levels i–iii. *Gait & posture*, 41(2):741–743, 2015.

[45] Johannes CM Schlachetzki, Jens Barth, Franz Marxreiter, Julia Gossler, Zacharias Kohl, Samuel Reinfelder, Heiko Gassner, Kamiar Aminian, Bjoern M Eskofier, Jürgen Winkler, et al. Wearable sensors objectively measure gait parameters in parkinson's disease. *PloS one*, 12(10):e0183989, 2017.

[46] Cosmin Stamate, George D Magoulas, Stefan Küppers, Effrosyni Nomikou, Ioannis Daskalopoulos, Marco U Luchini, Theano Moussouri, and George Roussos. Deep learning parkinson's from smartphone data. In *Pervasive Computing and Communications (PerCom), 2017 IEEE International Conference on*, pages 31–40. IEEE, 2017.

[47] Ioan Stanculescu, Christopher KI Williams, and Yvonne Freer. A hierarchical switching linear dynamical system applied to the detection of sepsis in neonatal condition monitoring. In *UAI*, pages 752–761, 2014.

[48] Christina Strohrmann, Shyamal Patel, Chiara Mancinelli, Lynn C Deming, Jeffrey J Chu, R Greenwald, Gerhard Tröster, and Paolo Bonato. Automated assessment of gait deviations in children with cerebral palsy using a sensorized shoe and active shape models. In *Body Sensor Networks (BSN), 2013 IEEE International Conference on*, pages 1–6. IEEE, 2013.

[49] Antonio Suppa, Ardian Kita, Giorgio Leodori, Alessandro Zampogna, Ettore Nicolini, Paolo Lorenzi, Rosario Rao, and Fernanda Irrera. L-dopa and freezing of gait in parkinson's disease: Objective assessment through a wearable wireless system. *Frontiers in neurology*, 8:406, 2017.

[50] Colin Vroome, Christopher M Maulucci, George M Ghobrial, and James S Harrop. Gait abnormalities. *Differential Diagnosis in Spine Surgery*, page 131, 2016.

[51] Shaohua Wan, Yan Liang, Yin Zhang, and Mohsen Guizani. Deep multi-layer perceptron classifier for behavior analysis to estimate parkinson's disease severity using smartphones. *IEEE Access*, 6:36825–36833, 2018.

[52] Tsung-Hsien Wen, Milica Gasic, Nikola Mrksic, Pei-Hao Su, David Vandyke, and Steve Young. Semantically conditioned lstm-based natural language generation for spoken dialogue systems. *arXiv preprint arXiv:1508.01745*, 2015.

[53] Serena Yeung, Olga Russakovsky, Ning Jin, Mykhaylo Andriluka, Greg Mori, and Li Fei-Fei. Every moment counts: Dense detailed labeling of actions in complex videos. *International Journal of Computer Vision*, 126(2-4):375–389, 2018.

[54] Martha L Zequera and Stephan Solomonidis. Performance of insole in reducing plantar pressure on diabetic patients in the early stages of the disease. In *Engineering in Medicine and Biology Society (EMBC), 2010 Annual International Conference of the IEEE*, pages 2982–2985. IEEE, 2010.

[55] Albert Zeyer, Patrick Doetsch, Paul Voigtlaender, Ralf Schlüter, and Hermann Ney. A comprehensive study of deep bidirectional lstm rnns for acoustic modeling in speech recognition. In *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*, pages 2462–2466. IEEE, 2017.