Multi-Speed Transformer Network for Neurodegenerative disease assessment and activity recognition

Authors

Mohamed Cheriet[1], Vincenzo Dentamaro[2], Mohammed Hamdan[1], Donato Impedovo[2] and Giuseppe Pirlo[2]

Affiliations

1. École de technologie supérieure, ÉTS, 1100 Notre-Dame St W, Montreal, Quebec H3C 1K3, Canada

2. Università degli studi di Bari "Aldo Moro", Department of computer science, via Orabona 4, Bari, 70125, Italy mail:name.surname@uniba.it

Corresponding author: vincenzo.dentamaro@uniba.it

Keyword

Beside gait dataset; multi-speed; deep learning; LSTM; Temporal Convolutional; Attention; Neurodegenerative;

Abstract

Background and Objective:

Neurodegenerative diseases are the most frequent age-related diseases. This type of disease, if not discovered in the initial stage, will compromise the quality of life of the affected subject. Thus, a timely diagnosis is of paramount importance. One of the most used tasks from neurologists to detect and determine the severity of the disease is analysing human gait. This work presents the dataset named "Beside Gait" containing timeseries of coordinates of extracted body joints of people with neurodegenerative diseases in various stages of the disease as well as control subjects. In addition, the novel Multi-Speed transformer technique will be presented and benchmarked against several other techniques making use of deep learning and Shallow Learning. The objective is to recognize subjects affected by some form of neurodegenerative disease in early stage using a computer vision technique making use of deep learning that can be integrated into a smartphone app for offline inference with the aim of promptly initiate investigations and treatment to improve the patient's quality of life.

Methods:

The recorded videos were processed, and the skeleton of the person in the video was extracted using pose estimation. The raw time-series coordinates of the joints extracted by the pose estimation algorithm were tested against novel deep neural network architectures and Shallow Learning techniques. In this work, the proposed Multi-Speed Transformer is benchmarked against other deep neural networks such as Temporal Convolutional Neural Networks, Transformers, as well as Shallow Learning techniques making use of feature extraction and different classifiers such as Random Forests, K Nearest Neighbours, Ada Boost, Linear and RBF SVM. The proposed Multi-Speed Transformer architecture has been developed to learn short and long-term patterns to model the various pathological gaits.

Results:

The Multi-Speed Transformer outperformed all other existing models reaching an accuracy of 96.9%, a sensitivity of 96.9%, a precision of 97.7%, and a specificity of 97.1% in binary classification. The accuracy in multi-class classification for detecting the presence of the disease in various stages is 71.6%, the sensitivity is 67.7%, and the specificity is 71.8%. In addition, tests have also been conducted against two other different activity recognition datasets, namely SHREC and JHMDB, in the exact same conditions. Multi-Speed Transformer has demonstrated to beat always all other tested techniques as well as the techniques reviewed in the state-of-the-art with respectively of accuracy 91.8% and 74%. Having those datasets more than two classes, specificity was not computed.

Conclusions:

The Multi-Speed Transformer is a valuable technique for neurodegenerative disease assessment through computer vision. In addition, the novel dataset “Beside Gait” here presented is an important starting point for future research work on automatic recognition of neurodegenerative diseases using gait analysis.

Introduction

Neurodegenerative disease refers to a set of disorders having the characteristic of developing in a progressive and irreversible manner. Such diseases cause a gradual loss of neuronal cells in certain areas of the central nervous system. Neurons are in fact constituent elements of the nervous system, which includes the brain and spinal cord. The factors that cause neurodegenerative diseases are many and have roots of origin:

- genetic
- hereditary
- environmental

The slow loss of neurons causes disorders of the motor system (ataxia), resulting in dysfunction of balance, movement and symptoms that can be summarized as resting tremor, muscle rigidity, impaired postural reflexes, blocked walking.

In addition, this loss of neurons causes mental disorders (dementia), resulting in memory loss, disorientation, and language deficits.

The number of individuals with neurodegenerative disorders is increasing, due, primarily, to the increase in the aging population, hereditary factors, and in the second instance to a spectrum of parameters of the life of a human being such as lifestyle, stress levels, and food.

A worrying characteristic of this type of disease is the difficult diagnosis: these diseases present asymptomatically, while symptoms become evident only when the disease reaches an advanced stage. In addition, there is no valid cure that allows the patient to recover completely from the disease, but only therapies that can alleviate the symptoms and delay its progression.

One of the most relevant models of analysis in this area is the Gait Analysis, which allows to highlight a strong correlation between neurological and movement-related disorders. In particular, clinical evidence supports the effectiveness of Gait Analysis in quantifying the level of functional limitation due to the onset of neurodegenerative disease and its evolution over time. Therefore, this tool provides a valuable aid to experts in clinical decision-making and the selection of a rehabilitation program.

Gait analysis consists of the punctual study of the walking characteristics of a human being [1].

When we walk, we alternately use our legs to provide support on one side and propulsion on the other side, ensuring that at least one foot always remains in contact with the ground. A person's walking style and manner may vary with mood, age, and health status.

The Gait Cycle represents the set of actions that are repeated in a cyclic manner during walking and is bounded by two consecutive heel strikes of the same foot. The heel strike represents the moment when one foot touches the ground with the heel [2]. The entire cycle is divided into two macro-sections: the stance and swing phases, as shown in Fig.1.

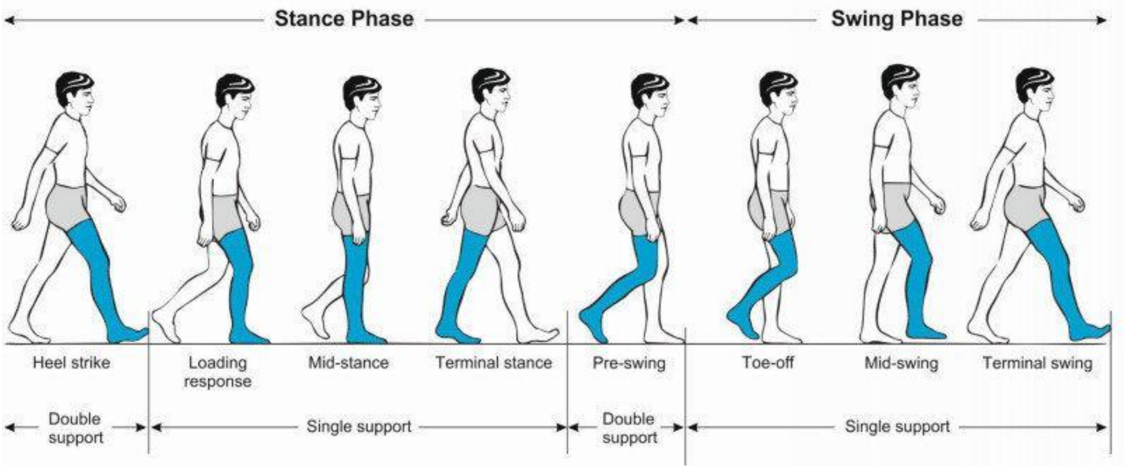


Figure 1: Gait Cycle

In this work, with the help of experienced neurologists and psychologists, a dataset called “Beside Gait” was constructed, containing videos of people walking from left to right and vice versa in front of a video camera. Each person is either healthy or has some form of neurodegenerative disease. For each sick person, a level of severity was also associated.

Our previous work in [3] shows how to post-process the videos, the body joints coordinates were extracted for each subject under analysis and for each frame, in order to construct a temporal vector of coordinates.

After this processing, the classic classification pipeline was applied: spatiotemporal features such as velocity, displacement, and acceleration of each joint were extracted. In addition, the sigma-lognormal features derived from the Kinematic Theory of rapid human movements were successfully tested. Instead, this paper publishes the dataset and focuses on testing the ability of deep neural networks to automatically learn appropriate spatiotemporal patterns to determine whether a patient is sick or is a healthy subject and the severity of the disease, if present.

The objective is to create a deep learning neural network which is accurate for diagnosis neurodegenerative diseases at early stage and with memory and processing constraints allowing to process the data within a mobile phone app and, in a completely offline scenario, provide inference triggering alerts for possible people suffering of some form of neurodegenerative disease. This future app can be used by general practitioners (primary care physicians) as well as family nurse and can, effectively, trigger the first signs of the disease and consequently promptly initiate investigations and treatment to improve the patient's quality of life. This work presents the model to be integrated within the app and reports its accuracies as well as memory and processing requirements. In a future work, the complete mobile application will be presented.

Therefore, the main contributions of this work are the followings:

- The “Beside Gait” dataset. It contains timeseries of body joints coordinates in 2D of patients and healthy control subjects walking from left to right and from right to left.

- Design of Multi-Speed Transformer architecture for neurodegenerative disease classification using raw time-series of joint coordinates. This architecture is composed of two parallel 1D temporal dilated convolution layers followed by a multi-headed attention layer with positional encoding. The first branch performs 1x1 convolution with 1 stride at a time while the second parallel branch performs 1x1 convolution with 3 strides at a time for capturing long and short-term dependencies. The intuition is to perform multi-scale learning in parallel with one branch looking at fine details of the gait, like the presence of tremors, while the other looking at longer time patterns like the presence of freezing events. The network has also a relatively low number of parameters allowing to be integrated into a future mobile app for offline inferencing.

- Comparison against the state-of-the-art Deep Learning techniques for time-series classification, such as temporal convolution and transformers, as well as a comparison with Shallow Learning techniques using feature extraction, feature selection, and classification, is also provided.

- Stress the generalization capabilities of the Multi-Speed Transformer, the Temporal Convolutional Network, and Transformer architectures on two well-known datasets for activity recognition: JHMBD and SHREC.

The work is organized as follows: Section 2 sketches the state of the art in computer vision techniques for neurodegenerative disease assessment as well as techniques for skeleton-based activity recognition. Section 3 presents the dataset and its preprocessing phase. Section 4 contains the Methods describing the Temporal Convolutional architecture, the Transformer architecture, the Multi-Speed Transformer architecture, the features extracted for the Shallow Learning procedure, and the experimental setup. Results is sketched in Section 5 and a discussion is proposed in Section 6. Conclusions and future remarks are in Section 7.

State of the art review

Neurodegenerative disease assessment by means of pattern recognition techniques, answer the question:

What is the behavioral biometric pattern of a neurodegenerative disease across people having it?

In a similar manner, the activity recognition problem seeks to answer the research question: What is the behavioral biometric pattern of people performing the same activity?

As it is possible to assert, neurodegenerative disease assessment can be classified as a particular case of the broader problem named activity recognition since, from a pure pattern recognition perspective, the algorithm is designed to seek fine and specific behavioral biometric patterns of the movement.

In this section state of the art is organized in the following way: first, it is described state of the art concerning computer vision techniques for neurodegenerative disease assessment; in the second instance, it reviews the state-of-the-art techniques for activity recognition using computer vision with a particular focus on skeleton-based activity recognition.

Concerning neurodegenerative disease classification, a specific focus is applied to classification techniques that make use of the inter-patient separation scheme.

The inter-patient separation scheme uses samples from some people for training and samples from other, different people as the test set. This scheme is designed for medical tasks because it correctly answers the previously mentioned research question, and in the second instance, it is used to prevent the spread of biases from training set to the testing set.[3]

2.1 Computer Vision Techniques for Neurodegenerative disease assessment

An important aspect that affects Gait Analysis's performance is the tools used for data acquisition.

In past centuries, gait analysis was performed through the observation performed by experts, capable of extracting quantitative and qualitative data (such as gait speed, cadence and distance) from it. The obtained data was good for finding severe disorders, but it was not sufficient in finding minimal but important variations in the disease progression. This hampered accurate identification and diagnosis, of the actual severity level of a given disease with subsequent planning of the treatment. [2]–[4]

In recent decades, the development of this area of research and new technologies has led to the introduction of tools that are able to derive more accurate and objective data from gait analysis, which have in turn allowed increasingly precise diagnoses to be made and more targeted cures and treatments to be identified for each individual disease. [5]

These new instruments are mostly used within laboratories, set up specifically for gait analysis, under the guidance of experts whose task is to position, set up, calibrate, and monitor these instruments. The great complexity involved in setting up these laboratories, combined with the high cost of the instruments used, has meant that gait analysis still has very limited use today, especially in developing countries, despite its enormous importance.[5]

Of particular interest, because of the high precision and often the low invasiveness, are tools for optical motion capture.

The standard tools used to extrapolate kinematic information are based on optical motion capture: these are professional cameras capable of tracking spatial information and human motion, recording data with high accuracy (e.g., Vicon). [4]

These tools turn out to be very accurate in data mining; however, they also have some disadvantages:

- they can be used only within specially set up laboratories
- they require the careful supervision of experts
- they are usually very expensive
- they do not allow the patient to be analyzed in his or her daily routine, thus preventing the derivation of data reflecting his or her actual condition. [5]

Among these systems is possible to differentiate between systems using markers worn on the human body, and systems that do not make use of markers.

The first approach involves the application of markers, which can be "active," such as bright LEDs, or "passive," and thus be simply reflective. These markers are placed at the parts of the body whose coordinates are to be tracked during motion and allow for high-precision results. [6] The data processing process extrapolates frames from the video that is recorded through the cameras, and subjects each of them to filters that allow the image to be darkened, leaving only the areas marked by the markers in evidence: it will therefore be easier to derive the coordinates of these areas with extreme precision. [7]

The use of markers, however, also has disadvantages, among which mainly emerge the high costs, arising from their purchase and installation, and the limitations they impose on walking and moving freely. [6]

The second approach does not require the use of markers. In this type of approach, the frames extracted from the video recorded through the cameras must undergo numerous transformations to derive the coordinates of body joints. First, each frame is converted to a grayscale image. These conversions are essential to perform the process that is known as "background subtraction”. Doing so produces a binary image, in which the silhouette is clearly separated from the background. Thus, having obtained the silhouette of the moving subject, the coordinates of the body parts of interest for the study are computed, taking advantage of the results of anatomical studies, which have made it possible to estimate the position of some body parts (such as neck, hip, knee, ankle) in relation to the subject's height.

This type of approach allows data to be obtained with sufficient accuracy, but less than the approach using markers. However, compared with the latter, it allows the subject to move freely and feel fully comfortable while performing the test. [6]

An example of a commonly used system that uses the marker less approach for optical motion acquisition is the Microsoft Kinect, which allows tracking of the human skeleton by extrapolating 3D coordinates of the joints. The system appears to be sufficiently accurate, easier to install, and cheaper than conventionally used camera systems. However, the Kinect sensor has two significant shortcomings. The first shortcoming is due to the shallow depth range that the sensor is able to detect (between 0.5m and 4m): this prevents the sensor from capturing a large enough area to perform meaningful gait analysis. On the other hand, the second is intrinsic to the data extrapolation process adopted: the Kinect deduces data for body parts that are occluded and therefore not clearly visible, causing a decrease in the accuracy of the subject's pose estimation.[8]

In particular cases, such as Telemedicine and remote control, a Kinect sensor may not be available. Thanks to the huge spread of smartphones and their modern cameras, it is possible to use the patient's smartphone to extract the body joints coordinates of the person performing the gait.

For doing so, many solutions perform pose estimation from video or frames of common RGB cameras, like the cameras present in modern mobile phones.

In [9] authors used the Convolutional Pose Estimation algorithm to extract joints' movement of each individual and later synthesize different features (e.g. velocity-based and space-temporal features). The Random Forest algorithm was trained on these features to detect the presence and predict the severity the Parkinson's Disease (PD). The F1 score was 0.906 with an interpatient scheme.

Authors in [10] used a famous pose estimation technique called Open Pose [11] to extract joint coordinates and synthesize features like the length of the step, the stride amplitude, the stride velocity, the turning velocity, and many more. These features are easy to interpret by experts. The system achieved an overall accuracy of 94% but without the interpatient separation scheme. Authors in [12] used the camera of a normal smartphone as well as a printed surface to perform background subtraction and silhouette extraction utilizing thresholding and binarization techniques. The system is very fast and accurate for foot track, but no classification was conducted.

The authors of [13] used a multimodal fusion strategy that integrates gait data and eye fixational motion patterns without the use of any kind of marker for Parkinson's Disease prediction. The technique extracted frames from the video and in parallel frames of eye movements. Then the two types of frames are fed into a Convolutional Deep Neural network named MobileNetV2 and the embeddings are extracted constituting the feature vector. The classification achieved 100% accuracy but without the inter-patient separation scheme.

In [14] authors proposed a 3D body pose estimation technique capable of extracting the body joints' coordinate time series in 3 dimensions. These coordinates are then fed into a deep neural network classifier for Parkinson’s recognition of patients with orthopedic problems and post-stroke patients. Because of the use of deep learning, feature engineering was not performed. The average classification accuracy was 71.25%. The authors used the inter-patient separation scheme.

Other approaches successfully used the pose estimation technique for Knees Abduction Moment estimation [15] using temporal body joint coordinates.

2.2 Computer Vision Techniques for Activity Recognition

Concerning activity recognition by computer vision and pose estimation, the authors in [16] used probability maps extracted from the pose estimation algorithm to generate new frames with colored clouds of the probability maps for each frame. The frames of each video are fed into a 3D CNN for classification aims. The system was very good achieving 67.9% accuracy on JHMDB dataset using pose only.

Another solution was proposed by [17] where authors used the extracted joint coordinates, linearize them in a single vector, and stacked them temporarily on top of the other, encoding each coordinate, with a different color. In this way, the authors treated Spatio-Temporal coordinate data as an image and fed these images within a Convolutional Neural Network for further classification. The system achieved an accuracy of 65.5% on JHMDB using RGB information and pose estimation only. Of particular interest is the work proposed in [18] where three semantic deep neural network modules have been proposed to face, separately, different aspects of the time series of body joint coordinates separately. The spatial pose CNN was designed to extract several modalities of pose heatmaps, the temporal pose convolution was adaptively developed to aggregate, the spatial heatmaps over time frames, thus generating a time series of pose representation, and the action CNN which was developed to recognize the human actions. The scores were fused in order to perform action recognition. Authors achieved 65.9% accuracy on JHMBD dataset.

The authors in [19] proposed a deep neural network that taking as input the raw time-series body joint coordinates, were capable of creating location-viewpoint invariant features (embeddings) of the skeleton sequences. The accuracy reached was the state of the art on the JHMDB reaching 77.2% of accuracy.

The work proposed in [20] used Graph neural networks to model relations of hand joints coordinates. The node and edges features are automatically learned via self-in both spatial and temporal domain, the accuracy reached was of 90.7 on SHREC dataset.

A modern technique for video activity recognition is developed in [21]. It uses temporal convolutions by stacking a 3D convolutional layer followed by 2D Long Short-Term Memory (LSTM) layer for modeling space and time dependencies. The accuracy achieved was very high on all three tested datasets. A similar solution was also proposed in [22]–[26]. The temporal convolution with Attention mechanism was proposed in [27] achieving state of art accuracies on SHREC dataset.

Another approach is to use pose estimation algorithms in both 2D and 3D, extract joints coordinate, plot the skeleton on new frames with a black background and feed a custom pre-trained deep neural network such as ResNet-50, DenseNet-201 and so on. This approach is developed in [28]. The accuracy reached was high with respect to other tested techniques.

An interesting approach is used in [29] where, in order to cope with the difficult parallelizability of LSTM and to increase sensitivity in handling temporal data by using 1D CNNs, authors proposed a single 1D CNN with dilated causal convolutions and multi-headed self-attention for activity recognition. The input to the network was 2D body joints coordinates extracted by a pose estimation algorithm using standard RGB cameras.

In general, it is possible to summarize the state of the art in human activity recognition using computer vision with three macro trends. The first trend tends to use video information as a whole, feeding the video into some particular deep neural network capable of capturing both spatial and temporal patterns. The second trend tends to perform pose estimation of the people in the video and then encode this information in various forms ranging from images to vectorial form and embeddings. The third trend tends to use raw 2D or 3D coordinates of the time series of body joints coordinates and feed them into some type of deep neural network which usually make use of 1D CNN or LSTM to manage temporal data, Graph Neural Networks to model human body as a graph and self-attention to automatically learn where to attend in time: i.e., weight more the time-space patterns that discriminate better one activity with respect to other.

Dataset and preprocessing

Beside Gait Dataset description

The dataset used is composed of 115 videos of 43 subjects containing patients with a neurodegenerative disease and controls subjects described as follows:

- 23 control subjects
- 20 subjects with varying neurodegenerative diseases as well as varying severity level.

22 subjects were woman and 21 were man.

The whole dataset is thus composed by:

- 61 videos of control subjects
- 54 videos of people with diseases.

These videos have varying time lengths and were recorded in different structures with different backgrounds and scenes. In each video a person walks, following a linear path in both directions, from left to right and from right to left. Videos were recorded at 25 fps.

It follows the guidelines used for capturing videos and depicted in Figure 2:

- The subjects walked for 4 meters following a straight line highlighted on the floor.
- The camera was positioned perpendicularly about 4 meters distant from the straight line with a height of 2 meters.

Subjects had different aging. Control subjects are between 30-75 years while patients are between 65-90 years. Thus, the dataset is age unbalanced. It is important to state that finding subjects in their 80s completely healthy was challenging.

For the multi-class classification, the stage of the disease is considered. For simplicity and with the help of trained neurologists and psychologists, the disease stage was assessed using the Mini-Mental State Examination. Each subject is classified as normal, mild, or severe stage.

The preprocessed dataset has the following nomenclature:

Data of patients are identified as follow: xxxx_NS_L_DATA, where xxxx is the identification number, N is the initial of the name, S is the initial of the surname, L is the initial of the location, DATA is the data of recording, whereas data of healthy patients are identified as follow: Nxxx_NS_L_DATA, so it adds N as Normal at the beginning of the string.

The dataset contains 2 files: the binary diagnosis excel file contains subjects identifiers and their condition, 0 if healthy, 1 if the subject suffers from some form of neurodegenerative disease. The multi-class diagnosis file contains less data because it was impossible to determine the correct stage for all patients. The multi-class distribution can be found in Fig.3.

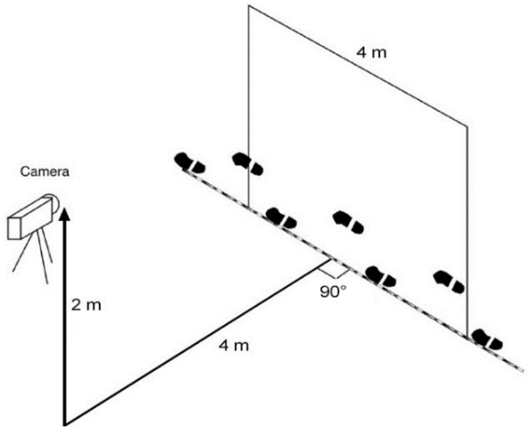


Figure 2: The camera setup for video acquisition

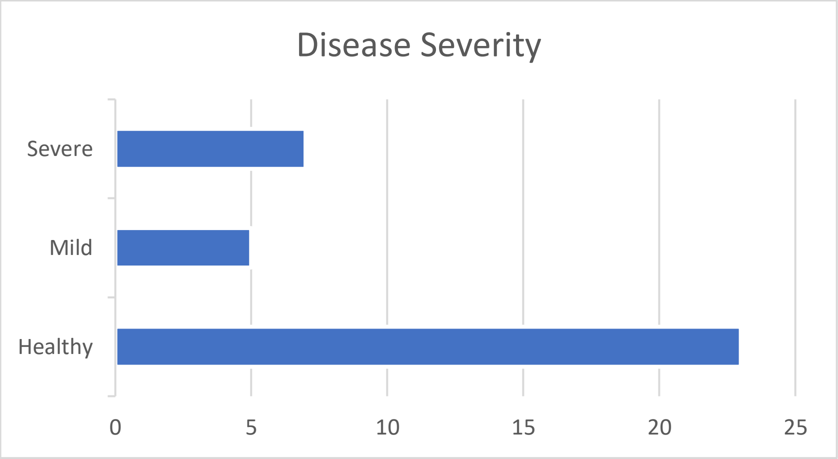
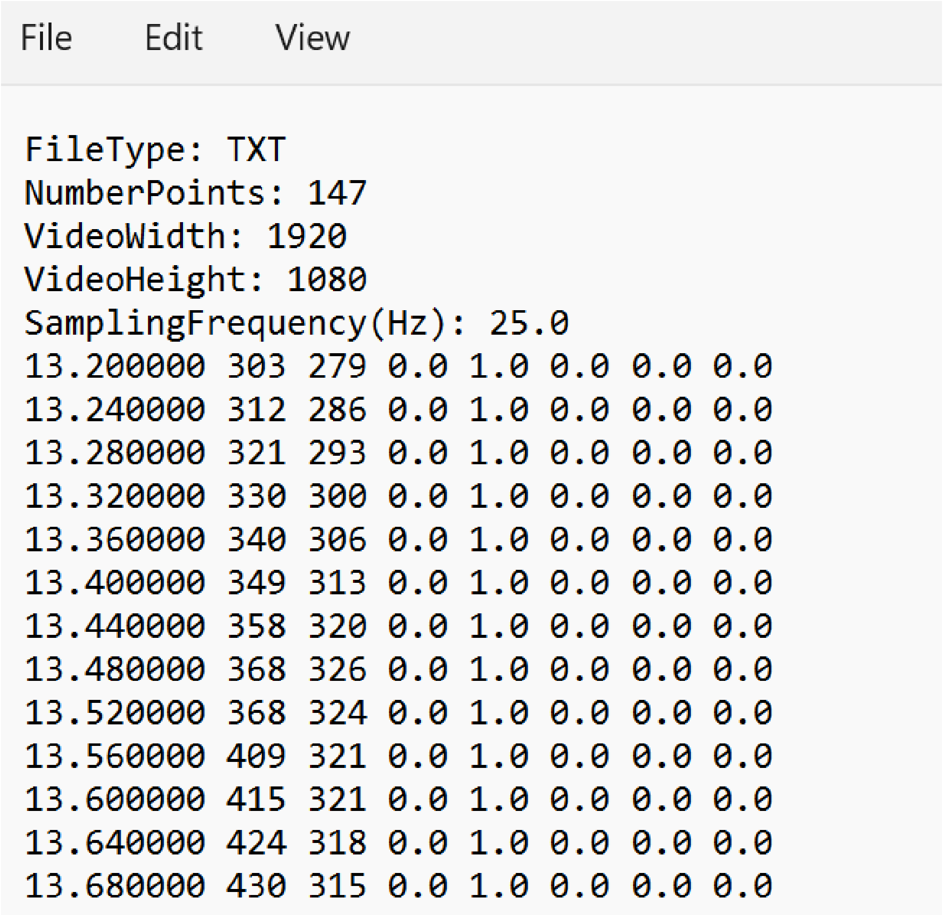


Figure 3: Disease severity distribution 23 healthy subjects, 5 patients in early conditions and 7 patients in severe stage.

Each subject data is organized in folders. Each folder has the name of the identifier specified above and contains two other folders: the left (sx) and right (dx) directions of walking. Each direction contains three other subdirectories namely: (i) original, containing unfiltered time series of body joints coordinates, (ii) no peaks containing time series of body joints coordinates filtered of big mistakes in pose estimation as it will be explained in the preprocessing step and (iii) estimated containing preprocessed data filtered of big mistakes and smoothed with Kalman Filter. Each of these folders contains the body part as a text file. Each text file is composed of coordinates as shown in Fig.4. The first row represents the file type, the second rows is the number of samples, the third and fourth rows represent the width and height of the original video. The fifth row is the sampling rate, in this case 25Hz. From the sixth column to the end there are the timestamp (1st column), x coordinate (2nd column), y coordinate (3rd column), all other columns are unused information maintained for compatibility reasons with previously developed tool for sigma-lognormal extractor, in particular column 5 is the pen down signal, that's because the tool was designed to work with online handwriting data.



```
FileType: TXT
NumberPoints: 147
VideoWidth: 1920
VideoHeight: 1080
SamplingFrequency(Hz): 25.0
13.200000 303 279 0.0 1.0 0.0 0.0 0.0
13.240000 312 286 0.0 1.0 0.0 0.0 0.0
13.280000 321 293 0.0 1.0 0.0 0.0 0.0
13.320000 330 300 0.0 1.0 0.0 0.0 0.0
13.360000 340 306 0.0 1.0 0.0 0.0 0.0
13.400000 349 313 0.0 1.0 0.0 0.0 0.0
13.440000 358 320 0.0 1.0 0.0 0.0 0.0
13.480000 368 326 0.0 1.0 0.0 0.0 0.0
13.520000 368 324 0.0 1.0 0.0 0.0 0.0
13.560000 409 321 0.0 1.0 0.0 0.0 0.0
13.600000 415 321 0.0 1.0 0.0 0.0 0.0
13.640000 424 318 0.0 1.0 0.0 0.0 0.0
13.680000 430 315 0.0 1.0 0.0 0.0 0.0
```

Figure 4: Body part text file.

SHREC Dataset

The SHREC Dataset is composed by 2800 samples divided in 1 official training split and 1official testing split, 3D data of hands coordinates. The gesture to be recognized are 14 or 28 depending on the test. The dataset is available online with the original publication in [30].

JHMDB Dataset

The JHMDB Dataset is composed of 928 samples divided in 3 official training splits and 1 official testing split, 2D data of body skeleton coordinates. The gesture to be recognized are 21. The dataset and original publication in [31] are present online.

Preprocessing pipeline

The preprocessing pipeline is proposed in Figure 5. This pipeline is described in detail in our seminar work in [3] and for clarity is also summarized here.

Coordinate Extraction

For the 2D body joints coordinate extraction it has been used Open Pose [31], an open-source system for real-time Pose Estimation making use of the Part Affinity Fields (PAFs) technique. The system used only 14 extracted body joints, namely: *Nose, Neck, Wrists (right and left), Elbows (right and left), Shoulders (right and left), Hips (right and left), Ankles (right and left), Knees (right and left).* A result is shown in Fig.6

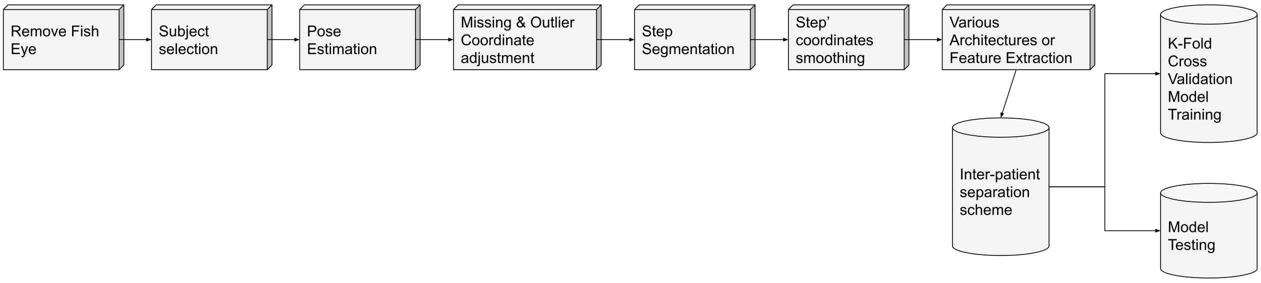

Remove Fish Eye
Subject selection
Pose Estimation
Missing & Outlier Coordinate adjustment
Step Segmentation
Step' coordinates smoothing
Various Architectures or Feature Extraction
Inter-patient separation scheme
K-Fold Cross Validation Model Training
Model Testing

Figure 5. Pipeline common to all proposed solutions

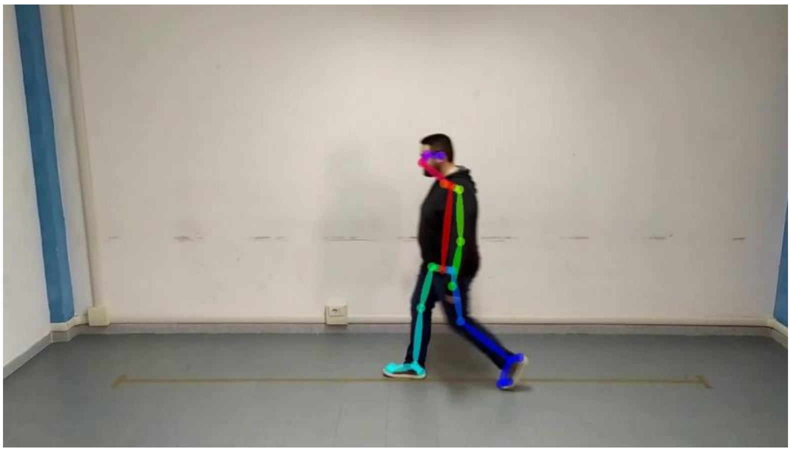


Figure 6. The extracted skeleton of a person performing gait

Sequence Creation

In case of the presence of more people in the videos, it is important to isolate the coordinates relative to the patient from those close to the other individuals present and track them in time. Thus, the developed system calculated the Middle Hip joint for each person in the video and then calculated the Euclidean distance between the current reference coordinate and the coordinate relative to the joint Middle Hip of each other people. In this case, before every computation, the user is requested to click on the hip of the person to analyze looking at the first frame. The closest reference coordinate in all other subsequent frames allows to track the person in the video.

Interpolation and Kalman Smoothing

It has been demonstrated in [2], [3], [11], [32] that pose estimation have high oscillations. In order to smooth oscillations, a linear Kalman filter [33] has been used on raw coordinates [3] performing 1D time-series smoothing and analyzing each coordinate (e.g. x or y) as a time-series. Formally the Kalman filter states are presented in eq. (1):

$$\begin{gathered} x_k = x_{k-1} + w_k \\ z_k = x_k + v_k \end{gathered} \quad (1)$$

Where $x_k \in N*1$ contains a series of coordinates (x or y) of a body joint. The random variables w_k and v_k represent respectively the process noise and the measurement noise. They are independent and normally distributed. z_k is a new measurement.

The time update (prediction) equation is shown in eq. (2):

$$\begin{gathered} \hat{x}_k^- = \hat{x}_{k-1} \\ P_k^- = P_{k-1} \end{gathered} \qquad (2)$$

While the time measurement update (correction) is shown in eq. (3):

$$\begin{gathered} K_k = \frac{P_k^-}{P_k^- + R} \\ \hat{x}_k = \hat{x}_k^- + K_k(z_k - \hat{x}_k^-) \\ P_k = (1 - K_k)P_k^- \end{gathered} \quad (3)$$

Where R is a constant representing the noise which was set to 0.1.

Methods

In this chapter, Shallow Learning techniques, i.e. the classic Machine Learning pipeline making use of feature extraction and non-deep neural networks classifiers, are benchmarked against deep learning techniques for neurodegenerative disease assessment through gait analysis. In particular, a novel deep neural network architecture called Multi-Speed Transformer is proposed and benchmarked against existing Temporal Convolutional architecture and vanilla Transformer architecture and several Shallow Learning techniques.

The Temporal Convolutional Neural Network

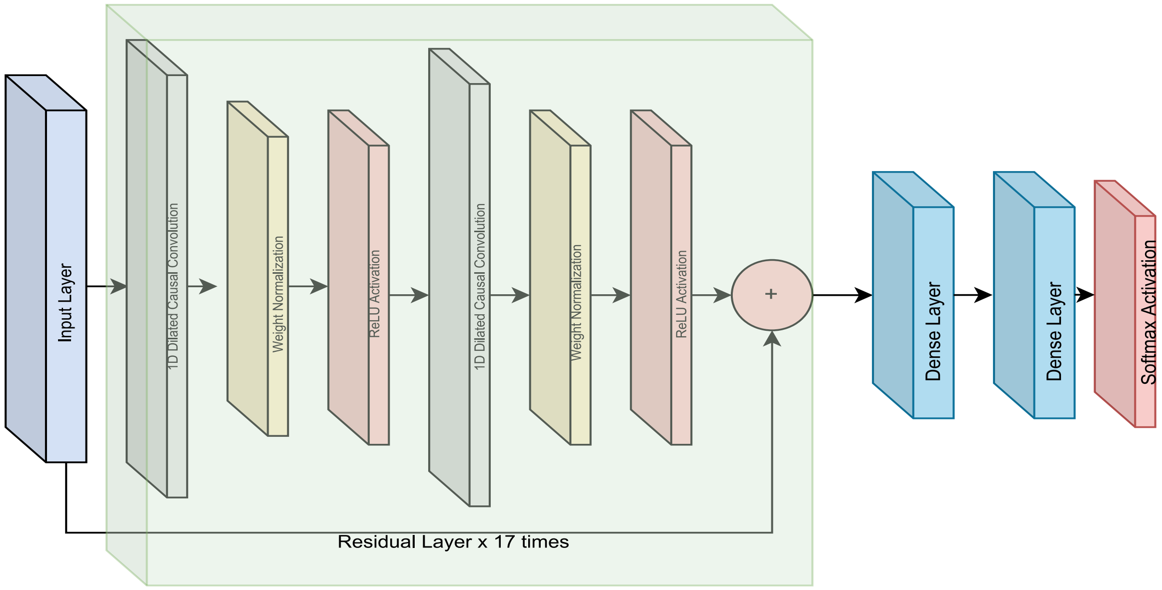


Figure 7. The Temporal Convolution architecture. The content of the green transparent box is repeated 17 times.

In sequence modeling tasks, a variation on Convolutional Neural Networks called a temporal convolutional network (TCN) was used in the experiments. [34] TCN has a memory significantly longer than recurrent architectures with the same capacity. Likewise, in different sequence modeling tasks, the Neurodegenerative disease assessment task consistently achieves better results than RNN, LSTM, and GRU architectures; since TCN is parallel, adaptable in terms of receptive field size, stability of gradients, low memory requirements during training, and acceptability of inputs of varying lengths. The architecture of this TCN is shown in Fig. 7.

Vanilla Transformer architecture

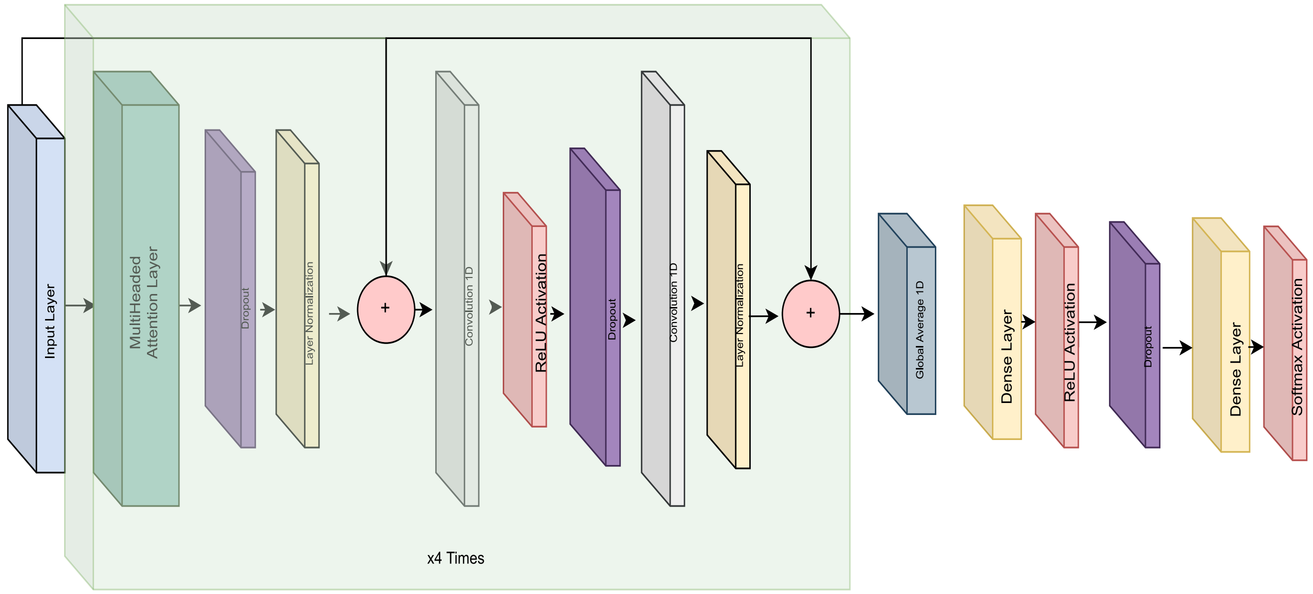


Figure 8 The Transformer architecture. The content of the green transparent box is repeated 4 times.

The original Transformer [35] architecture for machine translation performed better and faster than RNN encoder-decoder models. The Transformer is encoder-decoder-based. The encoder extracts features from input, and the decoder uses them to construct output. Encoder block employs self-attention to augment each token with correction context. Likewise, with the transformer model, it is possible to inject positional encoding into each embedding so that the model can know the position of the input without repetition. Transformer architecture processes input data without recurrence or convolution, which can handle any data sequence. In computer vision, the sequences are image patch sequences, whereas, in reinforcement learning, the sequences are sequences of states, actions, and rewards. In this specific case, being a sequence classification and not generation, only the encoder model has been used. In particular the encoder has been repeated 4 times as shown in Fig. 8.

Multi-Speed Transformer architecture

The Multi-Speed Transformer architecture is presented in Fig. 9. The intuition behind this architecture is the capability of learning meaningful time-dependent correlations and patterns at two scale levels, fast and slow. It uses the concept of multiscale learning [36]–[38], in which data is analyzed at different scales. An intuitive analogy can be made with a slide viewed by a microscope at various resolutions: at high resolution, it is possible to observe very small features, while at lower resolution it is possible to capture more broad concepts. In a similar fashion, Multi-Speed Transformer is composed of two parallel branches. The top branch performs 1D convolution with stride = 1 and a subsequently dilated convolution with dilation rate = 2. While in the second parallel branch the first convolution has a stride = 3. Apart from this change, the two parallel branches are equal: the top branch continues with a Positional Encoding Layer [35] which adds the output of the dilated 1D convolution with a positional signal, in this case, the same sine and cosine functions used in [35].

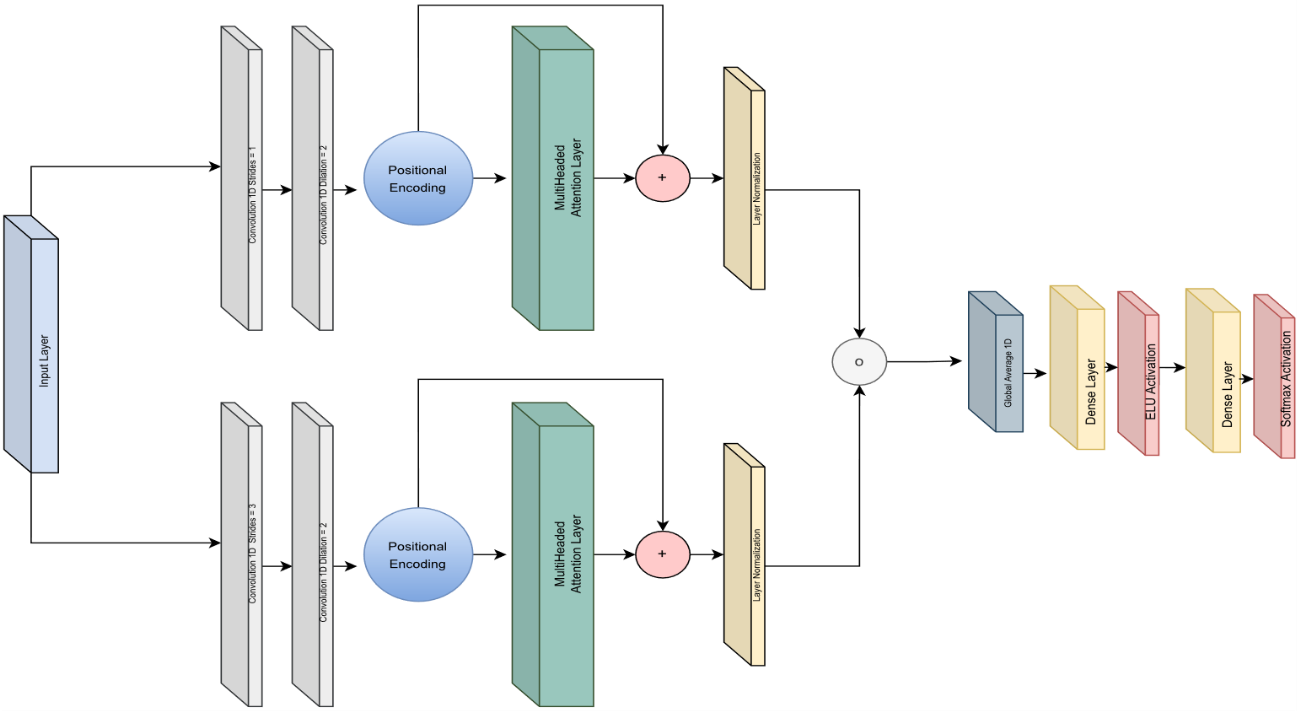

Input Layer
Convolution 1D Strides = 1
Convolution 1D Dilation = 2
Positional Encoding
MultiHeaded Attention Layer
+
Layer Normalization
Convolution 1D Strides = 3
Convolution 1D Dilation = 2
Positional Encoding
MultiHeaded Attention Layer
+
Layer Normalization
o
Global Average 1D
Dense Layer
ELU Activation
Dense Layer
Softmax Activation

Figure 9 The Multi-Speed Transformer Architecture

Positional Encoding is used to remodel the temporal dependency outputted by the dilated convolutional layer, which otherwise would be lost once injected into the Multi-Headed Attention Layer. The Multi-Headed Attention layer is defined with a key size of 128, 4 heads, and 0.2 of dropout. It is trained in a self-attention fashion, allowing to capture the correlations between the different elements of the same sequence. The output is summed with the output of the positional encoding allowing residuals to propagate forwards. Subsequently, a Z-Score normalization layer ends the two in parallel branches. Their output is concatenated and fed into a mono-dimensional global average layer, a dense layer with ELU [39] activation function, and a last dense layer with SoftMax. Thus, the merged representation of patterns extracted at various scales performs the decision.

The theoretical explanation behind the in-parallel varying speed is represented by two key components: the varying stride in parallel branches and the use of dilated convolutions [40] for expanding the receptor fields. Formally a 1D convolution with varying stride is shown in eq. (4):

$$y(n) = \begin{cases} \sum_{i=0}^{k} I(n+i)h(i), & \text{if } n = 0 \\ \sum_{i=0}^{k} x(n+i+(s-1))h(i), & \text{otherwise.} \end{cases} \quad (4)$$

Where *x* is the input to the convolutional layer having size *n*, the kernel *h* has length *k,* and being *s* the number of positions (number of strides) after each convolution operation. A stride *s* > 1 loose information. This can be interpreted as losing fine-grain details with the aim of capturing the big picture. It can be intuitively sough as a moving average filter with a non-overlapping window.

While the dilated convolution [40] can be formally defined as in eq. (5):

$$(\mathbf{x} *_l h)(y) = \sum_{i=0}^{f-1} \mathbf{x}_{y-l\cdot i} \cdot h(i) \quad (5)$$

where $*_l$ is the dilated convolution operator, *h* is the kernel function, *f* is the width of the convolutional filter, *l* is the dilation rate (in this case 2) and *y* is an element of the 1D sequence. Thus, the idea of dilated convolution is to skip some values of the input in order to cover a larger area: it can be viewed as a way to expand the field of view without increasing the computational cost and not of less importance, allowing to get rid of the pooling layer, which means that there is no loss of resolution in the output series. To recap, the top branch is engineered to capture fine details of the movement leveraging computational cost, while the lower branch is engineered to lose information with the benefit of having a global view of the movement.

Shallow Learning

Feature extraction is performed on 2D time-series coordinates of the body joints. Once the coordinate sequences have been obtained, features presented in Table 1 have been synthesized. These features range from Kinematic angles, Spatio temporal and sigma lognormal features. These last features derive from the Kinematic Theory of Rapid Human Movements. [41] . This theory is based on the intuition that human movements (movements of wrist , elbow, legs and etc…) is a mixture of primitives, whose acceleration and velocity profile is lognormal. A more detailed description of the use of sigma-lognormal features for neurodegenerative disease classification is provided in [3], [32], [42]–[44]. For each feature, being computed on a time-series of coordinates, the mean, the median, the standard deviation, 1 and 99 percentile statistics have been used to synthesize temporal features.

Table 1. Extracted Features

Category	Features
Temporal Space	Displacement x & y Velocity x & y Acceleration x & y Tangent angle
Sigma Log Normal features	Number of lognormal strokes $D, \mu, \sigma, \theta_s, \theta_e$ parameters for all lognormal strokes
Corners	Angle between: 1) Nose, Neck and Hip 2) Neck Hip and Knee 3) Shoulder, Elbow and Wrist 4) Hip, Knee and Ankle 5) Right Knee, Hip and Left Knee

Experiment

Tests have been conducted by using the previously explained interpatient-separation scheme. Both binary and multi-class classifications are repeated 10 times with a random internal 70-30 ratio ensuring the previously explained inter-patient scheme. Thus, at every fold, 28 subjects went in the training and 12 in the test. This cross-validation test is used to provide more truthful results given the small size of the dataset.

For Shallow Learning techniques, the feature selection was carried out using ExtraTrees making use of top 100 ordered most important features [45].

Features were standardized with Z-Score. For the Shallow Learning the following 5 classifier have been used:

1. KNN
2. Random-Forest (RF)
3. AdaBoost using DecisionTree
4. SVM linear
5. SVM RBF

The RF was configured with 50 trees as well as a maximum depth of 5 for as pre-pruning parameter. The K of the KNN was 5. Linear SVM was configured with C = 1. The SVM RBF was used with automatic gamma adjusting and C = 1. Instead, the AdaBoost was configured with 10 decision trees as week learner each with max depth pre-pruning parameter of 10.

For Deep Learning, time-series of raw x, y coordinates of body joints were used. The x coordinate was scaled with respect to the width of the video image, and y with respect to the height of the image, thus respectively by 1920 and 1080.

For what concerning SHREC and JHMDB datasets, only 2D (x,y) coordinates were used for the classification. Specifically, for the SHREC it was selected to classify 14 different actions, while for JHMDB 21 different actions. Both dataset were already normalized.

Results

Results for the various datasets are presented in Table 2 for the binary classification on the neurodegenerative disease dataset and Table 3 for multi-class classification. JHMDB dataset results are presented in Table 4 while SHREC dataset results are in Table 5. In bolded the techniques achieving the highest accuracies.

Neurodegenerative Gait Dataset Results

Table 2. Results on gait dataset with binary classification

Algorithm	F1	Sensitivity	Specificity	Precision	Acc	AUC
KNN	0.949	0.930	0.932	0.968	0.949	0.951
Random Forest	0.942	0.945	0.944	0.943	0.942	0.945
Ada Boost	0.932	0.936	0.938	0.943	0.942	0.943
Linear SVM	0.955	0.953	0.955	0.959	0.955	0.961
RBF SVM	0.954	0.959	0.961	0.952	0.954	0.955
Temporal Convolutional Network	0.917	0.919	0.921	0.933	0.919	0.919
Transformer Architecture	0.960	0.960	0.961	0.965	0.960	0.960
Multi-Speed Transformer	**0.967**	**0.969**	**0.971**	**0.977**	**0.969**	**0.969**

Table 3. Results on gait dataset with multi-class classification

Algorithm	F1	Sensitivity	Specificity	Precision	Acc	AUC
KNN	0.550	0.587	0.588	0.539	0.587	0.682
Random Forest	0.511	0.561	0.563	0.491	0.561	0.668
Ada Boost	0.534	0.574	0.572	0.529	0.574	0.656
Linear SVM	0.535	0.592	0.594	0.520	0.592	0.700
RBF SVM	0.472	0.508	0.511	0.461	0.508	0.622
Temporal Convolutional Network	0.640	0.715	0.718	0.614	0.715	0.738
Transformer Architecture	0.598	0.689	0.691	0.559	0.689	0.613
Multi-Speed Transformer	**0.671**	**0.716**	**0.718**	**0.677**	**0.716**	**0.739**

JHMDB Dataset

Table 4. Results on JHMDB Dataset Multi-Class Classification

Algorithm	F1	Sensitivity	Precision	Acc	AUC
Choutas et al. [16]	NA	NA	NA	0.679	NA
Ludl et al. [17]	NA	NA	NA	0.655	NA
Yan et al. [18]	NA	NA	NA	0.659	NA
Yang et al. [19] **with 64 filters**	**NA**	**NA**	**NA**	**0.772**	**NA**
Yang et al. [19] with 32 filters	NA	NA	NA	0.737	NA
Temporal Convolutional Network	0.580	0.570	0.590	0.570	0.749
Transformer Architecture	0.05	0.130	0.04	0.14	0.5
Multi-Speed Transformer	0.740	0.740	0.760	0.740	0.849

SHREC Dataset

Table 5. Results on SHREC Dataset Multi-Class Classification

Algorithm	F1	Sensitivity	Precision	Acc	AUC
Hou et al. [27]	NA	NA	NA	0.892	NA
Devineau et al. [25]	NA	NA	NA	0.912	NA
Lai et al. [26] average	NA	NA	NA	0.854	NA
Lai et al. [26] maximum reached	NA	NA	NA	0.853	NA
Temporal Convolutional Network	0.870	0.870	0.870	0.870	0.922
Transformer Architecture	0.651	0.680	0.670	0.680	0.767
Multi-Speed Transformer	**0.918**	**0.918**	**0.918**	**0.918**	**0.944**

Model Complexity Analysis

Table 6. Models' complexity comparison with respect to the number of parameters

Algorithm	# Parameters
Yang et al. [19] with 64 filters	1.8M
Yang et al. [19] with 32 filters	0.5M
Temporal Convolutional Network	0.44M
Transformer Architecture	0.261M
Multi-Speed Transformer	0.303M

Table 6 reports the known number of parameters of the compared architectures, these are the trainable elements, the weights of the network that need to be changed to perform the learning. Generally speaking, it is difficult to define a unique complexity metric to compare the intrinsic complexity of the proposed solutions as well as state-of-the-art existing solutions. This is due mainly to the various internal transformation. Thus, it is usually used a simple but effective complexity metric like the Akaike Information Criterion briefly AIC [46] defined in eq. (6):

$$\boldsymbol{AIC = -2\log L + 2 \cdot d} \quad (6)$$

d is the number of parameters of the model and L is the maximized value of the likelihood function of the model. Regarding deep neural networks and their huge number of parameters, wanting to reason in the limit, it is clear that the equation is governed almost solely by d thus a simple but effective complexity comparison can be made in terms of the number of trainable parameters. Unfortunately, very few works cite the number of parameters or any form of complexity.

Discussion

The capabilities of the Multi-Speed Transformer of learning multi-scale patterns at different speed has allowed the increase in the accuracy of classification of neurodegenerative diseases in both binary and multi-class classification.

It is important to state that, in this case, the *precision* identifies the percentage of subjects classified as sick that are really sick (ground truth). The *sensitivity* identifies the percentage of correctly classified subjects with respect to all people that have the disease. Whereas the *specificity* identifies the percentage of correctly classified healthy subjects, among the healthy population.

The Multi-Speed Transformer reached the state-of-the-art accuracy in the binary classification of 0.969 with a precision of 0.977, a sensitivity of 0.969 and a specificity of 0.971 as shown in Table 2. Such high ratio of precision suggests that the system is very good at finding sick people and it is good in discriminating people that are clearly healthy. Sick people that fall in the mild severity level could be more difficult to identify, for this reason multi-stage classification has been performed and results are presented in Table 3. Also in this case, the Multi-Speed Transformer outperforms all other tested techniques by a large margin. In particular, it reached an accuracy in multi-class classification of 0.716 with a precision of 0.677, a specificity of 0.718 and a sensitivity of 0.716. Even though it is the most accurate model among the tested, its precision may seem relatively low but it has enough predictive power in discerning people with mild condition from healthy. This plays a major role when it comes to neurodegenerative disease assessment. This is because a timing detection of the neurodegenerative disease, for example when it is in its mild condition, increases the quality of life.[47] Therefore, to some extent, the Multi-Speed Transformer can be used as a biomarker for detecting the early signs of the disease. The large margin of accuracy difference of over 25% from binary to multi-class classification means that it is very difficult to discern sick people in mild conditions from healthy subjects, especially if healthy subjects are elderly. Of course, the dataset size limits an accurate estimate of the metrics.

The difference with Shallow Learning is also visible in the multi-class scenario reaching a +12% of increased accuracy of Multi-Speed Transformer with respect to the linear SVM. Interestingly the Temporal Convolution did not perform better than Shallow Learning on the binary classification scenario but outperformed the Shallow Learning techniques in the multi-class scenario. The transformer architecture was successful in the binary classification case, but unsatisfying in the multi-class scenario, reinforcing the intuition that multi-speed learning is required to correctly model long and short-term patterns such as example freezing gaits (long-term patterns) and tremors (short term patterns).

It is important to state that, at moment of writing, there are very few datasets for neurodegenerative disease assessment for multi-stage (varying severity) classification that make use uniquely of videos of people performing gait as input. Other works [48]–[50] tested on different neurodegenerative disease gait datasets made of videos with multi-stage (multi-class) classification setting achieved a multi-class classification accuracy of about 0.6.

In the cases of JHMDB and SHREC datasets, it is possible to see, respectively from Table 4 and Table 5, that using *only* the 2D time-series coordinates, the system was on par or higher than the state of the art. In particular, on the SHREC dataset in Table 5, Multi-Speed Transformer outperformed the state of the art. While on the JHMDB the Multi-Speed Transformer is outperformed by Yang et al. [19]. Authors of [19] specified that the network reaching the highest accuracy have a complexity of 1.82 million parameters using 64 filters while using 32 filters reduced the number of parameters to 0.5M as shown in Table 6 but also decreased the accuracy to 0.737. Multi-Speed Transformer reached the accuracy of 0.740 with just 0.303M parameters, thus at a lower complexity level, the multi-speed Transformer performs better. The number of parameters of Multi-Speed Transformer is just a fraction of the number of parameters of the MobileNetV3 network [51] used for inference in various mobile apps which has been proved to perform inference in milliseconds. This is of paramount importance when it comes to offline inference for mobile apps. With the aim of testing the Multi-Speed Transformer on a modern smartphone with memory and processing constraints, and being it built using TensorFlow v2.10, the trained model has been successfully converted into a TensorFlow Lite model using the official TensorFlow model converter into a model ready to be used for inferencing by a smartphone. The memory consumption of the model is 5818 Kbytes and the overall memory usage (including TensorFlow lite dependencies) in android is about 26 Mbytes well below the amount of ram available on modern Android smartphones. For inferencing time result, we tested it by using a test app, capable of testing only the model on a previously exported pose estimation of a person walking for 5 seconds from left to right. Tests were conducted on a OnePlus Nord 2 with 6gb of Ram and CPU MTK Dimensity 1200-AI, the inference only took in average 186ms with a standard deviation of 32ms without optimization. So, to conclude, the model can perform inference on a modern smartphone without any problems.

Limitations

Despite the low footprint in memory and time consumption for inferencing of the Multi-Speed Transformer, the pre-processing steps require pose estimation the be computed within a smartphone. Currently, modern solutions of pose estimation, such as Google ML Kit [52] report about 30-45 fps, but in our tests, the most accurate model performs pose estimation at about 10-15 fps. At this time it should be added the computation of the Kalman filter. Of course, this result shows that processing on a mobile smartphone is feasible, but in order to be compatible with the majority of smartphones, the pose estimation, as well as the Kalman Filter estimation and the inference with the Multi-Speed Transformer could not be in real-time. In theory, it will take just a few seconds of computation depending on the duration of the gait analysis. Other limitations of the study are the orientation of the camera and its distance from the subject. To mitigate this bias, it is required to strictly following the indication in Fig. 2.

Conclusions

In this work, the Multi-Speed transformer architecture is proposed. The intuition behind the development of this architecture is to develop a system capable of learning short and long-term patterns to model the various gaits associated with various diseases and the varying severity of the disease. The technique has been benchmarked against the here proposed “Beside Gait” neurodegenerative gait dataset as well as on two activity recognition datasets named SHREC and JHMDB. Accuracies are compared with different techniques belonging to the Shallow Learning or deep learning scenario. The Multi-Speed Transformer demonstrated state-of-the-art accuracy on the neurodegenerative disease dataset, as well as high accuracies on SHREC and JHMBD datasets. Multi-Speed Transformer has a small complexity with about 0.3M parameters making it suitable to be integrated, for inferencing purposes, within smartphones and small edge AI solutions with limited memory and computing power, thus not requiring an internet connection and allowing a real-time response.

In the future, the Multi-Speed Transformer will be applied to different other problems such as, but not limited to violence recognition and in general, time-series classification and prediction.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Ethical approval

All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Informed consent

Informed consent was obtained from all individual participants included in the study.

G.D.P.R. Compliancy

The dataset is provided totally anonymously and does not contain videos containing people's faces or sensitive biometrics but only spatiotemporal coordinates of body parts extracted by a pose estimation technique.

References

[1] M. W. Whittle, “Clinical gait analysis: A review,” *Hum. Mov. Sci.*, vol. 15, no. 3, pp. 369–387, Jun. 1996.

[2] G. Cicirelli, D. Impedovo, V. Dentamaro, R. Marani, G. Pirlo, and T. D’Orazio, “Human Gait Analysis in Neurodegenerative Diseases: a Review,” *IEEE J Biomed Health Inform*, 2021, doi: 10.1109/JBHI.2021.3092875.

[3] V. Dentamaro, D. Impedovo, and G. Pirlo, “Gait Analysis for Early Neurodegenerative Diseases Classification Through the Kinematic Theory of Rapid Human Movements,” *IEEE Access*, vol. 8, pp. 193966–193980, 2020.

[4] S. Chen, J. Lach, B. Lo, and G. Z. Yang, “Toward Pervasive Gait Analysis With Wearable Sensors: A Systematic Review,” *IEEE J Biomed Health Inform*, vol. 20, no. 6, pp. 1521–1537, Nov. 2016, doi: 10.1109/JBHI.2016.2608720.

[5] S. Kumar, K. Gopinath, L. Rocchi, P. T. Sukumar, S. Kulkarni, and J. Sampath, “Towards a portable human gait analysis & monitoring system,” *2018 International Conference on Signals and Systems, ICSigSys 2018 - Proceedings*, pp. 174–180, Jun. 2018, doi: 10.1109/ICSIGSYS.2018.8372660.

[6] C. Prakash, R. Kumar, N. Mittal, and G. Raj, “Vision based Identification of Joint Coordinates for Marker-less Gait Analysis,” *Procedia Comput Sci*, vol. 132, pp. 68–75, 2018, doi: 10.1016/J.PROCS.2018.05.060.

[7] G. Gao, M. Kyrarini, M. Razavi, X. Wang, and A. Graser, “Comparison of Dynamic Vision Sensor-Based and IMU-based systems for ankle joint angle gait analysis,” *2016 2nd International Conference on Frontiers of Signal Processing, ICFSP 2016*, pp. 93–98, Dec. 2016, doi: 10.1109/ICFSP.2016.7802963.

[8] M. Pathegama *et al.*, “Moving Kinect-Based Gait Analysis with Increased Range,” *Proceedings - 2018 IEEE International Conference on Systems, Man, and Cybernetics, SMC 2018*, pp. 4126–4131, Jan. 2019, doi: 10.1109/SMC.2018.00699.

[9] M. H. Li, T. A. Mestre, S. H. Fox, and B. Taati, “Vision-based assessment of parkinsonism and levodopa-induced dyskinesia with pose estimation,” *J Neuroeng Rehabil*, vol. 15, no. 1, pp. 1–13, Nov. 2018, doi: 10.1186/S12984-018-0446-Z/TABLES/8.

[10] A. Zanela *et al.*, “Using a Video Device and a Deep Learning-Based Pose Estimator to Assess Gait Impairment in Neurodegenerative Related Disorders: A Pilot Study,” *Applied Sciences 2022, Vol. 12, Page 4642*, vol. 12, no. 9, p. 4642, May 2022, doi: 10.3390/APP12094642.

[11] Z. Cao, T. Simon, S. E. Wei, and Y. Sheikh, “Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields,” *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, vol. 2017-January, pp. 1302–1310, Nov. 2016, doi: 10.48550/arxiv.1611.08050.

[12] W. Zhu, B. Anderson, S. Zhu, and Y. Wang, “A computer vision-based system for stride length estimation using a mobile phone camera,” *ASSETS 2016 - Proceedings of the 18th International ACM SIGACCESS Conference on Computers and Accessibility*, pp. 121–130, Oct. 2016, doi: 10.1145/2982142.2982156.

[13] J. Archila, A. Manzanera, and F. Martínez, “A multimodal Parkinson quantification by fusing eye and gait motion patterns, using covariance descriptors, from non-invasive computer vision,” *Comput Methods Programs Biomed*, vol. 215, p. 106607, Mar. 2022, doi: 10.1016/J.CMPB.2021.106607.

[14] R. Mehrizi, X. Peng, S. Zhang, R. Liao, and K. Li, “Automatic Health Problem Detection from Gait Videos Using Deep Neural Networks,” Jun. 2019, doi: 10.48550/arxiv.1906.01480.

[15] M. A. Boswell *et al.*, “A neural network to predict the knee adduction moment in patients with osteoarthritis using anatomical landmarks obtainable from 2D video analysis,” *Osteoarthritis Cartilage*, vol. 29, no. 3, pp. 346–356, Mar. 2021, doi: 10.1016/J.JOCA.2020.12.017.

[16] V. Choutas, P. Weinzaepfel, J. Revaud, and C. Schmid, “PoTion: Pose MoTion Representation for Action Recognition,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2018. doi: 10.1109/CVPR.2018.00734.

[17] D. Ludl, T. Gulde, and C. Curio, “Simple yet efficient real-time pose-based action recognition,” in *2019 IEEE Intelligent Transportation Systems Conference, ITSC 2019*, 2019. doi: 10.1109/ITSC.2019.8917128.

[18] A. Yan, Y. Wang, and Z. Li, “PA3D : Pose-Action 3D Machine for Video Recognition,” *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[19] F. Yang, Y. Wu, S. Sakti, and S. Nakamura, “Make skeleton-based action recognition model smaller, faster and better,” *1st ACM International Conference on Multimedia in Asia, MMAsia 2019*, Dec. 2019, doi: 10.1145/3338533.3366569.

[20] Y. Chen, L. Zhao, X. Peng, J. Yuan, and D. N. Metaxas, “Construct Dynamic Graphs for Hand Gesture Recognition via Spatial-Temporal Attention,” *30th British Machine Vision Conference 2019, BMVC 2019*, Jul. 2019, doi: 10.48550/arxiv.1907.08871.

[21] Y. A. Andrade-Ambriz, S. Ledesma, M. A. Ibarra-Manzano, M. I. Oros-Flores, and D. L. Almanza-Ojeda, “Human activity recognition using temporal convolutional neural network architecture,” *Expert Syst Appl*, vol. 191, p. 116287, Apr. 2022, doi: 10.1016/J.ESWA.2021.116287.

[22] S. Bai, J. Z. Kolter, and V. Koltun, “An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling,” Mar. 2018, doi: 10.48550/arxiv.1803.01271.

[23] D. Srivastav, A. Bajpai, and A. Singhal, “A Temporal Convolutional Neural Network Based Activity Recognition Model using a Real-Time Two-Dimensional Single Pose Estimation Framework,” 2022. doi: 10.1109/confluence52989.2022.9734159.

[24] S. K. Yadav, K. Tiwari, H. M. Pandey, and S. A. Akbar, “Skeleton-based human activity recognition using ConvLSTM and guided feature learning,” *Soft comput*, vol. 26, no. 2, 2022, doi: 10.1007/s00500-021-06238-7.

[25] G. Devineau, F. Moutarde, W. Xi, and J. Yang, “Deep learning for hand gesture recognition on skeletal data,” *Proceedings - 13th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2018*, pp. 106–113, Jun. 2018, doi: 10.1109/FG.2018.00025.

[26] K. Lai and S. N. Yanushkevich, “CNN+RNN Depth and Skeleton based Dynamic Hand Gesture Recognition,” *Proceedings - International Conference on Pattern Recognition*, vol. 2018-August, pp. 3451–3456, Nov. 2018, doi: 10.1109/ICPR.2018.8545718.

[27] J. Hou, G. Wang, X. Chen, J. H. Xue, R. Zhu, and H. Yang, “Spatial-temporal attention res-TCN for skeleton-based dynamic hand gesture recognition,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11134 LNCS, pp. 273–286, 2019, doi: 10.1007/978-3-030-11024-6_18/FIGURES/7.

[28] N. Tasnim, M. K. Islam, and J. H. Baek, “Deep learning based human activity recognition using spatio-temporal image formation of skeleton joints,” *Applied Sciences (Switzerland)*, vol. 11, no. 6, 2021, doi: 10.3390/app11062675.

[29] R. A. Hamad, M. Kimura, L. Yang, W. L. Woo, and B. Wei, “Dilated causal convolution with multi-head self attention for sensor human activity recognition,” *Neural Comput Appl*, vol. 33, no. 20, pp. 13705–13722, Oct. 2021, doi: 10.1007/S00521-021-06007-5/TABLES/14.

[30] Q. de Smedt *et al.*, “SHREC’17 Track: 3D Hand Gesture Recognition Using a Depth and Skeletal Dataset,” *Eurographics Workshop on 3D Object Retrieval, EG 3DOR*, vol. 2017-April, pp. 1–6, Apr. 2017, doi: 10.2312/3DOR.20171049.

[31] Z. Cao, G. Hidalgo, T. Simon, S. E. Wei, and Y. Sheikh, “OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields,” *IEEE Trans Pattern Anal Mach Intell*, vol. 43, no. 01, pp. 172–186, Jan. 2021, doi: 10.1109/TPAMI.2019.2929257.

[32] N. Convertini, V. Dentamaro, D. Impedovo, and G. Pirlo, “Sit-to-Stand Test for Neurodegenerative Diseases Video Classification,” *https://doi.org/10.1142/S021800142160003X*, vol. 35, no. 12, Sep. 2021, doi: 10.1142/S021800142160003X.

[33] G. Bishop, G. W.-P. of SIGGRAPH, undefined Course, and undefined 2001, “An introduction to the kalman filter,” *axon.cs.byu.edu*, 2001, Accessed: May 26, 2022. [Online]. Available: https://axon.cs.byu.edu/~martinez/classes/778/Papers/Kalman.pdf

[34] S. Bai, J. Z. Kolter, and V. Koltun, “An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling,” Mar. 2018, doi: 10.48550/arxiv.1803.01271.

[35] A. Vaswani *et al.*, “Attention Is All You Need,” *Adv Neural Inf Process Syst*, vol. 2017-December, pp. 5999–6009, Jun. 2017, Accessed: Nov. 11, 2021. [Online]. Available: https://arxiv.org/abs/1706.03762v5

[36] G. C. Y. Peng *et al.*, “Multiscale Modeling Meets Machine Learning: What Can We Learn?,” *Archives of Computational Methods in Engineering*, vol. 28, no. 3, pp. 1017–1037, May 2021, doi: 10.1007/S11831-020-09405-5/FIGURES/7.

[37] M. Alber *et al.*, “Integrating machine learning and multiscale modeling—perspectives, challenges, and opportunities in the biological, biomedical, and behavioral sciences,” *npj Digital Medicine 2019 2:1*, vol. 2, no. 1, pp. 1–11, Nov. 2019, doi: 10.1038/s41746-019-0193-y.

[38] Y. Wang, S. W. Cheung, E. T. Chung, Y. Efendiev, and M. Wang, “Deep Multiscale Model Learning,” *J Comput Phys*, vol. 406, Jun. 2018, doi: 10.48550/arxiv.1806.04830.

[39] A. Shah, E. Kadam, H. Shah, S. Shinde, and S. Shingade, “Deep residual networks with exponential linear unit,” in *Proceedings of the Third International Symposium on Computer Vision and the Internet*, Sep. 2016.

[40] F. Yu and V. Koltun, “Multi-Scale Context Aggregation by Dilated Convolutions,” *4th International Conference on Learning Representations, ICLR 2016 - Conference Track Proceedings*, Nov. 2015, doi: 10.48550/arxiv.1511.07122.

[41] C. O’Reilly and R. Plamondon, “Design of a neuromuscular disorders diagnostic system using human movement analysis,” *2012 11th International Conference on Information Science, Signal Processing and their Applications, ISSPA 2012*, pp. 787–792, 2012, doi: 10.1109/ISSPA.2012.6310660.

[42] V. Dentamaro, D. Impedovo, and G. Pirlo, “An Analysis of Tasks and Features for Neuro-Degenerative Disease Assessment by Handwriting,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 12661 LNCS, pp. 536–545, 2021, doi: 10.1007/978-3-030-68763-2_41.

[43] V. Dentamaro, P. Giglio, D. Impedovo, and G. Pirlo, “Benchmarking of Shallow Learning and Deep Learning Techniques with Transfer Learning for Neurodegenerative Disease Assessment Through Handwriting,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 12917 LNCS, pp. 7–20, 2021, doi: 10.1007/978-3-030-86159-9_1.

[44] D. Impedovo, G. Pirlo, F. Balducci, V. Dentamaro, L. Sarcinella, and G. Vessio, “Investigating the sigma-lognormal model for disease classification by handwriting,” *Lognormality Principle And Its Applications In E-security, E-learning And E-health, The*, pp. 195–209, Jan. 2020, doi: 10.1142/9789811226830_0009.

[45] P. Geurts, D. Ernst, and L. Wehenkel, “Extremely randomized trees,” *Machine Learning 2006 63:1*, vol. 63, no. 1, pp. 3–42, Mar. 2006, doi: 10.1007/S10994-006-6226-1.

[46] P. Stoica and Y. Selén, “A review of information criterion rules,” *IEEE Signal Process Mag*, vol. 21, no. 4, pp. 36–47, 2004, doi: 10.1109/MSP.2004.1311138.

[47] P. Batista and A. Pereira, “Quality of Life in Patients with Neurodegenerative Diseases,” *J Neurol Neurosci*, vol. 7, no. 1, 2016, doi: 10.21767/2171-6625.100074.

[48] A. Sabo, S. Mehdizadeh, A. Iaboni, and B. Taati, “Estimating Parkinsonism Severity in Natural Gait Videos of Older Adults With Dementia,” *IEEE J Biomed Health Inform*, vol. 26, no. 5, pp. 2288–2298, May 2022, doi: 10.1109/JBHI.2022.3144917.

[49] A. Sabo, S. Mehdizadeh, K. D. Ng, A. Iaboni, and B. Taati, “Assessment of Parkinsonian gait in older adults with dementia via human pose tracking in video data,” *J Neuroeng Rehabil*, vol. 17, no. 1, pp. 1–10, Jul. 2020, doi: 10.1186/S12984-020-00728-9/TABLES/6.

[50] Z. Zhang *et al.*, “Deep Learning based gait analysis for contactless dementia detection system from video camera,” *Proceedings - IEEE International Symposium on Circuits and Systems*, vol. 2021-May, 2021, doi: 10.1109/ISCAS51556.2021.9401596.

[51] A. Howard *et al.*, “Searching for mobileNetV3,” *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2019-October, pp. 1314–1324, Oct. 2019, doi: 10.1109/ICCV.2019.00140.

[52] “Pose detection | ML Kit | Google Developers.” https://developers.google.com/ml-kit/vision/pose-detection (accessed Nov. 07, 2022).