International Journal of Computational Intelligence Systems (2023) 16:98
https://doi.org/10.1007/s44196-023-00280-z

RESEARCH ARTICLE

# Motion Tracking in Diagnosis: Gait Disorders Classification with a Dual-Head Attentional Transformer-LSTM

**Mohsen Shayestegan[1] · Jan Kohout[2] · Kateřina Trnková[3] · Martin Chovanec[3] · Jan Mareš[1,2]**

Received: 27 January 2023 / Accepted: 27 May 2023


## Abstract

Gait and motion stability analysis in gait dysfunction problems is a very interesting research area. Usually, patients who undergo vestibular deafferentation are affected by changes in their dynamic balance. Therefore, it is important both patients and physicians are able to monitor the progress of the so-called vestibular compensation to observe the rehabilitation process objectively. Currently, the quantification of their progress is highly dependent on the physician's opinion. In this article, we designed a novel methodology to classify the gait disorders associated with unilateral vestibular deafferentation in patients undergoing vestibular schwannoma surgery (model of complete vestibular loss associated with imbalance due to vestibular nerve section and eventual labyrinthectomy). We present a dual-head attentional transformer-LSTM (DHAT-LSTM) to evaluate the problem of rehabilitation from gait dysfunction, which is observed by a Kinect. A system consisting of a key-point-RCNN detector is used to compute body landmark measures and evaluate gait dysfunction based on a DHAT-LSTM network. This structure is used to quantitatively assess gait classification by tracking skeletal features based on the temporal variation of feature sequences. The proposed deep network analyses the features of the patient's movement. These extracted high-level representations are then fed to the final evaluation of gait dysfunction. The result analytically demonstrates its effectiveness in classification evaluation when used in conjunction with state-of-the-art pose estimation and feature extraction techniques. An accuracy greater than 81% was achieved for given sets of individuals using velocity-based, angle-based, and position features for both the whole body and the symmetric features of the body.

**Keywords** Deep learning · Classification · Gait disorders · Vision transformer · LSTM · TPCNN

## Abbreviations

| | |
|---|---|
| Adam | Adaptive moment estimation |
| CNNs | Convolutional neural networks |
| DHAT-LSTM | Dual-head attentional transformer-LSTM |
| FLOPs | Floating-point operations |
| LSTM | Long short-term memory |
| MACs | Multiply-accumulate operations |
| MLP | Multilayer perceptron |
| MS-COCO | Microsoft common objects in context |
| RCNN | Region-based convolutional neural networks |
| ReLU | Rectified linear unit |
| ResNet | Residual network |
| RNNs | Recurrent neural networks |

Mohsen Shayestegan, Jan Kohout, Kateřina Trnková, Martin Chovanec and Jan Mareš contributed equally to this work.

✉ Jan Mareš
maresj@vscht.cz

Mohsen Shayestegan
mohsen.shayestegan@upce.cz

Jan Kohout
jan.kohout@vscht.cz

Kateřina Trnková
katerina.trnkova@fnkv.cz

Martin Chovanec
martin.chovanec@fnkv.cz

[1] Faculty of Electrical Engineering and Informatics, University of Pardubice, Nam. Cs. Legii 565, 530 02 Pardubice, Czech Republic

[2] Department of Mathematics, Informatics and Cybernetics, University of Chemistry and Technology Prague, Technicka 1905/5, 166 28 Praha, Czech Republic

[3] Department of Otorhinolaryngology, Charles University Prague, 3rd Faculty of Medicine, University Hospital Kralovske Vinohrady, Šrobárova 1150/50, 100 34 Praha, Czech Republic

Published online: 06 June 2023

| | |
|---|---|
| TPCNN | Triple parallel convolutional neural network |
| ViT | Vision transformer |

## 1 Introduction

Sequential data are common in a wide variety of domains, including medical analysis [1], weather prediction [2], and renewable energy systems [3]. Classification is one of the most frequently requested tasks in the analysis of sequential data due to its importance for industrial and scientific purposes. Several machine learning approaches can be used to solve sequential classification problems, including decision tree classifiers [4], ANNs [5], autoencoders and echo state networks [6], recurrent neural networks (RNNs) and convolutional neural networks (CNNs) [7], and long short-term memory (LSTM) [8, 9].

Recently, deep neural networks have been widely applied to model sequential data. Complex patterns have been extracted for sequential trends using deep learning methods based on convolutional neural networks (CNN) [10] and recurrent neural networks (RNN) [11]. However, RNN suffers from vanishing and explosion gradients when processing long sequences [12]. Long-short-term memory (LSTM) methods have been developed to address these problems [13, 14]. Transformer [15] was also recommended as a method for processing a sequence of data that uses the attention process. In contrast to RNN-based methodologies, the transformer does not evaluate the input in an ordered sequence. Instead, it analyses the entire sequence of the input and then uses self-attention processes [16] to learn the temporal connections of the sequence. This makes it more capable of identifying patterns with long-term dependencies on sequential data, which are challenging for sequence models.

### 1.1 Gait Disorder Analysis

Recently, biometrics of the human body have been analysed in clinical diagnosis, such as handwriting [17], speech [18], and gait [19]. Since each patient has a specific walking style, gait analysis can be a great statistic to determine pathological performance. Body or brain damage, ageing, or related disorders [19, 20], are factors that can directly affect the patient's movement and cause gait disorders. Therefore, accurate gait analysis can be very beneficial in a variety of diagnostic procedures in many branches of medicine (e.g. neurology and otorhinolaryngology, head and neck surgery, and orthopaedics). Therefore, gait analysis has recently been studied by many research groups because it combines deep knowledge in medicine and information engineering.

Gait disorder analysis is significantly dependent on temporal information and involves variable input sequences. Thus, analysis requires a model that is good at learning long-term dependencies. Although LSTM is used to work with the input of the sequence, but according to [21], it might not be sufficient to solve long-term dependencies. Alternatively, the Transformer has been broadly used in natural language processing projects and has reached state-of-the-art results. However, it has not been used in gait disorders analysis systems. According to [22], the use of positional encoding in the Transformer can create a high computational cost. Looking at the advantages and disadvantages of both algorithms, a hybrid model named DHAT-LSTM has been implemented to solve the classification problem in the gait analysis system. LSTM is used to obtain the hidden state of the features, while the Dual Attention in the Transformer encoder layer (DHAT) is used to improve the learning of the temporal information. The combination of DHAT and LSTM is adapted to improve the ability of the model to learn long-term dependencies in the analysis of gait disorders.

All the features of the body are significant, although it is believed that some features would be more distinct than others [23]. Birch et al. [24] found that the swing of the arm was the most important feature that helped in the analysis, which is the part of the upper body. Therefore, they determined that the upper body would also be useful for analysis. This emphasises that evaluation of the whole body, rather than just the lower body, can increase the set of characteristics. Veres et al. [23] recommended that upon assessment of the outline of the body, the static component should be considered to be the essential component. However, this study ignored the dynamic component (swing of the arms and legs) because it is considered less important, and the results show that reduction of the accuracy rate. Control and stability of the gait cycle are provided by the rotational behaviour of the body, where Collins et al. [25], Herr et al. [26], and McIntosh et al. [27] observed the effects of this rotational behaviour on the gait adaptability of a person walking. Their results indicate that greater forces affected the hip, ankle, and knee of those with noticeably increased ankle motion. Jokisch et al. [28] recommended that frontal angle standing is superior to half-profile and profile view by determining the optimal viewing angles of the individual. Jokisch et al. and Larsen et al. [28, 29] concluded that the most significant characteristic analysed was the frontal view of the subject.

### 1.2 Biomedical Background

The human balance system is multisensory and consists of complex mutual coordination and communication of multiple organ systems, such as the brain, the vestibular apparatus of the inner ear, the visual system, and proprioception. Balance disorders not only increase the risk of falls with eventual fractures and other types of injury, but also significantly disrupt the quality of life of the patient,

who (depending on the severity of the disorder and its development) may be unable to perform in employment, sports, and also common physical tasks such as household chores or hygiene habits. Diagnoses that disrupt patient stability involve multiple disciplines (e.g. otorhinolaryngology, neurology, orthopaedics, and physical therapy) and require close cooperation.

With modern diagnostic methods, specialists from different disciplines are able to characterise the type of pathology responsible for the balance disorder (e.g. uncompensated vestibular lesion) and quantify the severity of the individual deficit of the concrete organ (e.g. presence of spontaneous nystagmus on videooculography, positive head shake test, saccades and gain on Video Head Impulse Test, deviation in subjective visual vertical or sway analysis in posturography).

Modern diagnostic methods are able to objectively assess only the static part of equilibrium (stabilometry, posturography) [30–32]. Questionnaires or clinical tests (e.g. Timed Up and Go Test or Six-Minute Walk Test) are the currently employed to evaluate dynamic stability. However, they are of limited utility for quantifying gait and motion stability in the normal life of the patient. Furthermore, these methods cannot be processed quantitatively and are subject to a certain degree of subjectivity by the examiner [33–36]. In clinical practice, there is as yet no examination that objectively evaluates the dynamic component of stability as a whole.

In the course of compensation of balance dysfunction, pharmacological treatment is initially important, but after its relatively short period of application, targeted rehabilitation of balance is the most important part. In case of disbalance related to peripheral vestibular disorders it is a training of slow and later also rapid head movements and image stabilisation initially. Gradually sitting independently and then standing, but walking is essential in everyday life. For each patient, these individual phases proceed at different speeds and some exercises are not yet suitable for him. That is why the patient with new lesion of the vestibular system will not initially cope with all parts of rehabilitation compared to the patient who has developed a lesion a couple of weeks or months ago. The development of an objective test to quantify balance in gait, to determine the appropriate rehabilitation procedure and standardise the recommended exercises in order to individualise the care of each patients is necessary. In addition, the evaluation using the clinical tests used so far is observer-dependent and does not allow to monitor the development over time in a meaningful way. For similar reasons, dynamic computed posturography has been introduced into clinical practice, for example, to enable objective assessment of postural stability in stance. Therefore, considerable efforts are centred on development of systems that would allow the assessment of stability in locomotion, especially in walking, that is one of the most characteristic human movements.

### 1.3 Proposed Method

Human movements range from simple activity through an arm or leg to the complex integrated activity of combined arms, legs, and body [37]. This movement is represented by a sequence of frames, which observers can understand by analysing the contents of multiple frames in sequence. In this paper, we classify gait disorders in a way similar to a clinician's observation of the patient's gait. We use DHAT-LSTM to consider the information of previously recorded frames to automatically understand movements in the currently recorded frames. To design the DHAT-LSTM, we were inspired by the Transformer encoder architecture [15] and the LSTM network [38, 39].

In the proposed method, selected features of recorded frames are analysed for gait disorders. The initial features of every frame are extracted using a pre-trained key-point-RCNN model based on the Mask RCNN paper [40]. The DHAT-LSTM architecture is then developed with two layers of dual-head attention to learn sequence information about the features of the recorded frames. This implementation of DHAT-LSTM has a high capacity to learn sequences and frame-to-frame changes in features due to small changes in the visual data of frames.

### 1.4 Main Aims

Our primary objective was to improve the accuracy of our system. To achieve this, our framework consisted of two main components: data processing using the recorded image dataset and the development of a deep learning algorithm for classification. Our aim was to enhance the practicality of our system by increasing its accuracy compared to existing evaluation methods.

The significant contribution of our Transformer approach lies in its evaluation of both global and local movement features of the patient's skeletal characteristics during three activities. This evaluation is accomplished using a dual-head attention transformer, which is integrated with an LSTM network for gait function assessment. Additionally, we introduce human key-point extraction to generate new and relevant features. These features include the speed and angle of each point in two continuous sequences , as well as the distance and angle of symmetric features within each sequence.

The proposed framework comprises the following key components:

1. Data processing scheme: This involves extracting features and mapping the input sequential data into a vector of specific dimensions. Furthermore, feature engineering is applied to the extracted features to optimise their qual-

ity and improve the performance of the deep network for the final classification.

2. Dual parallel self-attention transformer mechanism: This mechanism operates on the split input vector and is integrated with LSTM networks . It calculates a temporary vector that captures the temporal dynamics of the data.

By combining these components, we aim to achieve higher accuracy and enable more effective classification in our system.

## 2 Materials and Methods

### 2.1 Measurement Scheme

The measured process consists of the different walking exercises presented in Table 1. Similar to standard clinical testing of gait three conditions with different requirements on balance system were subject of analysis. Normal walking is the least demanding, in the clinical setting it is known as the gait test I. Walking along the line with visual control (clinical test of tandem gait) is devoted to assess both vestibular and cerebellar functions. Finally, walk with the eyes closed correspond to clinical test of gait II. It is crucial to assess the impact of visual feedback on balance functions.

A patient was asked to perform these exercises in a hospital hallway (length 5 m) and the frames of the patient's gait were recorded. Length of recording was dependent on the velocity of walking to perform a specific test along the entire length of the corridor. In general, probands were asked to walk at a leisurely pace so that they will feel supremely balance confident. Furthermore, patients rehearsed the individual tests before actually performing the recording.

**Table 1** Hospital control rehabilitation exercises

| Order | Exercise |
|---|---|
| 1 | Normal walking |
| 2 | Walking along a line |
| 3 | Walking with closed eyes |

The patient is scanned at the same time with a Kinect camera, and the exercise sequences are stored in data files for further analysis. Kinect is placed on a mobile robotic platform that was developed at UCT Prague and can track a patient moving behind it. The whole measurement scheme is shown in 1. The robotic platform is described in more detail in our previous work [41].

The Kinect v2 camera was used because of its ability to capture video and create skeletons. However, it is not able to work correctly in real-time. This is the reason why preprocessing has to be employed (see 2.4). A typical patient during measurement (doing three exercises on the hospital corridor) is shown in Fig. 2. The first reason for doing these three exercises is that the category of disorders can be classified by observing these three exercises, as the clinicians are measured, so we should have all three exercises in each class. The reason for combining them as input data for the network is to make the input data have more sequences of features. As can be seen from Fig. 2, the type of walking in different exercises is slightly different. For example, for exercise 3, the patients walk with closed eyes, and they may not walk in a straight line, or they may use their hands to make their movement balance. Therefore, the combination of these data would help to have a long sequence of their walk in a model to train to find the pattern of difference. Gait data in the form of skeletal key-point landmarks were obtained during the course of the patient's postoperative rehabilitation.

### 2.2 Dataset

The dataset contains 84 successful rehabilitations in 37 patients (Table 2). All patients (23 males/14 females; age 21-77 years) were indicated for surgery for vestibular schwannoma (24 left side/13 right side; Koos classification: 10 grade 1/10 grade 2/6 grade 3/11 grade 4a tumours; International classification: 10 grade T0/9 grade T1/8 grade T2/9 grade 3/0 grade 4/1 grade 5 tumours).

All patients underwent standard neurootological tests (i.e. clinical neurootologic examination, Videoculography, Caloric Testing, Video Head Impulse Test, Ocular and Cervical Vestibular Evoked Myogenic Potentials, Subjective

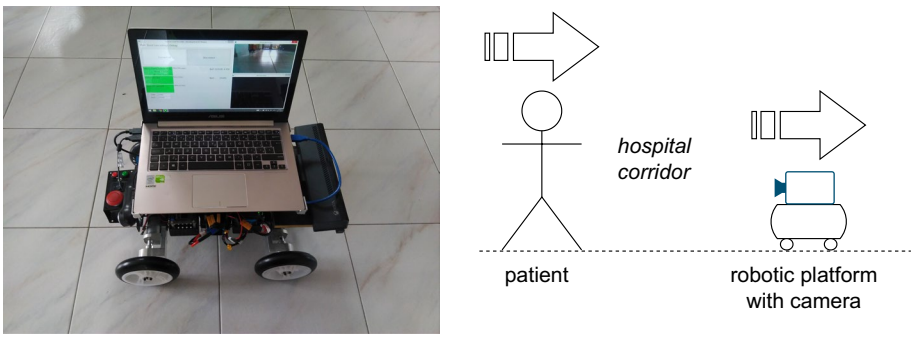


**Fig. 1** Measurement scheme

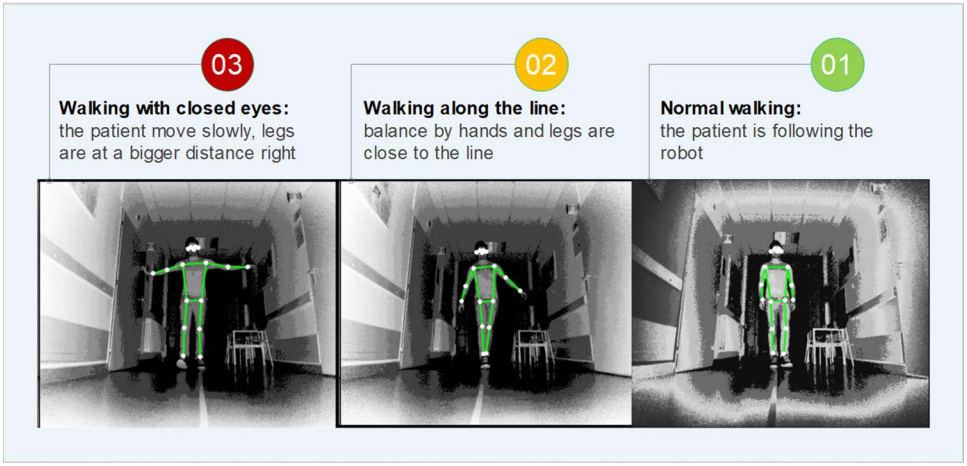


**Fig. 2** Typical patient with three forms of exercises

**Table 2** Dataset overview

| | |
|---|---|
| Start date | 2018-01-16 |
| End date | 2020-12-13 |
| Number of patients | 37 |
| Number of sessions | 84 |
| Man | 23 |
| Woman | 14 |
| Average age | 56.6 |

Visual Vertical) to quantify vestibular deficit . Vestibular schwannoma is a tumour that gradually disrupts the patient's balance due to the developing vestibular deficit. Balance dysfunction usually affects patients with larger tumours, but it can affect patients even at the very beginning of tumour growth. Based on the clinical and neurotological testing patients were divided into 3 groups. Patients with no disbalance presented without the deficit on caloric testing, there was either normal function or eventual compensated pathology on Video Head Impulse Test, and no balance problems in Romberg and Tandem gait tests. Patients with light disbalance have presented with caloric hyporeflexia, partial but compensated pathology on Video Head Impulse Test and instability in Tandem gait test with eyes closed but no problems in Romberg test. Finally, patients with heavy disbalance presented with areflexia on caloric testing, spontaneous nystagmus, non-compensated pathology on Video Head Impulse Test and pronounced instability in both Romberg and Tandem gait test. To manage tumours, the retrosigmoid approach was performed in seven patients, the retrolabyrinthine approach was used in four patients, and the translybyrinthine approach was used in 26 patients. Patients with associated orthopaedic, neurologic, and ophtalmologic comorbidity with an impact on the patient's balance were not enroled in the study. To manage tumours, the retrosigmoid approach was performed in seven patients, the

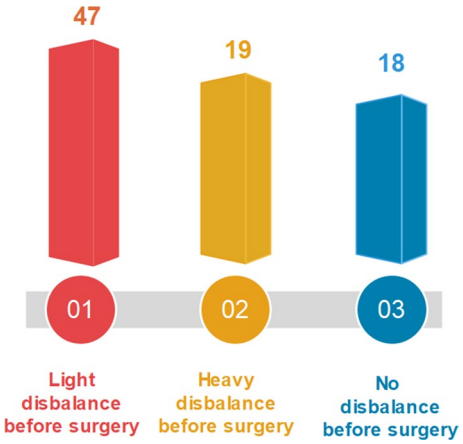


**Fig. 3** Classes distribution in our dataset

retrolabyrinthine approach was used in four patients, and the translybyrinthine approach was used in 26 patients. Patients with associated orthopaedic, neurologic, and ophtalmologic comorbidity with an impact on the patient 's balance were not enroled in the study.

## 2.3 Dataset Classes

Figure 3 shows the grade score distribution (classes) of our dataset. The samples in Fig. 4 are the amount of data, which totalled approximately 84 measurements, where each measurement included three exercises.

As can be seen in Fig. 4, 47 samples (34 samples for the training dataset and 13 samples for the test dataset) presented with a light disbalance before surgery (class 1), 19 samples (13 samples for the train dataset and 6 samples for test dataset) presented with heavy disbalance before surgery (class 2), and 18 samples (11 samples for train dataset and 7 samples for test dataset) presented with no disbalance before surgery (class 3). Each folder presents a patient's walking

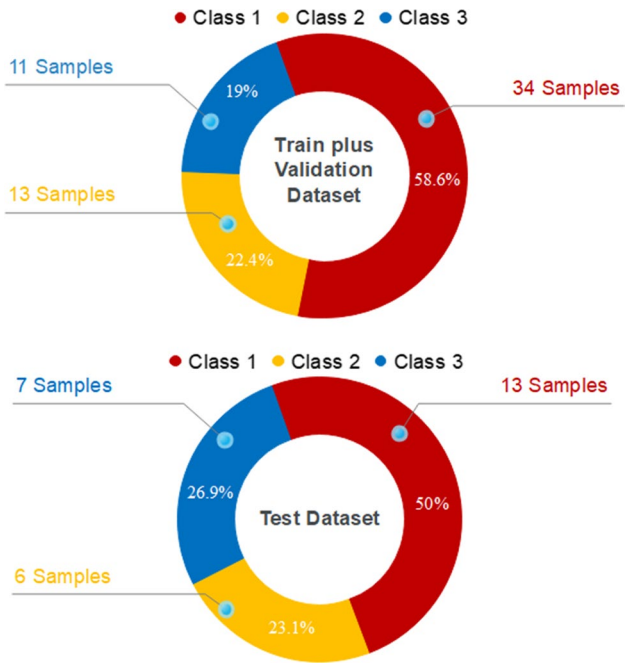


**Fig. 4** Number of samples for train and test data

along a straight line in a hallway. The patients had performed three exercises during the examination. Time steps or sequences are expressed as the number of frames recorded by the camera. Skeletal key-points are the number of features in sequences recorded from measurements that are the initial input before feature engineering.

## 2.4 Data Preprocessing

Figure 5 presents the overall framework of our research project. Figure 6 shows the architecture of the proposed data processing as the first three steps of the whole process before training our dataset, as follows: data preprocessing, feature engineering, and splitting and storing of the final dataset. The data preprocessing step applies different image filters to get better detection results. We have applied the Equalhist, GaussianBlur, and applyColorMap filters from the OpenCV library. By applying these filters, the contrast of the image will be adjusted, and unwanted noise be removed from an image which can improve the contrast and the visual quality and clarity of the image which results in more collected detection sequences as the detector perform better with filtered images in some frames. The filtered image is then used as an input for key-points detection to extract the skeleton key-points and then apply the Kalman filter to track missing

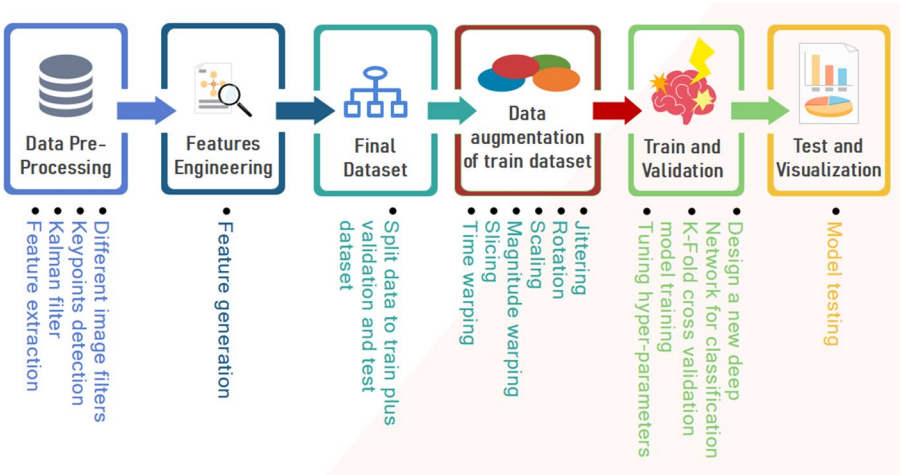


**Fig. 5** Overall framework of the proposed research project

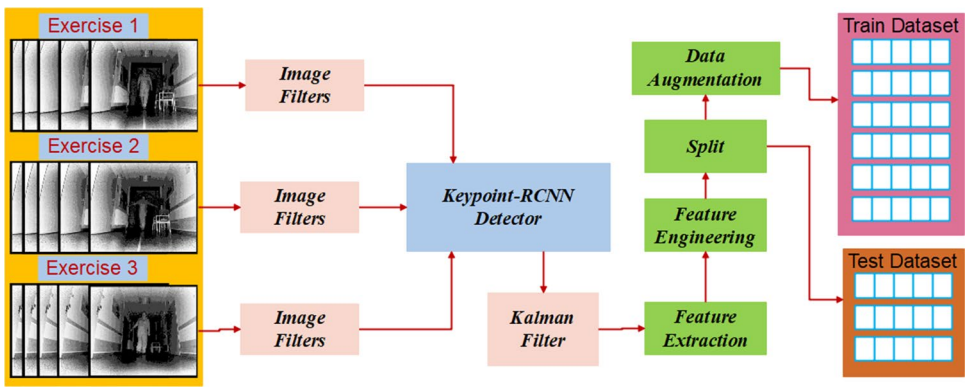


**Fig. 6** Architecture of the proposed data processing

key-points. These extracted features proceed to feature engineering and data augmentation. Before data augmentation (5), the data with extracted engineered features are split into training and test datasets (6), and are then stored for the next steps.

In this paper, a key-point-RCNN detector has been applied using the Torchvision library to collect the human skeletal key-points on images. The model is built on top of the ResNet-50 FPN backbone [42]. The purpose of applying key-point detection is to detect key-point positions of the body joints in a patient in an image. In the Mask-RCNN paper [40], the authors extended the model's capabilities to detect key-points in the person's body. A slight modification in the Mask-RCNN introduced a new solution for key-point detection. Figure 2 demonstrates how key-point detection helps in determining the right body part poses of a patient during exercises.

The key-point-RCNN is trained on the MS-COCO dataset [43], which proposed 80 classes for detection and segmentation, however, for key-point detection, the annotations are only presented for the person class. By running a key-point-RCNN detection model, if a patient is detected in the image, then there will be 17 key-points on the patient's body [40]; as illustrated in Fig. 7 and Table 3. We scheduled a window size as a maximum number of sequences that we want to collect the extracted features. The window size is configured for 100 sequences. We also applied the Kalman filter to track the feature key-points if the detection failed to detect all of the feature points.

As can be seen in Fig. 3, the class distribution in our dataset is imbalanced: there are more patients with class 1 than patients with other classes. To tackle unbalanced datasets, we applied the techniques that are recommended in [44] (e.g. optimum weights for each class in the training process). In addition, because the accuracy is not consistent for different iterations in the unbalanced dataset, we used the F-score metric, which is proper for an unbalanced dataset [44]. Due to our small dataset and to prevent overfitting, we used data augmentation methods as shown in Fig. 5 and Table 4, including jittering, slicing, scaling, magnitude warping, and time dimension warping, which is recommended in Ref. [44].

**Table 3** Points of interest (key-point marks)

| Point | Position | Point | Position |
|---|---|---|---|
| 1 | Nose | 10 | Left wrist |
| 2 | Left eye | 11 | Right wrist |
| 3 | Right eye | 12 | Left hip |
| 4 | Left ear | 13 | Right hip |
| 5 | Right ear | 14 | Left knee |
| 6 | Left shoulder | 15 | Right knee |
| 7 | Right shoulder | 16 | Left ankle |
| 8 | Left elbow | 17 | Right ankle |
| 9 | Right elbow | | |

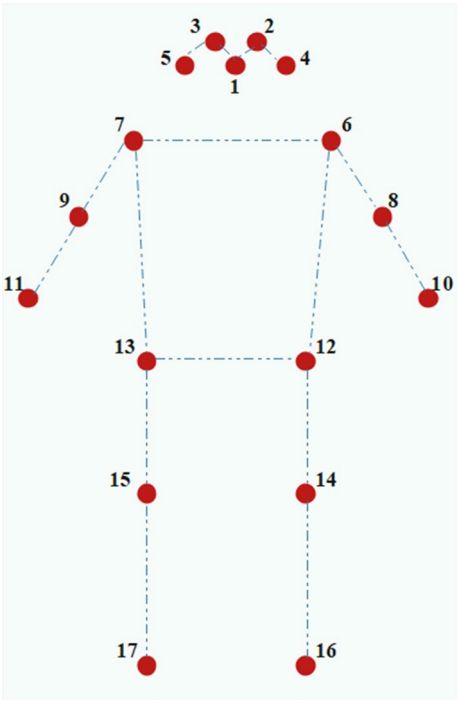


**Fig. 7** Skeleton positions and numbering of an individual used for data processing

Generally, the detector model takes images as an input and returns the 2D pixels coordinates of the skeletal key-points of the patients with respect to the image frame [40]. Initially, a network predicts 2D confidence maps of the positions of body parts and a set of 2D vectors for the related parts [40]. Each confidence map then returns the key-points of the body parts within the image [45].

In the last step of key-point extraction, a Kalman filter [46, 47] is applied to the detection result, of which every step is taken. This is needed because the result in frames may suffer from missing points and high coordinate variance, due to the small inaccuracies of the proposed pose detection algorithm. A Kalman filter [46, 48] is a valuable and frequently used estimation algorithm. Using past estimates of variables, it can predict the future state of the variables. Performing this task for all key points smooths the sequence of the coordinates [48].

By carefully observing the patient walking and referring to abnormal gait, all features are calculated for each frame. Sequences containing the measurements relating to each feature point are subjected to four statistical measurements (i.e. x, y, velocity and angle for each point in two different sequences). In addition to this data vector, an additional data vector was created by distance-based and angles between

**Table 4** Data augmentation methods

| Methods | Description |
|---|---|
| Jittering | One of the effective data augmentation methods |
| Rotation | Increase accuracy when combined with other augmentation methods |
| Scaling | Can change the global intensity of a time series |
| Magnitude warping | Warps the magnitude of a signal by a smoothed curve |
| Slicing | Slicing time steps off the ends of the pattern |
| Time warping | Perturbation of a pattern in the temporal dimension using a smooth warping |

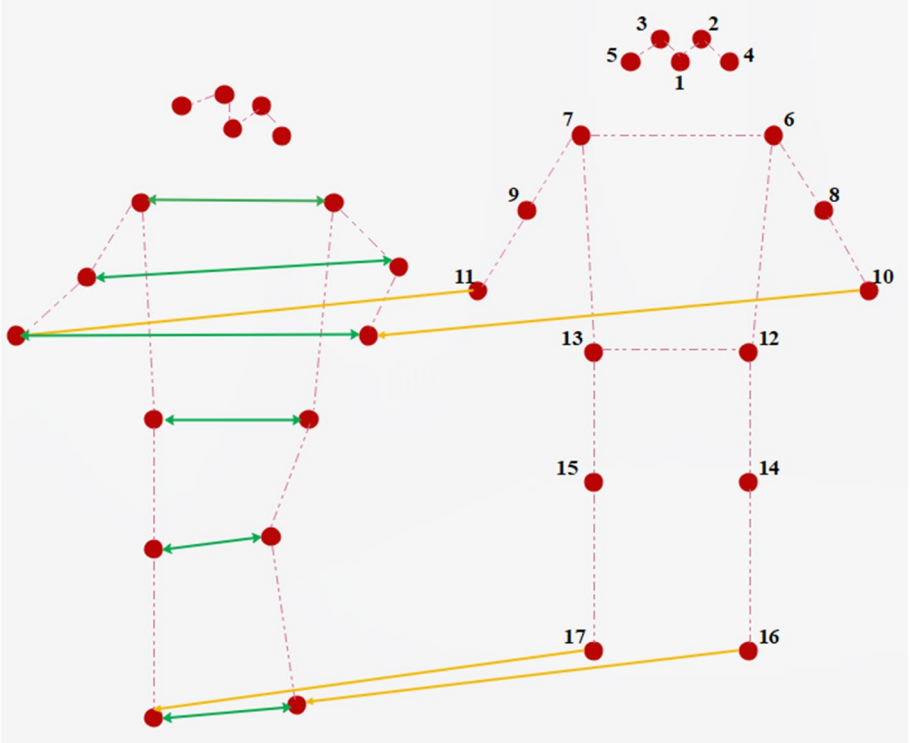


**Fig. 8** Movement of a patient and the skeletal key-points in two sequences

symmetric feature points. Figure 8 illustrates the movement of a patient in two sequences where each point moves to different pixels coordinates. The yellow arrows show the differences of each point in terms of position, angle, and velocity in two sequences, and the green arrows show the displacement and angle of symmetric key-points in the same sequence.

These spatio-temporal features are synthesised from the sequential of coordinates of each detection step by observing the walking of patients in the defined three exercises. These created features use the following statistical measures [48]:

$$d_i = \sqrt{\Delta x_i^2 + \Delta y_i^2} \tag{1}$$

$$v_i = \frac{d_i}{\Delta t_i} \tag{2}$$

$$\alpha_i = atan2(\Delta y_i, \Delta x_i) \tag{3}$$

where $d_i$ is distance-based of symmetric key-points, $v_i$ is velocity of key-points in two sequences and $\alpha_i$ is angle between detected key-points. As can be seen from Fig. 9, there are 17 key-points in each sequence and each key-point has two axes of data. The velocity and angle of these points are computed from two continuing sequences. This means that there are a total of 68 variables for each time step. In addition, the distance and angle of the symmetrical key-points pair are added to these features. The final feature vector included 84 features for each sequence. Furthermore, each series of data has been partitioned into 100 time steps (100 are sequences), and there are three exercises for each labelled case. This means that the total time steps are around $100 \times 3$, or 300 steps. Therefore, the sample data have $300 \times 84$ or 25,200 elements. This is exactly how we loaded the data, where one sample is one window of sequential data, each window has 300 time steps, and the time step

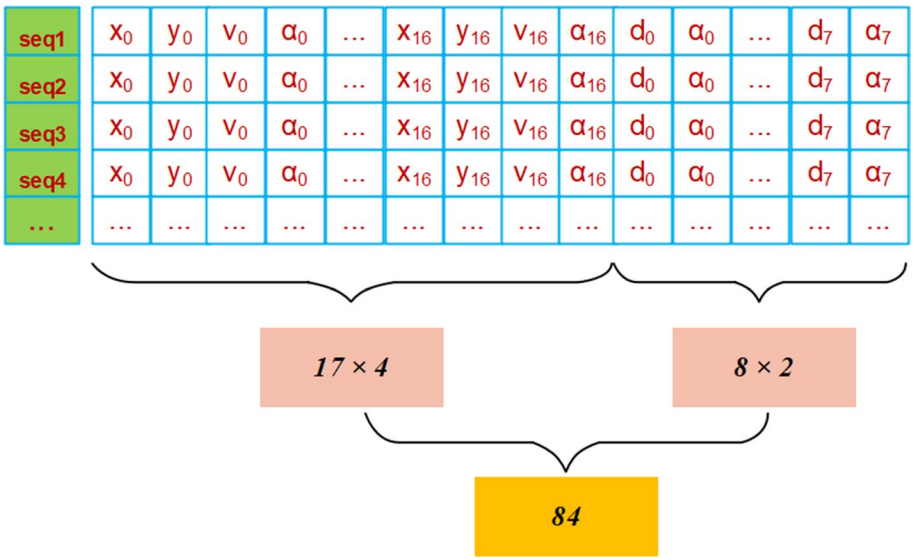


**Fig. 9** Created feature vector as input

has 84 features. The output of the model will be a three-element vector that contains a given window that belongs to each of the three classes.

## 2.5 Classification Methodology

During the clinical assessment, the dynamic stability of the patient is evaluated when walking under three standard exercises. To perform the analysis of gait dysfunction, we proposed a DHAT-LSTM to evaluate the high-level human skeletal movement features based on human body actions.

In this study, we introduce a dual-head attentional transformer integrated with the LSTM network. The proposed model consists of dual-head attention transformer blocks. The dual-head attention layer as a self-attention mechanism is applied to the sequence of extracted features. Instead of having one single self-attention block, we have two computations in parallel.

Figure 10 explains the framework of the proposed DHAT-LSTM. The proposed framework and its main mechanisms include feature extraction through key-point-RCNN and classification of gait movements from the sequence of frames. In the procedure for the classification of gait dysfunction, we

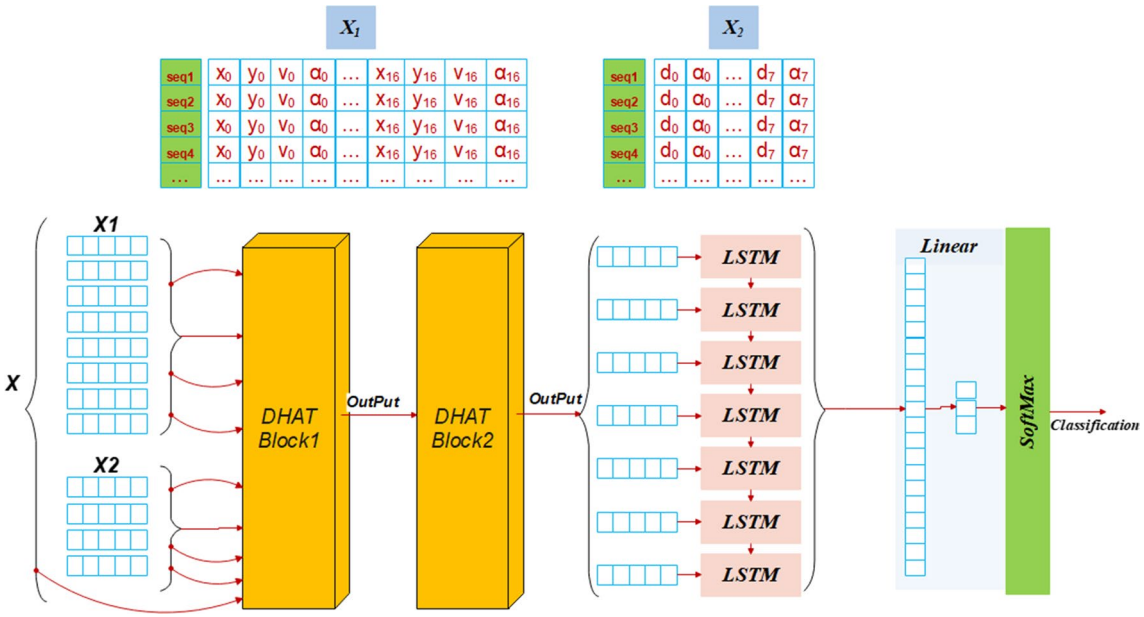


**Fig. 10** Representation of DHAT-LSTM framework

first extract features from the frames and then the generated features representing the sequence of action are fed to the proposed DHAT-LSTM. Finally, it is analysed for the final classification of the recorded frames.

Each individual frame is represented by key-point-RCNN features, which is followed by finding sequential information between them using DHAT-LSTM. From our observation and analysis of the detector performance on our dataset, hundreds of sequences of detected features are collected for each exercise, together with the specific recorded time of each frame as the input of the proposed network. The generated feature extraction input that is fed to the proposed model will be divided into two parts, which are created features from two sequential frames and created features extracted from symmetrical key-points. When they are input, these features are preprocessed by the data-processing scheme and fed as two different input vectors to the proposed network to extract the dynamic spatio-temporal features. The result of these two parallel stages is then merged and normalised as a final output to be used as input for the LSTM network.

Each DHAT block is composed of two sublayers: a concatenation of the double head attention sublayer and a fully connected MLP sublayer. Each sublayer is followed by a normalisation layer. Raw input data are processed by the data-processing scheme. The static and dynamic features measurements of the body are applied to assess the gait. The static features of the body include the distance between the patient's shoulder key-points or hip key-points.

Dynamic features are measurements related to gait. For example, the distance and angle between the left and right knees during gait. We take this as an output of the data-processing scheme and use it as a DHAT layer input. Then, the outputs of step 1 are fed to the DHAT block 1, including the double head attention layer. Because the transformer encoder layer usually uses positional encoding to remember the sequential data, it learns from the beginning of each sequence, and this leads to high computational costs [49]. In this research, a combination of the transformer concept and LSTM was investigated to solve the long-term dependency classification problem. We observe that the performance of other layers is rather similar by removing the positional encoding, and even a little enhances the performance of the proposed model. The combination of both blocks was designed for better learning of the temporal information to resolve the long-term dependencies of gait disorders.

In dual-head attentions, one head is used for the created features of two sequences and the other is used for created features of symmetrical key-points. This expands the model's ability to focus on different positions. With double headed attention, we have double sets of Query/Key/Value weight matrices. Each of these sets is initialised randomly. If we do the same self-attention calculation that we represented above, then just with different weight matrices, we end up with two different $Z$ matrices. Therefore, the results concatenate the matrices and then multiply them by an additional weight matrix $W_o$.

Addition and layer normalisation operations are then performed. The result is then fed to the MLP layer, and the addition and layer normalisation operations are performed once again to compute the final output, which is used as part of the input to the DHAT block 2. The same process as step 2 is applied to DHAT block 2 and its final output is used as the input to the LSTM layers. Finally, the results from LSTM layers are used as an input to the decoder layer, which consisted of a linear layer that produced the output, and are then passed to the Softmax layer to classify gait disorders.

### 2.5.1 DHAT Layer

The DHAT layer tries to capture temporal dependencies in the input feature vectors. As shown in Fig. 11, the attention mechanisms are the key parts of the DHAT block.

In the proposed framework, the output from the data-processing stage is divided into two parts, denoted as $X_1 = x_1, \ldots, x_{d_1} \in R^{d_1 \times N}$ and $X_2 = x_1, \ldots, x_{d_2} \in R^{d_2 \times N}$, and serve as the inputs of DHAT; as shown in Fig. 11. Here, $d_1$ and $d_2$ are the total number of features of the inputs vector 68 and 16, respectively, and N is the length of $x_i$ in $X_1$ and $X_2$, which is 300. DHAT expands the model's capability to focus on two positions by splitting the input features into two partitions, increasing the representation subspaces.

In the DHAT block, each self-attention is calculated to create three vectors from each of the input vectors of the encoder. Therefore, for each feature vector, we create a Query vector, a key vector, and a value vector [15, 50]. These vectors are created by multiplying the input vectors by the three matrices that we trained during the training process. We then need to score each feature vector of the input against each other to determine how much attention to place on other parts of the input vectors. The score is determined by getting the dot product of the query vector with the key vector of the relevant feature vector that is scoring. Then, to have more stable gradients, the scores need to be divided by the square root of the dimension of the key vectors used. Using the Softmax operation, we can determine how much each feature vector will be expressed. To keep the values of the feature vector that we want to focus on, we multiply each value vector by the Softmax score [15, 50]. In each subspace head, we calculate each attention head ($Ah_1$ and $Ah_2$) as follows and according to [15]:

$$Ah_1(Q_1, K_1, V_1) = Softmax\left(\frac{Q_1 \times K_1^T}{\sqrt{d_1}}\right) \times V_1 \tag{4}$$

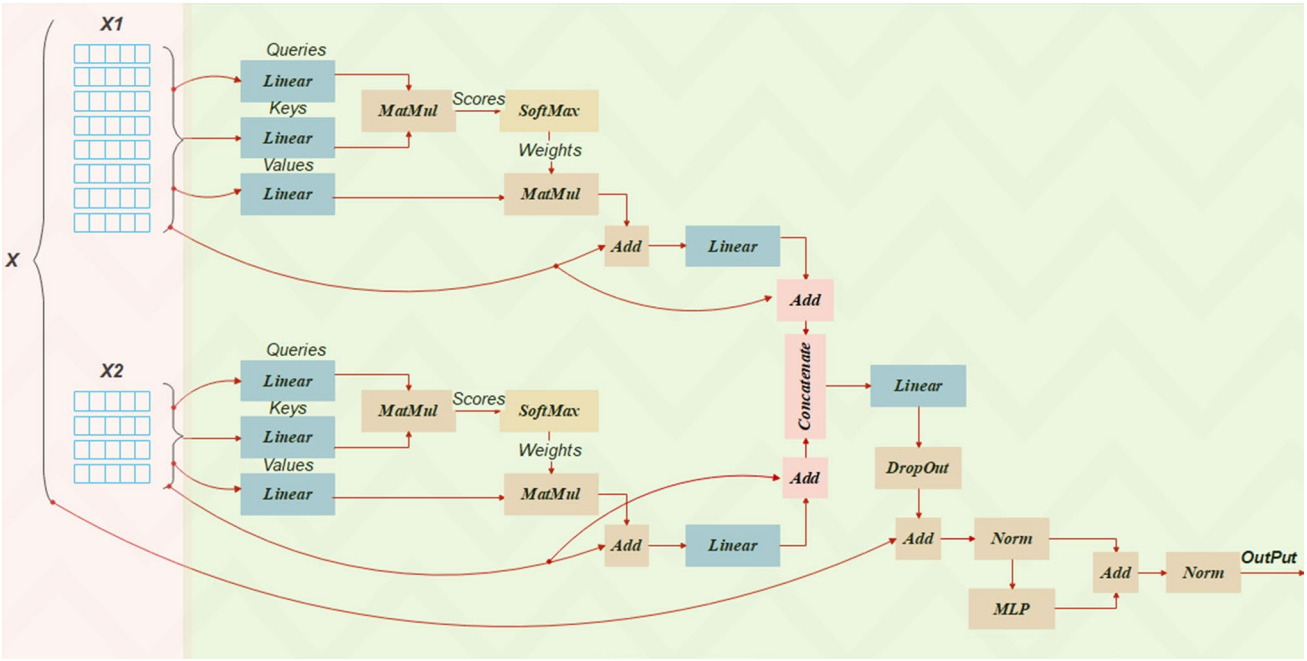


**Fig. 11** The structure of a DHAT block

$$Ah_2(Q_2, K_2, V_2) = Softmax\left(\frac{Q_2 \times K_2^T}{\sqrt{d_2}}\right) \times V_2 \tag{5}$$

Finally, these two head representations are concatenated together to produce the final output, as follows:

$$DHA(X_1, X_2) = Concat(Ah_1, Ah_2) \tag{6}$$

This result is then multiplied by an additional weights matrix (a linear after concatenation) and after that we apply a drop out to prevent overfitting. The result then adds to the whole input feature vector.

The DHAT has two Normalise layers and add, which add the output of the previous layer to the input of that layer through a residual connection [51], and then normalise the sum by applying layer normalisation [52]. It is possible to accelerate the training process by performing normalisation operations. The residual connections enable the model to transmit effective lower layer features to the higher layers [50]. For any vector v, the layer normalisation is computed as

$$LayerNorm(v) = \gamma \frac{v - \mu}{\sigma} + \beta \tag{7}$$

$$\mu = \frac{1}{d} \sum_{k=1}^{d} v_k \tag{8}$$

$$\sigma^2 = \frac{1}{d} \sum_{k=1}^{d} (v_k - \mu)^2 \tag{9}$$

in which $\mu$ and $\sigma$ are the mean and standard deviation of the elements of the vector $v$, where the scale $\gamma$ and the bias vector $\beta$ are parameters. The output of this Normalise layer is fed into an MLP neural network, which is a combination of a two-layer feed-forward network [50] with a ReLU activation function. Given a sequence of vectors $v_1, \ldots, v_n$, the computation of an MLP sublayer on any $v_i$ is defined as

$$MLP(v_i) = ReLU(v_i \times W_1 + b_1) \times W_2 + b_2 \tag{10}$$

where $W_1$, $W_2$, $b_1$, and $b_2$ are parameters that refer to the two fully connected layers.

### 2.5.2 LSTM Layer

The LSTM layers have been used in combination with DHAT blocks to improve the results. LSTM is a kind of RNN network that can overcome the exploding gradient and vanishing gradient problems in the long-time sequence [53]. Integrating LSTM with DHAT can improve the performance even further, which implies that the proposed model has a better understanding of temporal dependencies.

The LSTM block structure in a single sequence [49] is shown in Fig. 12. By adding gate units and a memory cell, the LSTM network can save and transmit information over a long time [49, 54]. The deep structure of LSTM networks

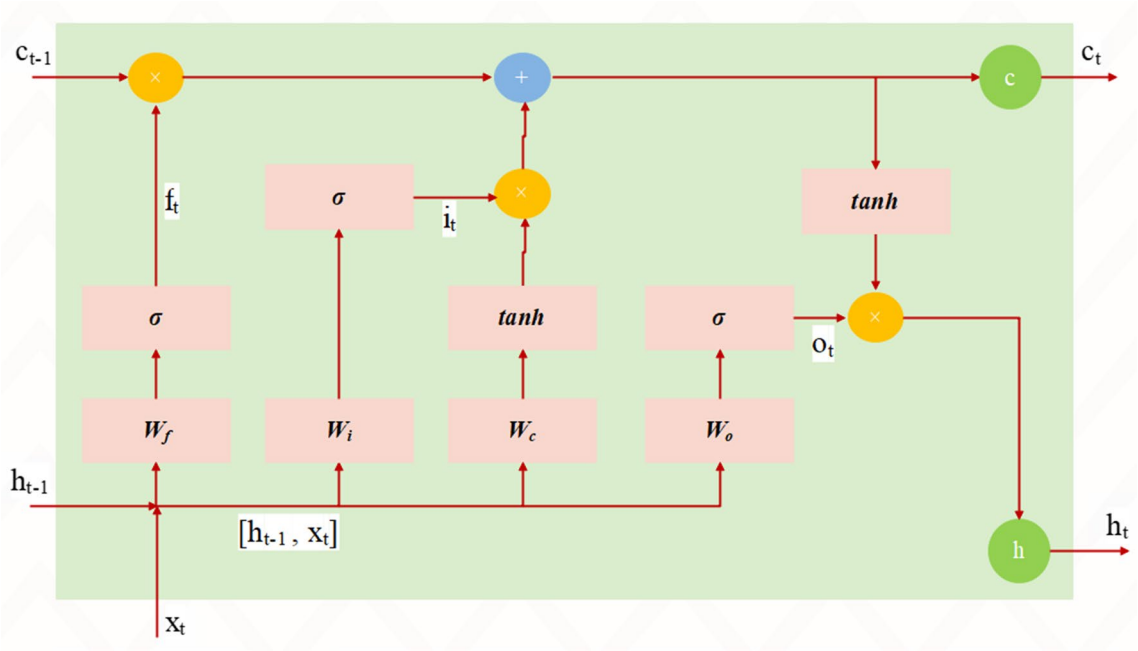


**Fig. 12** The structure of a LSTM block

automatically extracts the relationship between the various state parameters throughout training to obtain the prediction [49].

The final output of the LSTM is determined by the state of the output gate and the memory cell [49], as follows:

$$c_t = f_t \odot c_{t-1} + i_t \odot \tanh(W_c \times [h_{t-1}, x_t] + b_c) \tag{11}$$

$$h_t = o_t \odot \tanh(c_t) \tag{12}$$

where the input gate ($i_t$) determines the impact of $x_t$ on $c_t$, and the state $o_t$ is the output gate that is used to control the impact of $c_t$ on $h_t$ [49]. $W_c$, and $b_c$ are the weight matrix and the bias element of the input memory cell state, respectively. In accumulation processing, the forget gates limit the memory cell information, and additional information depends on the input gate, which is used for restriction. This accumulation of information depends on both the hidden state and the memory cell connection [49].

Finally, a feed-forward linear layer with an input size of 168 and output size of 3 has been included in the classifier, as shown in Fig. 10.

## 3 Results

In this section, to demonstrate the performance ability of the proposed approach, it has been compared with TPCNN [44] and Vision transformer (ViT) [55]. In paper [44], the authors designed a triple-path convolutional neural network (TPCNN) to evaluate the problem of facial paralysis. The ViT method was introduced by the Google team, who applied a pure transformer directly to sequences of image patches for image classification. For ViT, instead of using fixed-size patches of image (2D), we used the 1D network (with 2 layers and 2 heads) with 1D input feature vectors (84 features in each vector), and embed each of them and add the position embeddings. Finally, we feed the resulting sequence of vectors to a standard Transformer encoder by adding a classification layer, as illustrated in [55].

For the TPCNN classifier, the same structure of its network was attached to the input layer. Due to the structure of TPCNN [44], which included three parallel networks structure for different region of body (global region and local regions), we fed 84 feature length for one network line and fed the split features from the upper body features (created feature from points 1–11 shown in Fig. 7) and the lower body features (created feature from points 12–17 shown in Fig. 7) from the extracted features to other two network lines. The tenfold cross-validation process is applied to validate unseen key-point feature samples, while decreasing the possibility of overfitting to previously seen samples. This technique divides the dataset into 10 subsets. The datasets were shuffled randomly and split into tenfold of almost equal size. Each subset is included as validation data and the other nine subsets are used as training data. This procedure is repeated 10 times, and each class has the same probability of validation. In the last layer of the proposed framework, cross-entropy and soft-max are the loss and activation functions, respectively. The evaluation score provided by a clinician is used as a basis for training, and the extracted

feature vectors of the model are used to evaluate gait disorders. For the loss function, we have applied the standard multi-class cross-entropy for our model as follows:

$$L = \frac{-1}{M} \sum_{k=1}^{K} \sum_{i=1}^{M} y_i^k \log(\hat{y}_i^k) \tag{13}$$

where $y_i^k$ is the actual class for the $i_{th}$ sample and $\hat{y}_i^k$ is the predicted probability of the $i_{th}$ sample for class k, M is the total number of samples and K is the number of classes.

Due to the imbalance dataset, the loss function in multi-class cross-entropy equation penalises the miss-classification of all the classes, so the trained model may be biased toward classes with higher samples. To address this issue, we modified the loss function by assigning higher weights of classes with a few samples and lessen the weights for classes with higher samples to compensate for the misclassification by using a weighted cross-entropy loss, as follows:

$$L = \frac{-1}{M} \sum_{k=1}^{K} \sum_{i=1}^{M} w_k y_i^k \log(y_i^k) \tag{14}$$

$$w_k = \lambda \left(1 - \frac{M_k}{M}\right), \lambda = 2.5 \tag{15}$$

where $w_k$ and $M_k$ are the weight and the number of samples in class k, respectively. The choice of class weight $w_k$ depends on the number of samples of this class controlled by $\lambda$ and $\frac{M_k}{M}$. Class 1 has a greater number of samples, while classes 2 and 3 have fewer samples. Therefore, we assign the highest value to classes 2 and 3, and the lowest to class 1. Furthermore, we also used F-score as an evaluation metric because it is the key to measuring classification imbalance. The proposed model and models for comparison were implemented using PyTorch [56] as the programming framework. This framework is a deep learning open-source library that is written in Python and is based on the Torch library [56]. In this research, the experiments were performed on a system that provides an Intel(R) Core i7-CPU @ 2.60 GHz, 16 GB RAM, and NVIDIA GeForce GTX 1660 Ti. We use a batch-based ADAM algorithm. Finally, the best model is chosen among 150 epochs with $5e^{-4}$ optimised learning rate and adding regularisation by adding $1e^{-5}$ for weight decay to validate the performance of the proposed gait disorder classification network.

To evaluate the performance of the proposed framework, F-score, precision, and recall are used as the performing metrics. Since the performance of the neural network algorithm changes with initialisation, we set it with 21 different random seeds. Classification average precision, recall, F-score and accuracy are compared to different deep network structures in a tenfold cross-validation experiment in the training and validation stages as presented in Tables 5 and 6, respectively. As can be seen from the training results, all the methods show good performance during training, but in the validation results the highest scores are achieved with the DHAT-LSTM method, followed by TPCNN, where the ViT validation results show the lowest values among the other methods.

We also compared the performance of the proposed model and the published models with the test dataset, which is demonstrated in Fig. 13. Precision, recall, F-score, and accuracy are considered compared to the models presented. The proposed method has the highest accuracy and F-score values, 81% and 79%, respectively. The metrics performance of TPCNN and ViT give approximately a similar value, around 65% and 69% for accuracy, and 63% and 66% for F-score, respectively. The average precision and average recall values of the proposed approach are 82% and 77%, compared to TPCNN and ViT with 66%, 62%, 72% and 66%, respectively.

**Table 5** Comparison of training results (average ten times)

| Methods | Accuracy [%] | F-score [%] | Precision [%] | Recall [%] |
|---|---|---|---|---|
| DHAT-LSTM | 99.80 | 99.80 | 99.80 | 99.80 |
| TPCNN | 100.00 | 100.00 | 100.00 | 100.00 |
| ViT | 99.76 | 99.77 | 99.82 | 99.76 |

**Table 6** Comparison of validation results (average 10 times)

| Methods | Accuracy [%] | F-score [%] | Precision [%] | Recall [%] |
|---|---|---|---|---|
| DHAT-LSTM | 90.40 | 91.90 | 95.20 | 90.40 |
| TPCNN | 88.50 | 89.60 | 93.20 | 88.50 |
| ViT | 85.25 | 86.49 | 90.00 | 83.25 |

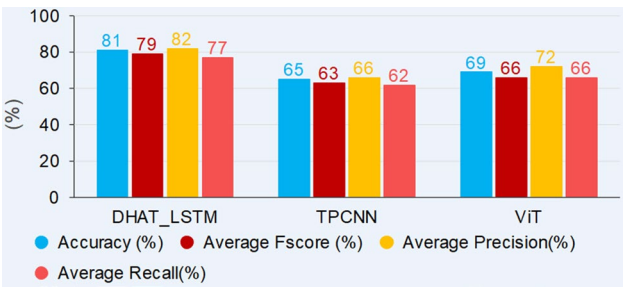


**Fig. 13** Performance comparison metrics

## 4 Discussion

Figure 14 shows the results of the DHAT model with and without the LSTM model. Considering that DHAT-LSTM introduces LSTM to enhance the analysis of sequential local information, its classification result is considerably better than that of DHAT. Therefore, the results of the experiment show the competence of the proposed model in learning long-term dependencies in classifying the analysis of gait disorders.

The proposed model showed an improvement in the classification rate of 73% to 81% compared to DHAT without LSTM. Other performance metrics also increased significantly, F-score from 71% to 79% and recall from 67% to 77%, where the precision is a nearly similar value. In addition, as we can see in Fig. 15, there is a large gap between the performance of the proposed model with two DHAT blocks and one DHAT block, whereby adding a second block the accuracy and F-score are increased around 12% greater than a model with one block.

It is obvious from the F-score of the classification in Fig. 16 that the DHAT-LSTM method shown the best results among other methods on the test dataset. The highest classification F-score is 86%, which is assigned to class 1, followed by 83% for class 3, while the lowest classification F-score is 67%, which is assigned to classes 2. The F-score of TPCNN and ViT methods show the similar value: 71% and 73% for class 1, whereas their results for class 2 and class 3 are 62%, 55%, 83% and 40%, respectively.

Table 7 compares the correct and incorrect (highlighted numbers) classification results of these models. The classes in the test dataset are not of the same size because the original data are also imbalanced, and therefore this test set is randomly selected with the proportion of their size. If the classes in the test set are of the same size, then class 1 will be more in the train dataset. This makes the train dataset worse because the gap between the size of the classes will be greater. The result of Table 7 shows that the proposed model has five incorrect classification results among 26 samples, while the TPCNN and ViT models have nine and eight fault classifications, respectively. As can be seen, in the four of five incorrect classifications of our proposed model, both other models or at least one of them also failed to classify correctly. For example, in samples number 9, 22 and 25, all methods have incorrect classification results; whereas in sample 15, the proposed model and TPCNN have incorrect classification results. Table 8 shows confusion matrix of mentioned models.

Figure 17 compares the model parameters, MACs, and FLOPs of the presented methods. For the proposed model, the total number of parameters is given as 242.259 K. The ViT model requires the least number of model parameters, where the minimum number of FLOPs and MACs is for TPCNN. The proposed framework has shown a larger number of parameter sizes, MACs and FLOPs, but significantly the highest accuracy among the other approaches. The accuracy of classification is more important in our application rather than a number of parameters and operations, as the model can be used for real applications. We reach significantly higher accuracy as it is more important in gait disorders classification applications. Our model has a higher number of operations, but it significantly has higher accuracy than other methods. As it can be seen from Fig. 13, the

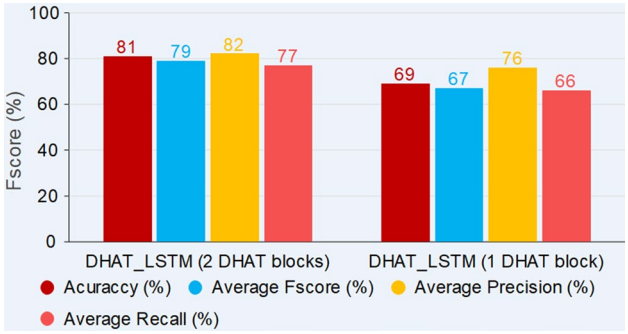


**Fig. 15** Performance comparison metrics of DHAT-LSTM with one and two blocks

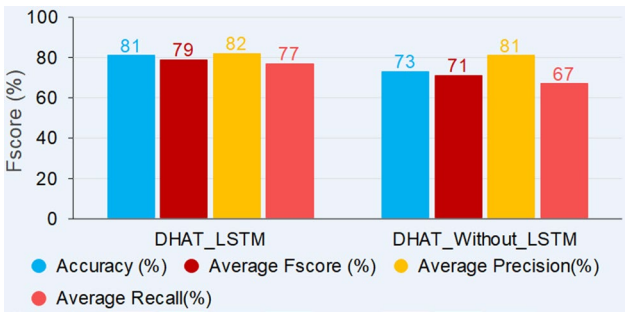


**Fig. 14** Performance comparison metrics of DHAT-LSTM and DHAT without LSTM

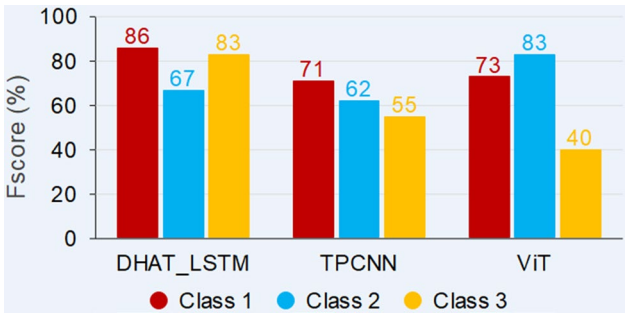


**Fig. 16** Average F-score comparison on the test dataset

**Table 7** Comparison of correct and incorrect (Bold) classification results

| Number | Ground-Truth | DHAT-LSTM | TPCNN | ViT |
|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 1 |
| 2 | 1 | 1 | **2** | **3** |
| 3 | 1 | 1 | 1 | 1 |
| 4 | 1 | 1 | 1 | 1 |
| 5 | 1 | 1 | 1 | 1 |
| 6 | 2 | **1** | 2 | 2 |
| 7 | 2 | 2 | **1** | 2 |
| 8 | 1 | 1 | 1 | 1 |
| 9 | 2 | **1** | **1** | **1** |
| 10 | 1 | 1 | 1 | 1 |
| 11 | 1 | 1 | 1 | 1 |
| 12 | 1 | 1 | 1 | 1 |
| 13 | 1 | 1 | 1 | **2** |
| 14 | 1 | 1 | 1 | 1 |
| 15 | 1 | **2** | **3** | 1 |
| 16 | 2 | 2 | 2 | 2 |
| 17 | 3 | 3 | **2** | 3 |
| 18 | 1 | 1 | **2** | 1 |
| 19 | 3 | 3 | 3 | **1** |
| 20 | 3 | 3 | **1** | **1** |
| 21 | 3 | 3 | 3 | 3 |
| 22 | 3 | **2** | **1** | **1** |
| 23 | 2 | 2 | 2 | 2 |
| 24 | 3 | 3 | 3 | **1** |
| 25 | 3 | **1** | **1** | **1** |
| 26 | 2 | 2 | 2 | 2 |

difference in accuracy and F-score is at least 12% higher than other methods. From these results, it can be seen that the TPCNN and ViT methods mostly show similar performance, but even our methodology is inspired by a standard transformer similar to the ViT method, although it shows how much the results have been improved by the proposed framework.

## 5 Conclusions

The proposed framework develops a standard transformer-based model for sequential classification that can perform spatio-temporal feature analysis on different types of variables. The DHAT-LSTM model is designed to classify the gait dysfunction of sequences based on created features from skeletal features key-points. To be more precise, the proposed framework modifies the self-attention mechanism to better capture the temporal dynamics of sequential input. Using the gait dysfunction classification as an experiment, our proposed model can produce state-of-the-art results.

Due to an imbalanced data sample and a lack of enough data samples for training, we applied techniques such as data augmentation, F-score metric evaluation, and optimum weights for each class in the training process. We have also created important features in the data-processing stage to improve the performance of the DHAT-LSTM classifier. Based on the DHAT-LSTM network, this method automatically learns spatio-temporal features to distinguish the difference between imbalance and balance movement with the created features.

We compared our algorithm with other deep networks in terms of accuracy and evaluation of the F-score. Our method has shown the best performance among other methods in evaluating gait disorders. The proposed method can help physicians make clinical decisions. This gait dysfunction classification method can be used as a classification system as an application to recommend that users visit a doctor if the result is classified as gait dysfunction.

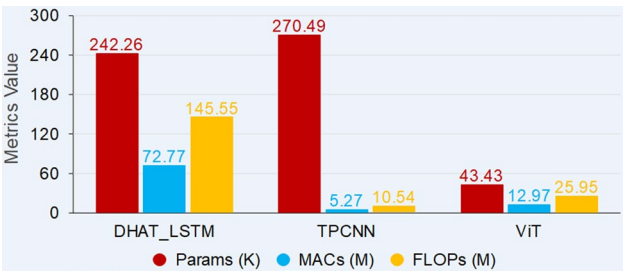


**Fig. 17** Metrics of model efficiency

**Table 8** Confusion matrix comparison

| | DHAT-LSTM | | | | TPCNN | | | | ViT | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | | 1 | 2 | 3 | | 1 | 2 | 3 |
| Clinician | **1** | 12 | 1 | 0 | **1** | 10 | 2 | 1 | **1** | 11 | 1 | 1 |
| | **2** | 2 | 4 | 0 | **2** | 2 | 4 | 0 | **2** | 1 | 5 | 0 |
| | **3** | 1 | 1 | 5 | **3** | 3 | 1 | 3 | **3** | 5 | 0 | 2 |

### 5.1 Future Work

The present study has some limitations. Although we have reached promising results with this small dataset, there are several gaps in which this study can be pursued. First, in future work, we intend to accumulate more datasets of real patients by working in partnership with rehabilitation centres and hospitals that treat patients with pathologies that affect the balance system. Second, we will add more effective features to the input feature vector, such as the angle of the lines of joint parts, such as the angle between arm and shoulder or the angle between elbow and arm. We are also interested in investigating the impact of each key-point or group of key-points on the performance of the classifier to propose the most effective key-points for feature gait disorder classification. Fourth, we will use the triple DHAT block in parallel for each exercise and will concatenate the result for classification and investigate the impact of multiple heads on the performance of classification. Finally, we will design a new methodology to analyse 3D data and compare its results with 2D performance.

**Acknowledgements** The work of J.K., and J.M. was supported by the Ministry of Education, Youth and Sports through a grant 'Development of Advanced Computational Algorithms for evaluating post-surgery rehabilitation' number LTAIN19007. The work of M.Ch. and K.T. was supported by the research project of Charles University Cooperatio 43 - Surgical Disciplines, 3rd Faculty of Medicine. This support is gratefully acknowledged.

**Author Contributions** J.K. and J.M.: research, signal processing, writing; M.S.: research, neural network design, writing; K.T. and M.Ch.: biomedical background, writing.

**Funding** The work was funded by a grant 'Development of Advanced Computational Algorithms for evaluating post-surgery rehabilitation' number LTAIN19007.

**Availability of Data and Materials** Data and materials are available on request.

## Declarations

**Conflict of Interest** The authors have no competing interests.

**Consent for Publication** All the authors read and consent the publication.

**Ethical Approval and Consent to Participate** This research was approved by the Ethics Committee of the University Hospital Královské Vinohrady Prague (EK-VP/4310120), where the measurements were made. Each patient signed an informed consent to the research conditions.



## References

1. Matsubara, Y., Sakurai, Y., Van Panhuis, W.G., Faloutsos, C.: Funnel: automatic mining of spatially coevolving epidemics. In: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 105–114 (2014)
2. Le Guen, V., Thome, N.: Shape and time distortion loss for training deep time series forecasting models. Adv. Neural. Inf. Process. Syst. **3**, 2 (2019)
3. Choi, E., Bahadori, M.T., Sun, J., Kulas, J., Schuetz, A., Stewart, W.: Retain: an interpretable predictive model for healthcare using reverse time attention mechanism. In: Advances in neural information processing systems. Springer, Cham (2016)
4. Ye, L., Keogh, E.: Time series shapelets: a new primitive for data mining. In: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 947–956 (2009)
5. Zhou, F., Gao, Y., Wen, C.: A novel multimode fault classification method based on deep learning. J. Control Sci. Eng. (2017)
6. Forney, E.M., Anderson, C.W., Gavin, W.J., Davies, P.L., Roll, M.C., Taylor, B.K.: Echo state networks for modeling and classification of eeg signals in mental-task brain-computer interfaces. Color. State Univ (2015)
7. Zhao, B., Lu, H., Chen, S., Liu, J., Wu, D.: Convolutional neural networks for time series classification. J. Syst. Eng. Electron. **28**(1), 162–169 (2017)
8. Sun, Z., Di, L., Fang, H.: Using long short-term memory recurrent neural network in land cover classification on landsat and cropland data layer time series. Int. J. Remote Sens. **40**(2), 593–614 (2019)
9. Kashiparekh, K., Narwariya, J., Malhotra, P., Vig, L., Shroff, G.: Convtimenet: A pre-trained deep convolutional neural network for time series classification. In: 2019 International Joint Conference on Neural Networks (IJCNN), pp. 1–8. IEEE (2019)
10. Jin, X., Yu, X., Wang, X., Bai, Y., Su, T., Kong, J.: Prediction for time series with cnn and lstm. In: Proceedings of the 11th international conference on modelling, identification and control (ICMIC2019), pp. 631–641. Springer, Cham (2020)
11. Tokgöz, A., Ünal, G.: A rnn based time series approach for forecasting turkish electricity load. In: 2018 26th Signal Processing and Communications Applications Confer-ence (SIU), IEEE, pp. 1–4 (2018)
12. Hochreiter, S.: The vanishing gradient problem during learning recurrent neural nets and problem solutions. Internat. J. Uncertain. Fuzziness Knowl. Based Syst. **6**(02), 107–116 (1998)
13. Noh, S.-H.: Analysis of gradient vanishing of rnns and performance comparison. Information **12**(11), 442 (2021)
14. Khedhiri, S., et al.: Comparison of sarfima and lstm methods to model and to forecast Canadian temperature. Reg. Stat. **12**(02), 177–194 (2022)
15. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł, Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems, vol. 30. Springer, Cham (2017)

16. Tay, Y., Bahri, D., Metzler, D., Juan, D.-C., Zhao, Z., Zheng, C.: Synthesizer: Rethinking self-attention for transformer models. In: International Conference on Machine Learning, PMLR, pp. 10183–10192 (2021)
17. Gupta, D., Sundaram, S., Khanna, A., Hassanien, A.E., De Albuquerque, V.H.C.: Improved diagnosis of parkinson's disease using optimized crow search algorithm. Comput. Electr. Eng. **68**, 412–424 (2018)
18. Wang, Y., Wang, A.-N., Ai, Q., Sun, H.-J.: An adaptive kernel-based weighted extreme learning machine approach for effective detection of Parkinson's disease. Biomed. Signal Process. Control **38**, 400–410 (2017)
19. Verlekar, T.T., Soares, L.D., Correia, P.L.: Automatic classification of gait impairments using a markerless 2d video-based system. Sensors **18**(9), 2743 (2018)
20. Zhou, C., Mitsugami, I., Yagi, Y.: Detection of gait impairment in the elderly using patch-gei. IEEJ Trans. Electr. Electron. Eng. **10**, 69–76 (2015)
21. Zihang, D., Zhilin, Y., Yiming, Y., Jaime G., C., Quoc V., L., Salakhutdinov, R.: Transformer-xl: attentive language models beyond a fixed-length context. CoRR abs/1901.02860 (2019). arXiv:1901.02860
22. Sakatani, Y.: Combining rnn with transformer for modeling multi-leg trips. In: WebTour@ WSDM, pp. 50–52 (2021)
23. Veres, G.V., Gordon, L., Carter, J.N., Nixon, M.S.: What image information is important in silhouette-based gait recognition? In: Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004., vol. 2. IEEE (2004)
24. Birch, I., Birch, M., Asgeirsdottir, N.: The identification of individuals by observational gait analysis using closed circuit television footage: comparing the ability and confidence of experienced and non-experienced analysts. Sci. Justice **60**(1), 79–85 (2020). https://doi.org/10.1016/j.scijus.2019.10.002
25. Collins, S.H., Adamczyk, P.G., Kuo, A.D.: Dynamic arm swinging in human walking. Proc. R. Soc. B **276**(1673), 3679–3688 (2009)
26. Herr, H., Popovic, M.: Angular momentum in human walking. J. Exp. Biol. **211**(4), 467–481 (2008)
27. McIntosh, A.S., Beatty, K.T., Dwan, L.N., Vickers, D.R.: Gait dynamics on an inclined walkway. J. Biomech. **39**(13), 2491–2502 (2006)
28. Jokisch, D., Daum, I., Troje, N.F.: Self recognition versus recognition of others by biological motion: viewpoint-dependent effects. Perception **35**(7), 911–920 (2006)
29. Larsen, P.K., Simonsen, E.B., Lynnerup, N.: Gait analysis in forensic medicine. J. Forensic Sci. **53**(5), 1149–1153 (2008)
30. Čakrt, O., Chovanec, M., Funda, T., Kalitová, P., Betka, J., Zvěřina, E., Kolář, P., Jeřábek, J.: Exercise with visual feedback improves postural stability after vestibular schwannoma surgery. Eur. Arch. Oto-rhino-laryngol. **267**(9), 1355–1360 (2010)
31. Govorun, M., Usachev, V., Kuznetsov, M., Golovanov, A.: The application of computed stabilometry for the diagnostics of vestibular disorders following stapedoplasty and for the estimation of the functional status in man. Vestn. Otorinolaringol. **4**, 57–58 (2012)
32. Shimizu, K., Imai, T., Oya, R., Okumura, T., Sato, T., Osaki, Y., Ohta, Y., Inohara, H.: Platform posturography of patients with peripheral vestibular dysfunction in the non-acute phase of vertigo. Auris Nasus Larynx **48**(4), 577–582 (2021)
33. Mégnigbéto, C., Sauvage, J.-P., Launois, R.: Validation clinique d'une échelle du vertige: Eev (european evaluation of vertigo). Revue de laryngologie, d'otologie et de rhinologie (1919) **122**(2), 95–102 (2001)
34. Steffeni, T., Hacker, T., Mollinger, L.: Age-anid gender-related test performance in community-dwelling elderly people: Six-minute walk test, berg balance scale, timed up & go test, and gait speeds. Phy's TI/er **82**, 128–37 (2002)
35. Zur, O., Carmeli, E.: The university of california los angeles dizziness questionnaire: Advantages and disadvantages. J. Vestib. Res. **23**(6), 279–283 (2013)
36. Graham, M.K., Staab, J.P., Lohse, C.M., McCaslin, D.L.: A comparison of dizziness handicap inventory scores by categories of vestibular diagnoses. Otol. Neurotol. **42**(1), 129–136 (2021)
37. Ullah, A., Ahmad, J., Muhammad, K., Sajjad, M., Baik, S.W.: Action recognition in video sequences using deep bi-directional lstm with cnn features. IEEE Access **6**, 1155–1166 (2017)
38. Staudemeyer, R.C., Morris, E.R.: Understanding lstm–a tutorial into long short-term memory recurrent neural networks. arXiv preprint arXiv:1909.09586 (2019)
39. Cui, Z., Ke, R., Pu, Z., Wang, Y.: Stacked bidirectional and unidirectional lstm recurrent neural network for forecasting network-wide traffic state with missing values. Transp. Res. Part C **118**, 102674 (2020)
40. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2961–2969 (2017)
41. Kohout, J., Crha, J., Trnkova, K., Sticha, K., Mares, J., Chovanec, M.: Robot-based image analysis for evaluating rehabilitation after brain surgery. Mendel **24**, 159–164 (2018)
42. Johnson, J.W.: Adapting mask-rcnn for automatic nucleus segmentation. arXiv preprint arXiv:1805.00500 (2018)
43. Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.-R.: Microsoft coco: common objects in context. In: European conference on computer vision, pp. 740–755. Springer, Cham (2014)
44. Shayestegan, M., Kohout, J., Štícha, K., Mareš, J.: Advanced analysis of 3d kinect data: supervised classification of facial nerve function via parallel convolutional neural networks. Appl. Sci. **12**(12), 5902 (2022)
45. Cao, Z., Simon, T., Wei, S.-E., Sheikh, Y.: Realtime multi-person 2d pose estimation using part affinity fields. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7291–7299 (2017)
46. Li, Q., Li, R., Ji, K., Dai, W.: Kalman filter and its application. In: 2015 8th International Conference on Intelligent Networks and In-telligent Systems (ICINIS), pp. 74–77. IEEE (2015)
47. Dalaison, M., Jolivet, R.: A kalman filter time series analysis method for insar. J. Geophys. Res. **125**(7), 2019–019150 (2020)
48. Dentamaro, V., Impedovo, D., Pirlo, G.: Gait analysis for early neurodegenerative diseases classification through the kinematic theory of rapid human movements. IEEE Access **8**, 193966–193980 (2020)
49. Song, H., Dai, J., Luo, L., Sheng, G., Jiang, X.: Power transformer operating state prediction method based on an lstm network. Energies **11**(4), 914 (2018)
50. Xiong, R., Yang, Y., He, D., Zheng, K., Zheng, S., Xing, C., Zhang, H., Lan, Y., Wang, L., Liu, T.: On layer normalization in the transformer architecture. In: International Conference on Machine Learning, PMLR, pp. 10524–10533 (2020)
51. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
52. Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer normalization. arXiv preprint arXiv:1607.06450 (2016)
53. Zhang, J., Bai, F., Zhao, J., Song, Z.: Multi-views action recognition on 3d resnet-lstm framework. In: 2021 IEEE 2nd International Conference on Big Data, Artificial In-telligence and Internet of Things Engineering (ICBAIE), pp. 289–293. IEEE (2021)

54. Rai, N., Kumar, D., Kaushik, N., Raj, C., Ali, A.: Fake news classification using transformer based enhanced lstm and bert. Int. J. Cognit. Comput. Eng. **3**, 98–105 (2022)
55. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M.-f., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
56. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gim-elshein, N., Antiga, L., et al.: Pytorch: an imperative style, high-performance deep learning library. In: Advances in neural information processing systems, vol. 32. Springer, Cham (2019)