IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 5 October 2023, accepted 6 November 2023, date of publication 28 November 2023, date of current version 1 December 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3336861

RESEARCH ARTICLE

Encoding Kinematic and Temporal Gait Data in an Appearance-Based Feature for the Automatic Classification of Autism Spectrum Disorder

B. HENDERSON, PRATHEEPAN YOGARAJAH, BRYAN GARDINER, (Member, IEEE), AND T. MARTIN MCGINNITY, (Senior Member, IEEE)
School of Computing, Engineering and Intelligent Systems, Ulster University, BT48 7JL Londonderry, U.K.

Corresponding author: B. Henderson (henderson-b7@ulster.ac.uk)

This work was supported by the Intelligent Systems Research Centre (ISRC), Ulster University, U.K.

ABSTRACT In appearance-based gait analysis studies, Gait Energy Images (GEI) have been shown to be an effective tool for human identification and gait pathology detection. In addition, model-based studies found kinematic and spatio-temporal features to be useful for gait recognition and Autism Spectrum Disorder (ASD) classification. Adapting the GEI to focus on the strong ASD features would improve the early screening of ASD by allowing the use of powerful appearance-based classifiers such as Convolutional Neural Networks (CNN). This paper introduces an enhanced GEI, by averaging images from a video sequence to produce a single image but by retention of a person's joint positions only, instead of the full body silhouettes. Depth is encoded into the binary images before they are averaged using colour mapping, a technique used in the Chrono-Gait Image. The Joint Energy Image (JEI) therefore embeds both the temporal and depth information of the joints into a 2D image. The image was preprocessed using Principal Component Analysis before being applied to a Multi-Layer Perceptron, and a Random Forest classifier. The JEI was also applied to a CNN directly and accuracy was improved when using a Test Time Augmentation (TTA) measure. The CNN achieved a TTA accuracy of 95.56% when trained on a primary dataset of 100 subjects (50 with ASD and 50 that are typically developed), and 80% TTA accuracy on a secondary dataset of 20 subjects (10 ASD and 10 typically developed) across multiple tests.

INDEX TERMS Autism spectrum disorder, gait analysis, neural networks, video analysis.

I. INTRODUCTION

Autism Spectrum Disorder (ASD) is a neurological developmental disorder with an estimated global prevalence of between 3 and 6 children per 1000 [1]. This level of prevalence prevails irrespective of culture, geography, and degree of industrialisation [2]. Developing countries report lower ASD prevalence and potential causes are still being discussed. These include a genuine low prevalence, deficits in diagnostic skills, mal-adaptation of diagnostic criteria in relation to culturally different behaviour, and under sampling. An increasing trend has also been noted with regards to the global prevalence of ASD, establishing that it is an ever-present and important disorder as it affects many people across the world.

The associate editor coordinating the review of this manuscript and approving it for publication was Byung-Gyu Kim.

Early diagnosis of a child with ASD gives these children and their families access to resources that can help improve their quality of life. Early Intervention (EI) programs were found to have moderate to large effects on child outcomes such as social, language, and nonverbal cognitive abilities [3]. An average age of 3.81 years was reported as optimal for improvements in social communication using EI. After this age the positive impacts diminish [4]. An early diagnosis also helps parents of children with ASD resist stigmatisation [5]. This highlights the importance of early and accurate diagnosis of ASD.

A. SYMPTOMS AND GAIT FEATURES

Research on ASD to date has mostly focused on the behavioural and social aspects of the condition. Clinical research sites across Europe use the Autism Diag-

TABLE 1. List of abbreviations in the order that they appear in the main body of this paper.

Abbreviation	Full Form
ASD	Autism Spectrum Disorder
EI	Early Intervention
ADOS	Autism Diagnostic Observation Schedule
ADI	Autism Diagnostic Interview
SRS	Social Responsiveness Scale
TD	Typically Developed
fMRI	Functional Magnetic Resonance Imaging
PCA	Principal Component Analysis
LDA	Linear Discriminant Analysis
SVM	Support Vector Machine
NBC	Naïve Bayes Classifier
ANN	Artificial Neural Network
QDA	Quadratic Discriminant Analysis
KNN	K-Nearest Neighbour
RF	Random Forest
DT	Decision Tree
CNN	Convolutional Neural Network
LSTM	Long-Short Term Memory
GEI	Gait Energy Image
CGI	Chrono-Gait Image
JEI	Joint Energy Image
TTA	Test Time Augmentation
MLP	Multi-Layer Perceptron
NASOM	National Autism Society of Malaysia
PMM	Predictive Mean Matching
DR	Dimensionality Reduction
ReLu	Rectified Linear Unit
TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative
CV	Cross Validation
ANOVA	Analysis of Variance

nostic Observation Schedule (ADOS) and the Autism Diagnostic Interview (ADI) [6]. Approximately half of the sites considered also used the Social Communication Questionnaire (SCQ) and the Social Responsiveness Scale (SRS). As these tools are based around human observation of behaviour and reflective questions, they are open to subjective responses by participants. Recent research has begun to investigate gross motor impairment as a potential set of symptoms for identifying ASD [7]. Kinematic [8], [9], kinetic [10], [11] and spatiotemporal [12] gait differences have all been found between children with ASD and children that are typically developed (TD). The advantage of using such characteristics for detecting autism is that they result from objective measurements of gait.

Kinetic features are measured almost exclusively from force plates that require physical touch from the participants [10], [13], [14]. The requirement for physical touch is a limiting factor because it is a more obtrusive method of data collection due to the need for a specific environment setup. Kinematic features like joint angles [15] and spatiotemporal features like cadence and velocity [16] have also been collected in ASD studies using similarly obtrusive methods such as marker-based tracking systems. Alternatively, non-ASD studies have confirmed the efficiency of using markerless-based systems like the Kinect V2 for collecting joint kinematics [17]. Spatio-temporal gait data e.g. stride-timings were also collected using an older version of the Kinect [18]. The Kinect has additionally been used successfully to collect whole-body movement kinematic data during video game play for ASD classification [19]. These studies highlight the viability of unobtrusive devices such as the Kinect to collect kinematic and temporal gait features for ASD classification.

B. CLASSIFICATION AND MACHINE LEARNING

A recent review of automated detection approaches for ASD using Human Activity Analysis [20] categorises the literature into three main areas; gaze analysis, repeated behaviour, and abnormal gait detection. The review also refers to Functional Magnetic Resonance Imaging (fMRI) studies where images of the brain obtained using MRI are used to detect ASD. Another review [21] summarises current ASD classification using fMRI literature. It finds the best outcome to have 83.00% accuracy so far [22] and given that as a field of research it is much younger, there is likely scope for further development. Despite the potential for early screening that this offers, it suffers from the same disadvantages as methods that require Motion Capture cameras. MRI machines aren't widely available in developing countries due to their cost and the expertise required for setup and usage, and so are limited to hospitals or laboratories. In addition, they can be uncomfortable for the participant being examined. Earlier reviews focused more directly on studies aimed at the classification of ASD using gait analysis [23]. Both of the reviews that included gait analysis include a subset of papers that made use of kinetic, kinematic and spatiotemporal gait features with machine learning models as a method of automatic classification. A study that was related to the second gait review paper [24], applied Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) as feature selection to a kinematic feature set before using a Support Vector Machine (SVM), Naïve Bayes Classifier (NBC) and Artificial Neural Network (ANN) as classifiers. All models performed similarly with the NBC achieving the highest accuracy. Kinetic and kinematic features were also used in combination with LDA and Quadratic Discriminant Analysis (QDA) as classifiers [13]. Here, the LDA outperformed the QDA when using kinetic features where the highest accuracy was achieved. However, the QDA performed better when the kinematic dataset was used. Another study compared SVM, K-Nearest Neighbours (KNN), Random Forest (RF) and Decision Tree (DT) classifiers on a small spatiotemporal dataset [25] with the RF classifier performing best when intra-subject variance was reduced. This supports the feasibility of machine learning in combination with gait analysis as a useful tool in the process of determining an ASD diagnosis.

C. APPEARANCE-BASED FEATURES

Recent studies have also considered video data in combination with neural networks to perform ASD classification. One

such study applied various Convolutional Neural Network (CNN) models to videos of people with ASD performing different actions such as object placement [26]. A set of Long-Short Term Memory (LSTM) models were then utilised for classification. Another study used videos from ADOS interviews to extract spatiotemporal facial features for classification [27]. In non-ASD literature, an appearance-based feature called the Gait Energy Image (GEI) [28], that encapsulates spatiotemporal characteristics of gait in a single 2D image, has been commonly paired with the CNN [29] in general gait-based classification problems. Although useful as a classifier itself, in one study the CNN has been used as an auto-encoder on GEIs before being applied to both a One Class – SVM and an Isolation Forest classifier for gait disorder detection [30]. A similar study took partial GEIs, affected by occlusion, and used them as input into a CNN auto-encoder before using a KNN classifier for human identification [31]. Again, in another study, the CNN was applied to GEIs to extract features for an LDA classifier [32]. In this study, transfer learning was applied to a pre-trained CNN called VGG-19 before being retrained for gait pathology feature extraction. Even without a CNN, the GEI is useful for calculating gait based features for gait classification problems using other machine learning classifiers such as the SVM [33]. These studies together support both the power of the GEI for gait-related classification problems and the ability of the CNN to interpret appearance-based features for classification tasks.

Although the CNN has yet to be applied to gait analysis ASD classification, it shows promise when applied to gait-based image features. Similarly, the GEI has yet to be applied to ASD classification. In the current study, the GEI is adapted so that it also embeds depth and time, to enable the information contained in ASD-specific kinematic and spatiotemporal gait features previously discussed to be highlighted in an appearance-based feature. This new feature is therefore specifically aimed at differentiating ASD gait and is used alongside a promising new classifier for ASD classification. One of the adaptations to the GEI in the literature is the Volume Energy Image [34] which averages 3D voxel volumes instead of 2D images. Incorporating depth into the GEI improves it by virtue of working in 3-Dimensional space, for example making it possible to differentiate between left and right limbs. Another adaptation of the GEI is the Chrono-Gait Image (CGI) [35] which reduces the silhouettes that are averaged into a GEI to only their contours. A colour map is applied to the images based on the video frame (time) in which the individual image occurs, relative to other images in a video sequence. This therefore encapsulates time-based information into a single image. The two key aspects of these adaptations; the colour mapping from the CGI to embed time and depth information, and the averaging of gait images into a single gait image from the GEI, are utilised in the development of a new appearance-based gait feature presented in the current study.

D. CONTRIBUTIONS

The current study aims to address the gaps in the literature discussed in the previous section by the following contributions:

1) Development of a new appearance-based gait feature computed from 3D joint positions that incorporates temporal and depth-based gait data called a **Joint Energy Image** (JEI).
2) Comparative tests of the JEI as a feature for ASD classification using a classic machine learning model and two neural network classifiers.
3) Application of Test Time Augmentation (TTA) Accuracy to assess the viability of both the classifiers and the JEI to compete with state-of-the-art feature sets and performance.

The remainder of the paper is organised as follows: In Section II the dataset that is used to train and test the machine learning models is described. In Section III the methodology that pertains to the creation and processing of the JEI from 3D joint positions is presented in detail. This includes the methodology for creating variations of the JEI using different combinations of joints, planes of rotation, and gait cycle segmentation techniques. In Section IV the techniques used for training and testing the machine learning models are given, including a description of the different accuracy measures and how they are calculated. In Section V the accuracy results from multiple classification tests are presented and compared to results from other state-of-the-art features and models. Finally, Section VI summarises the completed work with conclusions and suggests future research directions.

II. DATASETS

Two datasets (referred to as the primary and secondary datasets) were used to train the machine learning models for testing. The primary dataset is an existing 3D gait and full body movement dataset of children with ASD. A previous paper introduced the dataset and then extracted kinematic and spatiotemporal features from the 3D joint positions and trajectories [36]. The features extracted were the distances between two of the joints, the distances from some joints to the ground, the range of motion for each joint, hand tip position, step length/width, distance between the feet, stride length, gait cycle time, stand time and swing time. The mean, variance, and standard deviation of each of these were then calculated, resulting in 1259 values before dimensionality reduction. The same features were calculated again after the original gait data was augmented to represent different recording conditions. The augmentations included; applying jittering to simulate additive sensor noise, scaling the joint positions up and down, translating the skeleton left and right, flipping the horizontal axis, and slicing to extract different continuous slices from the original time series. PCA was then applied to reduce the dimensionality of the dataset to 31 features, the top 11 of which were chosen for classification based on their standard deviation results. The data was then

shuffled by subject and split with a 7:3 train to test ratio. The features that were calculated using the augmented data were then removed from the test set. Finally, a basic Multi-Layer Perceptron (MLP) model was trained for ASD classification, with 11 input nodes, 6 nodes in the hidden layer and 2 output nodes, achieving an accuracy of 95%.

The secondary dataset is also a 3D gait dataset for children with ASD. The dataset is a subset of the data used in [37], kindly provided by the authors. In the cited study, the kinematic 3D marker position trajectories were recorded using a Vicon Motion Capture System (Vicon MX T-Series). From this raw data, a series of kinematic features were extracted, including sagittal joint angles during foot strike and foot off events. Between-group tests and step-wise discriminant analysis were used for dimensionality reduction, resulting in 9 and 4 final features for training respectively. A 3-layer artificial neural network was then trained using 10-fold cross validation, achieving a best accuracy of 91.7%.

This work exploits the data from both datasets for ASD classification as follows. Whereas the original papers made use of kinematic and kinetic features calculated directly from the 3D joint positions, the current study aims to use the same 3D joint positions as a starting point to produce an appearance-based feature in the form of a 2D image. This feature will encode both the depth and temporal information of the joints (and therefore the kinematic and temporal information) into the image using colour-mapping and averaging techniques. A CNN classifier will be used due to its performance on similar features such as the Gait Energy Image. An MLP and RF classifier will be used for comparisons with the original papers as well as to include a classic machine learning model found to be useful for ASD classification on gait data [25].

The relevant information on the datasets as outlined in their original papers is summarised below.

A. PARTICIPANTS

Individuals involved in the primary dataset initially included 68 children with ASD and 50 TD children recruited to perform straight walking gait trials in a controlled environment. In cases where the degree of ASD was severe, where there was a lack of response or great dispersion, the data were excluded or their movements were simulated. This resulted in the total number of ASD participants' data equalling 50. Children with ASD were located from 7 ASD childcare centres in 3 different cities and the TD children were located from 2 kindergarten centres in Iraq. All participants were free of any lower extremity injury, neurological disorders or diseases that would affect gait, except for ASD.

Individuals involved in the secondary dataset included 30 ASD and 30 TD children. The author's shared 10 children's data from each group for use in this study. All children in this dataset were aged 4 to 14. Those in the ASD group were diagnosed with a mild category of ASD and were recruited from the National Autism Society of Malaysia (NASOM) centre. Children in the TD group were recruited from the local communities or were family of members of faculty.

B. ENVIRONMENTAL SETUP AND PROCEDURE

The trials for the primary dataset were recorded using a Kinect V2 camera, positioned so that the children walked towards it during the trials. The Kinect was placed at a height of 0.75m from the ground and measures were taken to control the lighting (measured using a Light Meter app on a Samsung Galaxy Note 9) and temperature (measured using a mercury thermometer). Walking trials were repeated 10 times per subject with each trial containing approximately 2 gait cycles before a single valid trial was chosen.

The trials for the secondary dataset were recorded using an eight camera motion capture system called the Vicon (MX T-Series). A total of 35 retro-reflective markers were tracked after being attached to the children based on the full body Plug-In Gait model. The 3D marker trajectories were recorded at 100 Hz during straight barefoot walks over a walkway. The built in Woltring generalized cross-validatory spline algorithm was implemented to minimize noise. An average of 10 trials were recorded per subject, with a single valid trial being chosen for analysis.

C. DATA RECORDS

The full primary dataset contains 3D joint positions, a skeleton movement video, joint trajectory videos, and colour videos captured using a Samsung Note 9 rear camera. For this study, only the 3D joint positions were required. The Kinect V2 recorded joint positions in meters as co-ordinates (x, y, z). Values ranged from -6 (right) to 6 (left) for the x co-ordinate, from -5 (bottom) to 5 (top) for the y co-ordinate and from 0 (camera location) to 8 (maximum depth). At each time frame, the 3D positions of 25 joints were recorded.

The secondary dataset only contains the 3D marker positions, recorded in millimeters as co-ordinates (x, y, z) using the Vicon. Different points on the body were tracked in comparison to the primary dataset. While the primary dataset calculated the positions of joints in the body, the secondary dataset tracked markers placed following the Plug-In Gait model. The 3D positions of 35 markers were recorded at a speed of 100 Hz.

D. DATA PRE-PROCESSING

Distance between the feet was used as a signal from which a single gait cycle could be extracted from the full time series. Missing data in the primary dataset was replaced using Predictive Mean Matching (PMM) to determine the replacement value. Some joints in the secondary dataset could not be replaced in the same manner. Therefore, if any of these missing markers included key positions for normalising the skeleton (as discussed in the methodology section), then that subject's data was excluded. Key positions or joints were those that were used in the directional alignment calculations such as the spinal joints. For joints that were missing and were not in key positions for the alignment calculations, only

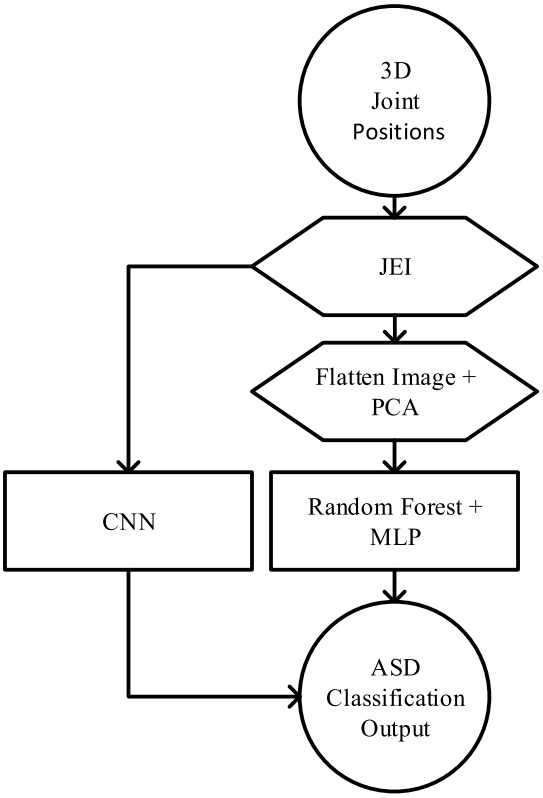

FIGURE 1. Complete JEI pipeline from input data to classification.

the missing joint was excluded. The data for each individual was processed by applying 7 transformations including translations of the skeleton to the left and right, applying Gaussian noise, flipping the skeleton horizontally, scaling the skeleton up and down and finally, slicing a different gait cycle from the original time series. With the augmented data, each individual was represented in the dataset, as used in this study, by 8 total gait cycles: one from the original recording as well as the additional 7 augmented versions. In total, the primary and secondary datasets produced 800 and 104 trials of 3D joint positions respectively, to be used for training and testing classifiers.

III. METHODOLOGY

The full pipeline for the proposed experiments is outlined in Fig. 1. Starting with the input data the 4 key stages are: creating the JEI feature; pre-processing; training machine learning models; and measuring its performance in the ASD classification problem. Each of these components contain several subprocesses that will be detailed in the following sections of the paper. First, the feature extraction stage describes the process of creating a JEI from a sequence of 3D joint positions from a complete gait cycle. It also includes identification and removal of duplicate images. Following this, the techniques used for pre-processing the images, namely flattening and PCA are detailed. This pre-processing stage is skipped when using the CNN classifier as it takes

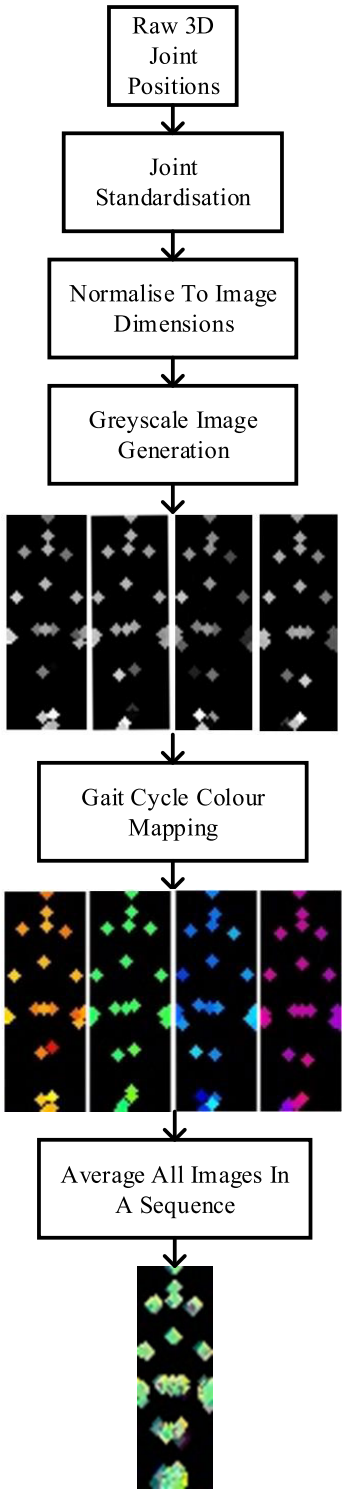

FIGURE 2. The process of creating a JEI from a sequence of 3D joint positions.

the complete image as input directly. The techniques used for determining an optimal model, as well as descriptions of the models themselves, are presented next in the training stage. Finally, the accuracy measures used to compare model performances are presented in the testing stage.

A. FEATURE EXTRACTION: JEI CREATION PROCESS

The 3D joint positions were recorded over a single gait cycle per person resulting in a set of 25 joint positions for every frame that was captured in the primary dataset. The number of frames for each subject's gait cycle ranged from 21 to 60, recorded at a rate of 30 frames per second. In the secondary dataset, the number of joint positions varied depending on missing joints with a maximum of 35. The number of frames for each subject's gait cycle in this dataset ranged from 67 to 136, recorded at a rate of 100 frames per second. The proposed JEI feature aims to convert a sequence of joint positions, which make up a skeleton, into a single 2D image that embeds the kinematic and spatiotemporal aspects of a person's gait cycle. The proposed process for creating the JEI from a sequence of 3D joint positions is outline in Fig 2, and can be broken down into 5 main steps:

1) Standardise the 3D joint positions by rotating the ''skeleton'' so that it is aligned with the frontal plane.
2) Normalise the value range of joint positions to fit predefined image dimensions and intensity values.
3) Generate a grayscale image, where joint positions are image coordinates, such that a simple shape can be drawn to represent the joint, and pixel intensity represents the depth values.
4) Apply a colour map to the grayscale images.
5) Average all colour mapped images from the same sequence into a single image.

The first step involves direction alignment of the 3D joint positions so that when a 2D image is generated from them, the overall skeleton shape is facing the ''camera'' and the image can be considered to have been captured from the frontal view. In the primary dataset being used, the joints positions were captured from the frontal view. It was noted that there was some slight variation between individuals in their direction of motion despite the common direction of movement. A similar pattern of variance in direction was noted in the secondary dataset, which first had to be rotated by 90 degrees to be in the same frontal view. The directional alignment of the joints aimed to remove this variation in direction to improve classification results.

The second step visually represents the depth of each joint in a grayscale image by changing the pixel intensity for each joint. This was achieved by taking the aligned joint positions and normalising them to the dimensions of the image so that their x and y positions, previously in either meters or millimeters depending on the dataset, could now be considered image coordinates. The z coordinate value is normalised to the full grayscale value range of an image so that the depth of each joint is maintained in the 2D image. This normalisation also removes variation due to distance from the recording device as all individuals will take up the same image space. Another key outcome of the normalisation is that the same image dimensions can be used consistently to represent all subjects' joints, regardless of their height and width. This is important when using images as input to machine learning models.

With the positions now in image coordinates, a grayscale image is generated. For each joint in a single frame, a shape is drawn on the image at its coordinates, using the normalised z coordinate for its pixel intensity. The result is a single black image with 25 white shapes of varying intensity resembling the human skeleton. The fourth step is to then segment the different frames from a sequence of images to represent different times in the gait cycle. A colour map is applied to every image contained in a segment, with different time segments using different colour maps. Each colour map represents a different time segment, and the range of colours within each colour map will represent the depth of the joint. Therefore, both time and depth information will have been embedded into the image. The final step averages all images from the same sequence, similar to the process applied in creating GEIs, to produce a single averaged image.

B. JOINT STANDARDISATION

The 3D positions of each joint in its raw format, are in a co-ordinate space where the device that captured them is the origin. To make the input more robust to device location, the joint positions are normalised to the coordinates of one of the joints, now making it the origin. The joint at the middle of the spine was chosen for this purpose. This removes any dependency of the joint positions from the device itself.

Manual inspection of the RGB videos in the dataset found that there was some deviation from the straight line during walking. Assuming the child starts in the same relative location to the camera, as they walk, their direction of motion may also deviate. Variation introduced by this deviation is adjusted for by considering all the joint positions as a single skeleton and rotating it back to the straight line.

The total rotation can be reduced to 3 smaller rotations, one each along the frontal, sagittal and transverse anatomical planes. The aim of these rotations is to align each of the named anatomical planes with the planes associated with a virtual camera such that the resultant image shows the person's skeleton as being upright and facing directly towards the camera.

Determining the angle of rotation along each plane was accomplished by first choosing a ''bone,'' defined as a line of best fit calculated when a subset of joints are considered as points on a graph. Combinations of three subsets of joints were considered for each test. The three sets were spinal joints, shoulder joints, and hip joints, the combinations of which are noted in Table. 2. A pivot joint is then chosen, around which the bone is to be rotated. Then the angle about this joint that the bone needs rotated by is calculated in order to align the bone line up with the required axis. The mid-spine and the spine shoulder joint were chosen as the pivot joints to be rotated around for the frontal and sagittal, and transverse rotations respectively. These were chosen because of their central location in the body along their respective axis.

For the frontal and sagittal rotation, the line of best fit, L_{BF} was aligned vertically whereas in the transverse rotation, it was aligned horizontally. To keep the calculation simple,

TABLE 2. Combinations of joints, rotations and gait cycle segmentation techniques used in each test.

Test	Joints	Rotations	Gait Cycle Segmentation
1	Spinal, Shoulder	All	Thirds
2	Spinal, Shoulder	All	Stance
3	Spinal, Shoulder, Hip	Sagittal Only	Quarters

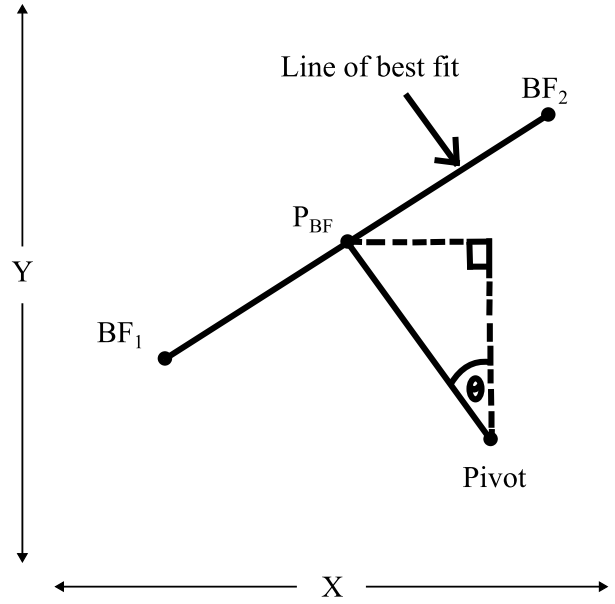

FIGURE 3. The points and angles used for the rotation calculations.

a new line, L_P that is perpendicular to L_{BF}, and passes through the pivot joint is calculated. The point where L_P intersects L_{BF} is, P_{BF}, as seen in Fig. 3. The intersect point, P_{BF} falls between two points on L_{BF}, BF_1 and BF_2. This is now a simpler scenario as instead of rotating BF_1 and BF_2 about the pivot until L_{BF} is aligned along the desired axis, only a single point, P_{BF}, will need rotated. Because L_P is guaranteed to pass through the pivot joint, and L_{BF} is perpendicular to L_P, when one is aligned horizontally, the other will be aligned vertically and vice versa. For the frontal and sagittal rotations, L_P is aligned horizontally to make the line of best fit vertical. In the transverse rotation, it is aligned vertically to leave the line of best fit in a horizontal position.

$$m = \frac{\sum_{i=1}^{25}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{25}(x_i - \bar{x})^2} \tag{1}$$

$$y_0 = \bar{y} - m\bar{x} \tag{2}$$

$$y_{0perp} = C_y - (\frac{C_x}{m}) \tag{3}$$

$$P_{BFx} = C_x - (\frac{y_{0perp} - y_0}{m + \frac{1}{m}}) \tag{4}$$

$$P_{BFy} = (m \times P_{BFx}) + y_0 \tag{5}$$

$$\theta = tan^{-1}(\frac{P_{BFx}}{P_{BFy}}) \tag{6}$$

Equations (1) to (6) above describe the steps taken to calculate the angle of rotation θ, around the pivot joint. Equations (1) and (2) are used to calculate the slope, m, and y intercept, y_0 of the line of best fit, L_{BF}, from a set of 2D coordinates where x_i and y_i are the x and y coordinates of joint i respectively. $\bar{x}$ and $\bar{y}$ are the means of these values. L_P is calculated once for each of the 3 anatomical planes, using positions from a subset of key joints projected on to the desired plane. The intercept of L_P is y_{0perp} in (3), and C_x and C_y represent the x and y values of the pivot point. These values are then used to determine the x and y coordinate of the point P_{BF} in (4) and (5). From here, the angle of rotation for the joints along the required plane is determined using (6), which utilises the x, P_{BFx}, and y, P_{BFy}, coordinates of the point P_{BF}. The combined result of these rotations is an upright skeleton with the shoulders facing out of the image.

One parameter that affected the output image was the subset of joints used to determine the line of best fit. This parameter can impact the outcome in two ways. The first is that by using more joints for the calculation, more of the full body shape is being considered when determining the angle that the joints by which the joints should be rotated. However, inclusion of joints that were not indicative of the direction of movement for the whole body in a single frame can result in an incorrect directional alignment. The second impact which the subset of joints can have, is that when more joints are being utilised in the rotation calculations, the rotation becomes less susceptible to cumulative noise error that may be present in each individual joint position. Ideally, a subset consisting of the maximum number of joints relevant to the direction of movement of the person is optimal for calculating the angle of rotation.

The output images were found to be particularly sensitive to this parameter when a large amount of noise was introduced to the joint positions, as was the case when considering the Gaussian augmentation. When noise is introduced frame by frame, as is the case in the Gaussian augmentation of the walking data, the rotation angle determined from a smaller set of joints can change substantially from one frame to the next. Additionally, a similar effect can occur where the rotation of the joints does not apply correctly to the extremities when the head or upper body is directed towards something outside of the path of motion. As a result, the output of such rotations would therefore not be aligned consistently.

The Gaussian augmentation applies noise to each individual joint position. When all of the joints were being used to align the skeleton contain positional noise, the alignment could suffer from the compounded error. This effect would be exaggerated with fewer joint positions being used to generate the line of best fit. Considering also that some of these joints are used in consecutive rotations to achieve the full standardisation, it is clear that small amounts of noise may accumulate into relatively large errors when less joints are being used. When applied frame by frame, the amount of noise present in the final output is so large that identifying the human skeleton shape can become difficult as seen in Fig. 4.

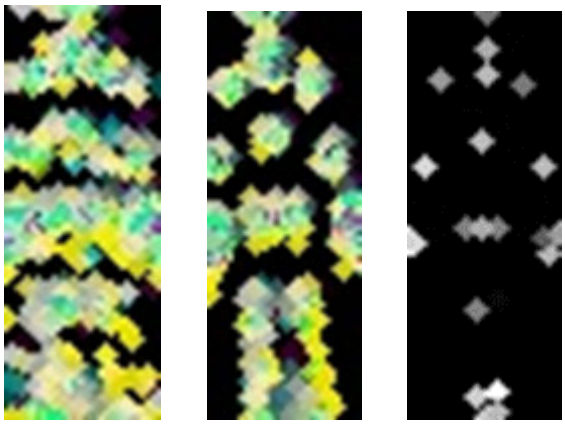

FIGURE 4. (a) Example of a JEI created using the Gaussian augmentation and the extreme case where only 2 joints are used. (b) Example of a JEI using the Gaussian augmentation and the case where a subset of more than 2 joints was used. (c) Example of grayscale image produced from joint positions.

This example represents the extreme case where only 2 joints are used to calculate a line of best fit.

The impact of noisy data can be reduced by using more than the 2 joints in the extreme case, while also limiting the joints to those that are relevant to the rotation angle being calculated. Therefore, rotations applied to the full skeleton result in outputs that still represent the initial gait data while also aligning the direction of the skeleton. The other effect of using the line of best fit is that the noise introduced to each joint, while still present, will no longer be amplified. The effects of this change can be seen in Fig. 4a and Fig. 4b, which compares a JEI created using the extreme case of only two joints with one where a larger selection of 9 joints is used for each rotation calculation.

C. JOINT POSITION NORMALISATION

Training a machine learning classifier requires that all the inputs have the same dimensionality. This presents a problem when the joint positions of different subjects are projected onto a 2D image as they will cover different portions of that image. To solve this, an image size is pre-defined, and all skeletons are projected onto it. Before the projection is applied, the raw joint positions are normalised so that the full range of x and y values are converted from meters to pixels. All the values are then scaled to fill the full x and y axis of the image. The skeleton filling the image in this way further standardises the skeleton image data to allow relative information to be used for machine learning classification while maintaining the input dimensionality constraint.

D. GRAYSCALE IMAGE GENERATION

Normalising the joint positions results in a set of 25 and 35 3D coordinates for the primary and secondary datasets respectively, where the x and y coordinates represent the x and y pixel coordinates for each joint in a 30×80 image. The third dimension, the normalised z coordinate, can now be used to represent the pixel intensity/gray value for each joint. With this information, a grayscale image can be generated for every frame.

Creating a grayscale image involves drawing a shape to represent the joint at its coordinates in the image. To this end, a circle with a radius of 2 pixels was chosen. The circle is a good shape to visualise the joints as it is simple and conceptually represents the joints found in the human body and their range of motion. The circle radius was chosen to cover enough of any single 30×80 image. In doing so, each set of joints can be identified when looking at the image.

To represent the depth of the joints visually in a grayscale image, the z coordinate of the joint was used for the pixel intensity. As the z coordinate was previously normalised to values from 10 to 255, it was already in the correct range for intensity values and could be used directly. The result is a single grayscale image with a black background and 25 white circles of the same size with varying intensities as seen in Fig. 4c. This process is applied for every frame of data, so each set of joints at each timestamp will result in a single intermediate grayscale image.

E. GAIT CYCLE SEGMENTATION AND DEPTH COLOUR MAPPING

Encoding temporal information into each image using colour can help produce a better feature for ASD classification by providing another dimension along which the data can be separated to emphasise different characteristics. This was successfully applied to the GEI's contours for the purpose of human identification, a multiclass classification problem [35]. This suggests that a similar technique could be beneficial for the JEI based on its similarity to the GEI.

Colour mapping allows the JEI to embed both time and depth into an image by using different ranges of colour for images that occur at different relative times. This is implemented by first choosing the number of gait segments per full gait cycle. Three different approaches for segmenting the gait cycle were tested; splitting a gait cycle into 3 segments of equal length (Thirds); splitting it into 4 segments of equal length (Quarters); and splitting the gait cycle into 4 segments using the distance between the feet to determine key gait events (Stance).

When using the distance between the feet to determine the segments, the first segment contained all the frames from the first frame until the distance between the feet was at its peak. The second segment included all frames from the end of the first segment until the distance between the feet was at its minimum (feet were together). The third and fourth segments repeated this pattern by taking all frames up until the feet were at their furthest and then closest points from each other, completing the gait cycle.

Encoding the depth is then achieved by mapping the pixel intensity of the grayscale image to a colour range. The result of gait cycle segmentation, is that frames from a gait cycle are assigned to either 3 segments, when the

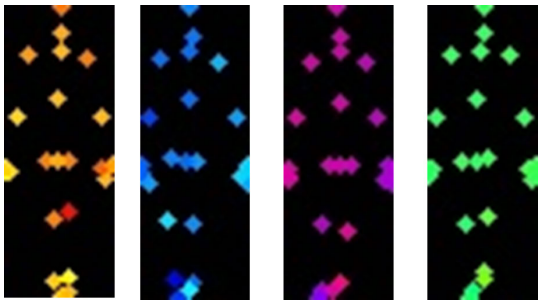

FIGURE 5. Example colour mapped frame, one from each segment of the gait cycle, showing how colour was used to represent both depth and time in a 2D image.

FIGURE 6. An example JEI, created by averaging across all colour-mapped images from the same gait cycle.

Thirds approach is used, or 4 segments, when the Quarters or Stance approaches are used. Each segment is then assigned a different colour map. The segment that a frame belongs to can be determined by which colour map was used, whereas the depth information is represented by different colours contained within each colour map. Fig. 5 shows 4 frames from different segments and therefore time ranges in the gait cycle.

As shown in Fig. 2, the result of the gait cycle segmentation and the colour mapping process is a set of images representing 1 gait cycle. Each image consists of a black background and 25 or 35 coloured circles with the colour range of the joints representing which segment of the gait cycle the frame comes from and the individual colour of each joint representing its relative depth. The final step entails averaging all the coloured frames from all segments into a single image, the output of which is a single RGB image representing a single gait cycle as seen in Fig. 6. In doing so, some information is lost. The exact joint positions from any single point in time can not be determined from only the resultant JEI. However, by averaging the images, we now have a visual representation of how the joint positions vary throughout a gait cycle in relation to each other.

F. REMOVING DUPLICATE IMAGES

Duplicate images can arise in the dataset because of the relationship between the augmented images and how the joint positions were pre-processed to obtain them. In the case of a simple translation in any direction for example, only the initial joint positions are changed in the original 3D space. Standardising the positions to a new coordinate space where the same joint is the origin removes any variation that would have been created by translating the original positions. Therefore, the two augmentations that involved a translation result in the exact same image being generated as in the original unaugmented data. Similarly, variation that would have been introduced through the scaling augmentations is removed after the joint positions are normalised to the same image size.

Allowing duplicates in the dataset would artificially increase the accuracy of any model trained on it. To address this, all the duplicate images were removed for every subject. This process of removing duplicate images was automated by using the MD5 algorithm [38] as a checksum. An image passed as input to the MD5 algorithm will produce a 128-bit hash value that will be unique to the series of pixel values that make up the image. The problem of identifying duplicate images is then reduced to comparing the hash values and looking for matches. For every matching pair, one of the associated images is removed from the training dataset. Once all of the images have been processed in this manner, only unique image instances should remain.

Removing duplicate images significantly reduces the number of total instances in our dataset. Depending on which combination of the 3 joint rotations were used during the standardisation stage, a different set of augmentations would produce duplicate images. The number of remaining instances after duplicates were removed for each test, is reported in Table. 3. Reducing the size of the dataset is not ideal as models trained on it can more easily suffer from overfitting. In the case of the datasets used in this study, having less augmented instances means the final models may not be able to generalise as well on the testing set. The advantage, however, is that the accuracy values obtained by these models are more representative of how they would perform in a real-world scenario as they are not impacted by duplicate data.

TABLE 3. Breakdown of augmentation instances and size of the dataset used in each test.

Dataset	Test	Augmentations Removed	Remaining Instances
P	1	3	500
P	2	2	600
P	3	4	400
S	ALL	5	39

IV. CLASSIFICATION

A. DIMENSIONALITY REDUCTION

The JEI images require some pre-processing before being applied to the machine learning models for training. Every image is 30 pixels wide, 80 pixels high and 3 colour channels

deep, resulting in 7200 total integer values as individual features representing 1 training instance. A single input to the classifier, if left unchanged, would therefore have 7200 dimensions.

Reducing the number of dimensions through Dimensionality Reduction (DR) is beneficial for two reasons. The first is that using such high dimensional data for training is inefficient as not all dimensions contain useful information for classification. Take for example a pixel on the image that does not have any joints near it, and therefore retains the same value across all input images. There is no useful information here for determining whether an image belongs to a subject with ASD or not as the pixel value never changes regardless of which class that subject belongs to. The outcome of DR is therefore a refined set of features more focused for the ASD classification problem.

The second benefit of reducing the number of dimensions for training is that it will require less processing power, and therefore time, to train the models. Although training a single model on such a small dataset is relatively quick when compared to larger datasets, employing a cross validated grid search (covered in the classification section) can take a lot of time and computing resources. This is dependent on the number of hyperparameters that will need to be tuned, with each new value to test increasing the number of models needed for training exponentially. Reducing the time required to train a single model will therefore compound the amount of time saved in achieving the optimal model due to the number of them being trained throughout the entire pipeline. As this work is not investigating the real-time application of any implementations, the timings for each of the tests were not important, and therefore not recorded.

PCA was chosen for DR of the primary and secondary datasets. To apply PCA to the image data, it must first be reshaped from a $30 \times 80 \times 3$ matrix into a 1-dimensional list of 7200 values. This process is called flattening the image. The cut-off point for the number of final components can have a large impact on the accuracy of any model trained on it when using PCA. Too few components and there may still not be enough useful information to accurately classify the inputs. Too many components however, and the benefit from reducing the number of dimensions for training is decreased. As it can be difficult to determine the optimal cut off value, it was instead encoded as a hyperparameter during the grid search phase of the classification process. This allows several different cut off values to be tested and validated, resulting in an optimal value being chosen from a range of values.

B. MACHINE LEARNING MODELS

Three model types, a RF, MLP and CNN, were selected to be trained in the ASD classification problem. The RF model was chosen due to its success and strong performance over other classic machine learning models in related studies on small ASD datasets [25]. The MLP was chosen because it was applied to a feature set derived from the same raw data, being the 3D joint positions, in another study and it achieved 95% accuracy [36]. Lastly, the CNN was chosen because of its ability to work well on image-based data in classification problems [32].

The base algorithm of the RF model uses an ensemble of weak learners, in this case DTs, in a voting system to produce a more generalised, strong learner. The number of weak learners that are trained is a tunable hyperparameter. For each DT within the RF classifier, the training samples are recursively split into two groups. A split occurs at each node Q_m between n_m samples in the DT, where n_m refers to the the number of samples n at the m^{th} node. A potential split, $\theta = (j, t_m)$ consists of a feature j, and a threshold t_m, where all of the samples at that node are split using the chosen feature at the chosen threshold into the subsets $Q_m^{left}(\theta)$ and $Q_m^{right}(\theta)$. The quality of any θ at node m, is calculated using an impurity function, H applied to both $Q_m^{left}(\theta)$ and $Q_m^{right}(\theta)$. They are weighted to the percentage of samples the two child nodes contain with respect to the total number of samples the parent node contained, resulting in a final evaluation of θ, $G(Q_m, \theta)$.

$$G(Q_m, \theta) = \frac{n_m^{left}}{n_m} H(Q_m^{left}(\theta)) + \frac{n_m^{right}}{n_m} H(Q_m^{right}(\theta)) \quad (7)$$

The impurity function is applied to all candidate splits and the split with the smallest G is chosen as the best split θ^*.

$$\theta^* = \text{argmin}_\theta \, G(Q_m, \theta) \quad (8)$$

This is then repeated for $Q_m^{left}(\theta^*)$ and $Q_m^{right}(\theta^*)$, until either the maximum depth is reached, the number of samples at a node drops below a threshold, or there is only one sample left if no threshold is provided. The RF algorithm can also use a bootstrapping method by randomly sampling the training instances and features with replacement from the complete dataset to supply to each DT. Each DT is therefore trained on a different combination of samples and features. This is why the individual trees are considered weak learners. Instead of deciding to use bootstrapping or not, it was treated as a trainable parameter along with the number of features to sample at each node.

The MLP follows a typical neural network structure of multiple layers of many interconnected perceptrons. Given a set of training examples $(x_1, y_1), \ldots, (x_n, y_n)$ where x is the set of features as inputs, and y is the output label, with one hidden layer and one hidden neuron, the network learns the function:

$$f(x) = W_2 g(W_1^T x + b_1) + b_2 \quad (9)$$

where W_1, W_2, b_1, and b_2 are model parameters. W_1 and W_2 represent the weights of the input layer and hidden layer, respectively; and b_1 and b_2 represent the bias added to the hidden layer and the output layer, respectively. During the grid search performed as part of the model training, the Hyperbolic Tan, Sigmoid, and the Identity activation functions were all tested as well as different numbers of hidden layers and nodes. The ASD classification problem is a type of binary classification with the two possible outputs

being either "ASD" or "TD". The output of the function $f(x)$), learned by the MLP, is therefore passed through the logistic function to obtain output values between 0 and 1:

$$g(z) = 1/(1 + e^{-z}) \tag{10}$$

A threshold value of 0.5 is then used to determine to which group the input instance is assigned. Starting with random weights, the Average Cross-Entropy loss function is minimised by repeatedly updating the weights through back propagation. The new weight is calculated using Gradient Descent as $W^{i+1} = W^i - \epsilon \nabla Loss^i_W$, where i is the iteration step, and ϵ is the learning rate. Learning stops when the preset maximum number of iterations, another trainable parameter, is reached.

The CNN follows a similar structure to the MLP, with an input layer, multiple hidden layers, and an output layer. Unlike the MLP, there are multiple types of hidden layers within a CNN. Firstly, the Convolutional Layer takes a 2-D (3-D if its the first hidden layer) matrix as input with M x N dimensions, and applies a series of 3 $\times$ 3 filters F in a convolution operation over the input I. At each step of the convolution t, the dot product of the filter and the input at t, $F.I(t)$, is calculated and stored in a matrix with a width of $M - 2$ and a height of $N - 2$. The filter is then moved across the image with a stride of 1, where the next dot product is calculated. The resultant matrix is then passed to a Max Pooling layer, where the same convolutional operation is used, instead taking the max element contained within the kernel at each step. The kernel is of size 2 $\times$ 2 with a stride of 2. The initial convolutional layer takes the 80 $\times$ 80 x 3 image matrix as input. It therefore applies each filter to the image 3 times, one for each colour channel, and the output at each convolution step is the summation of the dot products of all 3 channels at step t, $O = F.I_R(t) + F.I_G(t) + F.I_B(t)$. Lastly, after each Convolutional and Max Pooling Layer pair, a Flattening layer is applied that flattens the input from the last max Pooling layer into a 1-D array before being passed through a final fully connected layer before going to the output layer. The number of layers, the number of filters per layer, and the activation function for the output layer were all tuned as hyperparameters during the grid search. The activation functions of the output layer were limited to either the sigmoid function, so that a single output neuron could be utilised for an ASD or non-ASD binary output, or the soft max function, so that 2 output neurons could be utilised as probabilities for the ASD and TD prediction classes, respectively. All filters of the convolutional layers used the Rectified Linear Unit (ReLu) activation function and were updated using the same back propagation process as was used for the MLP.

A 5-fold cross validated grid search was employed during training. A hyperparameter grid was therefor created defining the different variables to consider. The parameter options for each model can be seen in Table. 4. The format of the Hidden Layer and Convolutional Layer parameter options, $(l_1, \ldots, l_n)$, show both the number of layers and the size of each layer where l_n is the size of the n^{th} layer.

TABLE 4. Parameter options used in grid search.

Model	Parameter	Options
MLP	Activation Function	Identity*, Sigmoid, Hyperbolic Tan
	Hidden Layers	(16, 8, 4)*, (30, 20, 15, 10)
	Max Iterations	150*, 300
	No. PCA Components	45, 64, 75*
RF	Split Criterion	Gini*, Shannon Entropy
	Max Features	3*, Square Root
	No. Estimators	50, 150*
	No. PCA Components	45*, 64, 75*
CNN	Dense Layer Activation	Sigmoid, Softmax*
	Epochs	10*, 20
	Convolutional Layers	(32,64,64)*, (8,16,32,64)

*Most commonly found in best model across all tests.

C. RF AND MLP TRAINING AND CLASSIFICATION PIPELINE

A pipeline was developed for training the RF and MLP models. The first step of the pipeline involved normalising the data using the L2 norm function so that all feature values were in the same range. Following this, the dataset was shuffled and split into training and testing sets. The training set consisted of 70% of the ASD subjects data and 70% of the TD subjects data. The testing set therefore consisted of 30% of each of the ASD and TD groups individually. Only original, unaugmented data instances were used in the testing set as the purpose of the augmented data is to increase the ability for the models to generalise to different conditions. The result is therefore that post train-test split, all augmented data instances were removed from the testing set, meaning only 30 instances were viable for the test set when using the primary dataset, and only 5 instances were viable when using the secondary dataset.

A 5-fold cross validated grid search was then employed during training. One hyperparameter that was contained in both models hyperparameter grids was the cut off value for the number of components to retain during PCA. A 5-fold cross validated grid search was then employed during training, which included both the PCA process as well as model training, to produce 3 different metrics for evaluation.

The first metric to be calculated was the model accuracy. The cross validated grid search was applied to the training set to obtain a set of hyperparameters for each model type that performed the best on unseen data. This model was then applied to the test set and the accuracy measured based on its predictions. This metric was used in the paper that accompanies the dataset used in this research and so it is useful for comparative purposes. The accuracy is calculated using Equation 11, where TP, and TN, are True Positive and True Negative, respectively. TP represents the number of

instances correctly identified as ASD, and TN is the number of instances correctly identified as TD. FP and FN therefore represent False Positive and False Negative instances, which are the cases where an ASD prediction was incorrectly given for a TD instance, and a TD prediction was incorrectly given for an ASD instance, respectively.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{11}$$

Although cross validation was used during the grid search, there remains one source of potential overfitting with the accuracy values it produces. As the best model from the grid search is applied to the testing set, the resultant accuracy can still be largely affected by which instances were retained for testing and which for training. This effect can be exaggerated by the fact the datasets being used are small with the testing sets consisting of only 30 and 5 instances in total. Therefore, to further evaluate the models, nested cross validation was also applied to the training set. In nested cross validation, for every iteration, once the validation fold is separated, and the remaining 4 folds (in 5-fold CV) are combined for training, the new training set is now applied to the k-fold cross validated grid search. For this research, the number of folds selected for this inner cross validation was 5. This process is then repeated for every iteration of the outer 5-fold cross validation loop. The result of nested cross validation is an accuracy value for each outer loop of cross validation. These are then averaged to produce a second metric for comparing models, the mean nested cross validation accuracy. It can be calculated using Equations 12 and 13. The values of k_{Inner} and k_{Outer} represent the number of folds in the inner and outer loops of the nested cross validation process. The accuracy in equation 12 $Accuracy_i$, is the accuracy obtained using Equation 11 on the i^{th} configuration of the folds during the inner CV loop. The output, $Accuracy_{Inner}$, is the mean accuracy of models trained and tested on the different inner folds. The output of Equation 13, is therefore the mean of the outputs from Equation 12, applied to each of the outer loop fold splits.

$$Accuracy_{Inner} = \frac{1}{k_{Inner}} \sum_{i=1}^{k_{Inner}} Accuracy_i \tag{12}$$

$$Accuracy_{Outer} = \frac{1}{k_{Outer}} \sum_{j=1}^{k_{Outer}} Accuracy_{Inner}j \tag{13}$$

The accuracies obtained from these equations can originate from models with different optimal hyperparameters. This means that an optimal hyperparameter selection for our models cannot be determined using this process, and these should be selected using the normal cross validation approach. The advantage of this approach is that the resultant accuracies give a less biased estimation on how the models will perform on unseen data.

A third and final metric was calculated to further evaluate the models. This metric, called the Test Time Augmented (TTA) accuracy, is designed to assess the average performance of the models in different conditions as defined by how the data is augmented. The augmented data is designed to generalise the dataset to a larger variation of conditions, the TTA accuracy therefore measures how well this dataset has generalised. To calculate the TTA accuracy, the optimal model from the original cross validated grid search was applied to the test set again. This time however, the augmented instances were reintroduced into the test set and were also applied to the optimal model. Following this, the predicted label for each instance and its augmentations were tallied in a voting system, with the most voted for label being assigned to the original instance as is visualised in Fig. 7. In the case of the same number of votes for the ASD label and the TD label for a set of instances, the vote that was assigned to the original instance was used. Predicted labels assigned to the original instance only were then used to calculate the final TTA accuracy value. Equation 11, was used on the voted predictions to determine the final TTA accuracy of the model. Fig. 8 shows the original versus augmented makeup of the train and test splits for each of the 3 metrics.

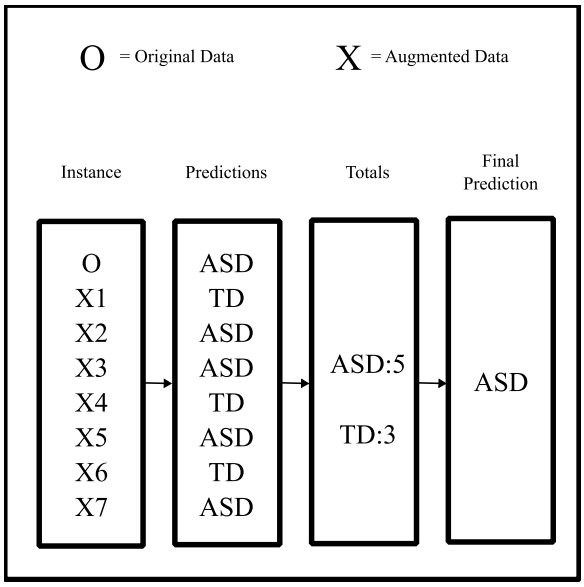

FIGURE 7. Diagram showing the voting process to obtain final predictions for the TTA accuracy metric.

This pipeline was then repeated for each model type 3 times, each time using a different train-test split of the data. This means that the 70% of data used for training consisted of different instances for each repetition. The same can then be said for the instances in the test set. By repeating the full pipeline 3 times and taking the average for each metric, any effect of bias due to the split of the data is further reduced.

A different pipeline was used when training the CNN model as can be seen in Fig. 1. A CNN can work directly with the multidimensional structure of an RGB image like a JEI. This is reflected in the adapted pipeline as the image data is not transformed into a 1D list, nor are the values normalised

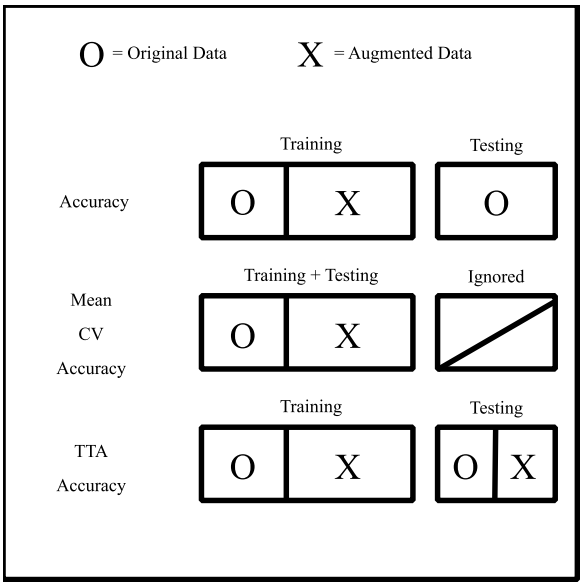

FIGURE 8. Diagram showing how the data was split and used for training and testing to obtain the 3 metrics.

like in the RF and MLP pipeline. The CNN models therefore take in matrices with dimensions of 30 × 80 x 3 for training and testing. This is advantageous because the CNN has been found to be useful for classification problems using images as input. It then uses a series of 2D convolutional layers, the number and size of which is defined by the Convolutional Layers parameter in Table. 4. A 3 × 3 kernel size was used in these layers. Each convolutional layer is followed by a max pooling layer for down-sampling the output. After the final max pooling layer, the output is flattened into a 1D array and passed through 2 densely connected layers. The second densely connected layer is used as the classification layer.

D. STATISTICAL ANALYSIS

The models were trained under the same conditions, three times for each test configuration. These tests are also repeated for the Secondary Dataset. Statistical tests were applied to the accuracy results in order to answer two questions:

1) Do any of the models perform significantly better than the others?
2) Do any of the JEI configurations utilised in the different tests lead to a notable enhancement to the classification results of the selected models?

To answer these questions, a Two-Way ANOVA test was used to determine if any significance could be observed between either the models accuracies or the test configurations accuracies in the Primary dataset. The Shapiro-Wilk and Levene tests were first applied to confirm the data satisfied the conditions of normality and homogeneity of variance required for ANOVA. The same process was used for applying the Two-Way ANOVA test to the results from the Secondary Dataset. If the Two-Way ANOVA results showed significance between any of the groups, then the Tukey

TABLE 5. An ANOVA table for results on the primary dataset.

Source	SS	df	MS	F	p
Models	79.91	2	39.96	1.65	0.2205
Tests	40.28	2	20.14	0.83	0.4521
Interaction	80.66	4	20.17	0.83	0.5229
Error	436.85	18	24.27		
Total	637.7	26			

TABLE 6. An ANOVA table for results on the secondary dataset.

Source	SS	df	MS	F	p
Models	4965.34	2	2482.67	6.70	0.0067
Tests	2998.24	2	1499.12	4.04	0.0354
Interaction	1623.98	4	405.99	1.10	0.3884
Error	6666.67	18	370.37		
Total	16254.23	26			

TABLE 7. A Tukey honest significant difference table for results between models on the secondary dataset.

Group 1	Group 2	MD	p-adj	Lower CI	Upper CI
CNN	MLP	-26.67	0.040	-52.20	-1.14
CNN	RF	-31.11	0.015	-56.64	-5.58
MLP	RF	-4.44	0.902	-29.98	21.09

TABLE 8. A Tukey honest significant difference table for results between test configurations on the secondary dataset.

Group 1	Group 2	MD	p-adj	Lower CI	Upper CI
CNN	MLP	-26.67	0.040	-52.20	-1.14
CNN	RF	-31.11	0.015	-56.64	-5.58
MLP	RF	-4.44	0.902	-29.98	21.09

Honestly Significant Difference post-hoc test was applied. The results of the Two-Way ANOVA test conducted on the primary and secondary datasets are displayed in Table. 5, and Table. 6, respectively. The Two-Way ANOVA found no significance with either group on the primary dataset and so no post-hoc test was applied. The Two-Way ANOVA did find significance for both the models and test configurations groups in the secondary dataset however. The results of these tests are displayed in Table. 7 and Table. 8, respectively.

V. RESULTS

Three sets of tests were applied, each producing the 3 calculated accuracy metrics for each of the 3 classifiers. Each test used a different combination of the joint positions for the line of best fit calculation, rotations and gait cycle segmentation techniques. The combinations that were used to produce the JEI for each test are defined in Table. 2. Each test was repeated 3 times with the results averaged for each classifier to provide a more reliable measure of performance. A total of 15 tests had been used for the primary dataset to determine the optimal combination of methods for generating

TABLE 9. Accuracy results from a subset of the executed tests from all 3 models.

Test Number	RF Accuracy (%)	RF Nested CV Accuracy (%)	RF TTA Accuracy (%)	MLP Accuracy (%)	MLP Nested CV Accuracy (%)	MLP TTA Accuracy (%)	CNN Accuracy (%)	CNN Nested CV Accuracy (%)	CNN TTA Accuracy (%)
1	64.44	85.33	91.11	66.67	82.67	90.00	88.89	76.86	95.56
2	58.89	83.16	85.56	74.44	86.26	91.11	84.44	79.44	92.22
3	74.44	83.10	87.78	83.33	84.52	92.22	86.67	77.14	88.89

TABLE 10. Accuracy results from a subset of the executed tests from all 3 models using the secondary dataset.

Test Number	RF Accuracy (%)	RF Nested CV Accuracy (%)	RF TTA Accuracy (%)	MLP Accuracy (%)	MLP Nested CV Accuracy (%)	MLP TTA Accuracy (%)	CNN Accuracy (%)	CNN Nested CV Accuracy (%)	CNN TTA Accuracy (%)
1	60.00	60.83	40.00	60.00	81.67	40.00	70.00	64.00	70.00
2	60.00	48.89	40.00	80.00	76.67	40.00	93.33	60.00	80.00
3	66.67	74.00	66.67	80.00	85.00	80.00	80.00	65.87	80.00

the JEI. Results from 3 of these were determined to provide optimal results depending on the classifier used. Only these three tests were then repeated for the Secondary dataset. The averaged results from these 3 repeated tests for both datasets are presented in Table. 9 and Table. 10 respectively. The averaged TTA accuracies are additionally presented in Fig. 9 and Fig. 10, respectively.

In first comparing the accuracy measure from each of the 3 classifiers across all 3 of the above tests for the primary dataset, the CNN classifier achieves consistently higher accuracy values. The CNN also achieves the best baseline accuracy of 88.89% in the first test. When applied to the secondary dataset, the CNN also performed consistently well, achieving the best baseline accuracy of 93.33% across the 3 tests. Comparatively, the RF and MLP classifiers had very low accuracies in test 1 at 64.44% and 66.67% respectively. This would indicate that their ability to generalise to unseen data is not as strong as in the CNN for ASD gait datasets. However, in terms of the nested cv accuracies, both the RF and MLP classifiers outperform the CNN across all 3 tests achieving a highest of 85.33% in test 1 and 86.26% in test 2, respectively. This holds true for the MLP classifier when using the secondary dataset, although the nested cv accuracy of the RF classifier fell below the CNN in these tests. As the nested cv accuracies are calculated based on performance on the training set, which contained the augmented data, it is possible that the RF and MLP models are more sensitive to the variation introduced through the augmented data. This would explain both their higher nested cv accuracy measures and their low base accuracies. For the CNN, the opposite may be true as it has performed better on the unaugmented test data.

Further evidence can be found that the CNN is best suited to the ASD classification problem from consideration of the TTA accuracy when using the JEI feature, as it was able to achieve a TTA accuracy of 95.56% in test 1, now outperforming the reported accuracy from the datasets original paper of 95%. This shows that the models can be

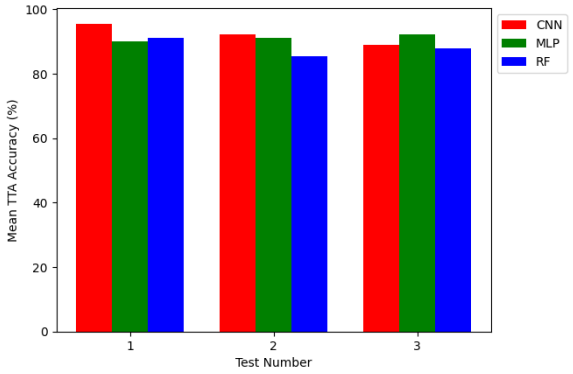

FIGURE 9. Graph showing the TTA accuracies of models trained on the primary dataset across all tests.

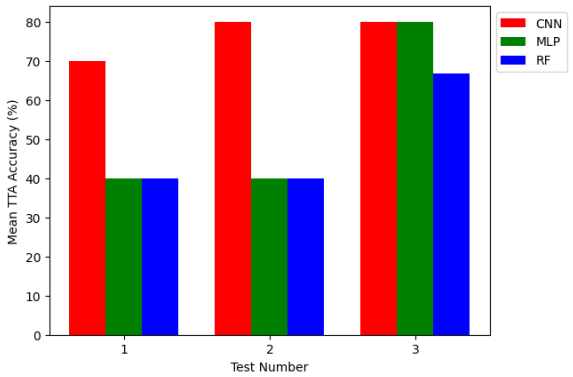

FIGURE 10. Graph showing the TTA accuracies of models trained on the secondary dataset across all tests.

greatly improved by using the augmentations at test time in a voting system to acquire the final predictions for the original instances. As real-world data can be augmented in the same way as the original instances in this dataset, it would simply

TABLE 11. Accuracy results from a range of published works.

Paper	Model	Accuracy (%)	No. Participants
Gait Analysis			
Current (Primary)	CNN	95.56*	100
Current (Secondary)	CNN	93.33	20
[36]	MLP	95.00	100
[13]	LDA	82.50	48
[37]	ANN	91.70	60
[39]	SVM	100.00	44
[40]	ANN	95.80	44
Repetative Behaviors			
[41]	CNN	98.29	6
[42]	LSTM	95.20	-
Gaze Pattern			
[43]	XGBoost	99.80	28
fMRI			
[22]	DT	83.00	92
[44]	LSTM	80.00	221

*TTA accuracy is reported.

become a further step in the classification pipeline before predictions are made in a real-world application.

Considering the test results from the secondary dataset in Table. 10, the CNN classifier also obtained a best baseline accuracy of 93.33%. It therefore outperformed the artificial neural network in [37], where a best baseline accuracy of 91.7% was reported. The secondary dataset used in the current study was derived from a subset of the raw data collected in the cited paper. Due to the similarity of the data used for training, it can be deduced that the CNN combined with the JEI is a competitive feature for ASD classification to raw temporal and kinematic features being used with other neural network architectures.

Accuracy performances of different models from other studies performing the ASD classification problem are presented for comparison in Table. 11. Although the presented methods best accuracy of 95.56% falls below some of the cited accuracies, the higher number of participants increases its reliability when applied to real-world scenarios within the gait analysis field. For example, [39] and [40] achieved 100 and 95.8% respectively while using a dataset of model-based features collected from 44 participants. Another similar study [13], again using model-based kinematic and kinetic gait features, achieved 82.5% accuracy with 48 participants. Under half the number of participants' data were used in the training and testing of their models when compared to the primary dataset, making overfitting to the data more likely. Additionally, the data in the cited studies was collected using a combination of marker-based tracking systems and force plates in more controlled environments. This shows that the model from this paper is able to remain competitive despite the data being collected in less controlled recording conditions and with less invasive and expensive equipment. Additionally, the performances recorded when the JEI was applied to the secondary dataset show the adaptability of such a feature even when using different joint position configurations and recording equipment. Comparing to other methods of ASD classification in Table. 11, the JEI and CNN combination achieve similar results. The benefit of gait analysis methods using the Kinect are that they don't require the presence of repetitive behaviours, nor intrusive methods of data collection as is the case for fMRI scans. Gaze patterns analysis provides a viable alternative to gait analysis based on the accuracy of 99.80% achieved in [43], however, they also only use data from 28 participants. Both approaches, Gaze and Gait, could benefit form testing on larger datasets before being directly compared. Alternatively, a combination of these methods could lead to even more reliable results.

The statistical analysis results between the model types found that in the Primary dataset, none of the models performances were significantly better than the others. The CNN performance in the Secondary dataset however was found to be significantly higher than both the RF and MLP performances. This implies that when enough data is available, all three models can perform similarly in terms of significance, however for more difficult datasets such as the Secondary datasets when less overall data is available, the CNN combined with the JEI performs significantly better. Statistical analysis between Tests 1, 2, and 3, found that no significant performance difference could be observed between the three JEI configurations. This could indicate that choosing specific joint selection, rotations, and gait cycle segmentation do not increase classification accuracy by a meaningful amount.

VI. CONCLUSION AND FUTURE DIRECTIONS

This report presents an appearance-based feature that can be created from 3D joint positions that are commonly used for model-based feature sets. The JEI embeds both temporal and depth information of the joints into a 2D image using colour maps. In total, 3 classifier model types were trained across 3 tests that varied the implementation of the components that the JEI is composed of. The models were RFs, MLPs and CNNs and were applied to a primary and secondary dataset separately. All models were assessed using 3 accuracy metrics. The baseline accuracy measures performance on a test set using the best parameter results from a 5-fold cross validated grid search. The nested mean cross validated accuracy measures performance on unseen data using the training set, using 5-folds for both the outer and inner loops of the algorithm. Lastly, TTA accuracy was measured on the test set, making use of the generalised models to improve the classification accuracy by using a voting system on instances and their augmentations.

The CNN obtained 95.56% TTA accuracy and 88.89% baseline accuracy in test 1 of the primary dataset tests. In test 2 of the secondary dataset tests, it achieved 80% TTA accuracy and 93.33% baseline accuracy. For the primary dataset, this outperformed both the RF and MLP models. The results achieved by the CNN show both its own viability as well as the viability of the JEI to compete with state-of-the-art feature sets and performances. These results also highlight

the effectiveness of using the CNN on appearance-based data for ASD classification. This is further supported by the fact that the 95.56% accuracy was achieved using only 4 out of 7 of the augmentations for training, where the other augmentations were made redundant as the JEI already accounts for translation and scaling changes.

In test 1, where the highest baseline accuracy occurred, one of the augmentations was removed due to its incompatibility with the JEI, which when included, was reducing accuracies of models significantly. The augmentation introduced noise to each joint position, showing that this version of the JEI will perform best with clean data. Another iteration of the JEI dealt with the sensitivity to noise, achieving 86.67% baseline accuracy in test 3 when the augmentation was included. A reduction of only 2% accuracy.

The high accuracy of 93.33% being achieved on the secondary dataset is also important due to the variation in the datasets' recording equipment and joint position configuration compared to the primary dataset. Whereas the primary dataset was recorded with a Kinect device in an uncontrolled environment, the secondary dataset used a high accuracy Vicon motion capture system in a more controlled setting. Additionally, the Kinect tracked 25 joints whereas the Vicon tracked 35 retro-reflective markers. These differences, combined with the high reported accuracies when using both setups, shows the adaptability of the JEI and CNN for ASD classification.

In terms of nested cross validation accuracy, the CNN did not perform as well as the MLP and RF models which had an accuracy of 86.26% in test 2 and 85.33% in test 1, respectively. Although the CNN did not perform as well along this metric, achieving only 79.44% in test 2, the CNN has the capability to be improved. This could be achieved by using deeper and more complex network structures than the 3 or 4 convolutional layers arranged sequentially as tested in this study.

The main contributions of this study are the development of the JEI, a new appearance based feature that incorporates depth and temporal data. Machine learning models were trained on the JEIs produced from two different datasets, for the ASD classification problem. The models were assessed using the TTA accuracy metric, finding that it is competitive with state-of-the-art model performances and feature sets.

A. LIMITATIONS

Limitations of the contribution include the small dataset sizes and the limited exploration of different structures for the neural networks. Although the size of the datasets allow for comparison to other ASD classification studies investigating gait analysis, they are limited in the fact that they only represent a very small proportion of potential cases of ASD considering the high global prevalence. However, the use of augmented data to improve the accuracy of models on unseen data has shown to be beneficial, albeit exposure to a greater number of real-world instances would assist the model's exposure to more variation, ultimately impacting the level of confidence that can be given to these models when applied to real-world scenarios.

B. FUTURE WORK

Future work should focus on improving the size of gait datasets so that they include more of the variation present in the gait of the world's ASD population. The statistical test results presented in Section V, of this paper, show that dataset size could be more important when the aim is to obtain optimal ASD classification results. Therefore, by collating more data into one common dataset for training and testing, better models can be produced. Variation in the demographic of such a dataset should also be prioritised. With a demographically balanced dataset, the outcomes would be more reliable, due to most studies usually only using data from the same geographical areas. Future work would therefore benefit from a cross-community or multi-nation effort. Care should also be taken to increase the balance between male and female participants. Although female participants are harder to enlist for ASD studies, their representation is important so that results aren't artificially altered by their exclusion.

Most of the limitations of this study could be addressed by collecting more data. However, it is important to note that the current lack of larger datasets is likely due to the difficulty in collecting gait data from ASD participants. A contributing factor to the difficulty of data collection is the use of intrusive collection devices like the Vicon and it's retro-reflective marker system. Alternatives like the Kinect used in this paper's primary dataset could present a solution to this problem. Future data collection efforts could be aided by the use of these less intrusive depth cameras. To this end, the pose estimation quality of such devices should be checked for viability with use on ASD gait data. This would begin to answer the question of whether the reduction in pose accuracy from using less intrusive devices affects the classification accuracy of models by a small enough degree to be worth the advantage of being able to more easily collect additional data in different environments.

Larger datasets would also present another avenue of future work. As more data is collected, the complexity of the data also increases. The extra complexity comes from more gait samples from different people being included, increasing the variation between participants. This would be a good opportunity to allow the performance of more complex model architectures to be investigated under more realistic conditions. The CNN architecture used for this study is relatively simple compared to the depth of other successful CNNs in other image classification tasks such as the VGG-19 [45] and its 19 layers. Although suitable for this study, as the datasets increase in size, different architectures should be explored.

ACKNOWLEDGMENT

The authors would like to thank Dr. Che Zawiyah Che Hasan for their open communication and willingness to help by providing additional data for use in this study.

REFERENCES

[1] N. H. Mohamed and A. B. M. Kassim, "The global prevalence and diagnosis of autism spectrum disorder (ASD) among young children," *Southeast Asia Psychol. J.*, vol. 7, pp. 1–14, Oct. 2019.

[2] A. Y. Onaolapo and O. J. Onaolapo, "Global data on autism spectrum disorders prevalence: A review of facts, fallacies and limitations," *Universal J. Clin. Med.*, vol. 5, no. 2, pp. 14–23, Dec. 2017.

[3] R. J. Landa, "Efficacy of early interventions for infants and young children with, and at risk for, autism spectrum disorders," *Int. Rev. Psychiatry*, vol. 30, no. 1, pp. 25–39, Jan. 2018, doi: 10.1080/09540261.2018.1432574.

[4] E. A. Fuller and A. P. Kaiser, "The effects of early intervention on social communication outcomes for children with autism spectrum disorder: A meta-analysis," *J. Autism Develop. Disorders*, vol. 50, no. 5, pp. 1683–1700, May 2020, doi: 10.1007/s10803-019-03927-z.

[5] D. Farrugia, "Exploring stigma: Medical knowledge and the stigmatisation of parents of children diagnosed with autism spectrum disorder," *Sociol. Health Illness*, vol. 31, no. 7, pp. 1011–1027, Nov. 2009.

[6] K. L. Ashwood, J. Buitelaar, D. Murphy, W. Spooren, and T. Charman, "European clinical network: Autism spectrum disorder assessments and patient characterisation," *Eur. Child Adolescent Psychiatry*, vol. 24, no. 8, pp. 985–995, Aug. 2015, doi: 10.1007/s00787-014-0648-2.

[7] L. A. Wang, V. Petrulla, C. J. Zampella, R. Waller, and R. T. Schultz, "Gross motor impairment and its relation to social skills in autism spectrum disorder: A systematic review and two meta-analyses," *Psychol. Bull.*, vol. 148, nos. 3–4, pp. 273–300, 2022.

[8] A. Pradhan, V. Chester, and K. Padhiar, "Classification of autism and control gait in children using multisegment foot kinematic features," *Bioengineering*, vol. 9, no. 10, p. 552, Oct. 2022. [Online]. Available: https://www.mdpi.com/2306-5354/9/10/552/htm and https://www.mdpi.com/2306-5354/9/10/552

[9] A. N. Olivas, M. R. Kendall, A. Parada, R. Manning, and J. D. Eggleston, "Children with autism display altered ankle strategies when changing speed during over-ground gait," *Clin. Biomech.*, vol. 100, Dec. 2022, Art. no. 105804.

[10] E. Biffi, C. Costantini, S. B. Ceccarelli, A. Cesareo, G. M. Marzocchi, M. Nobile, M. Molteni, and A. Crippa, "Gait pattern and motor performance during discrete gait perturbation in children with autism spectrum disorders," *Frontiers Psychol.*, vol. 9, pp. 1–13, Dec. 2018.

[11] C. Z. C. Hasan, R. Jailani, N. Md Tahir, and S. Ilias, "The analysis of three-dimensional ground reaction forces during gait in children with autism spectrum disorders," *Res. Develop. Disabilities*, vol. 66, pp. 55–63, Jul. 2017, doi: 10.1016/j.ridd.2017.02.015.

[12] B.-O. Lim, D. O'Sullivan, B.-G. Choi, and M.-Y. Kim, "Comparative gait analysis between children with autism and age-matched controls: Analysis with temporal-spatial and foot pressure variables," *J. Phys. Therapy Sci.*, vol. 28, no. 1, pp. 286–292, 2016.

[13] C. Z. C. Hasan, R. Jailani, N. M. Tahir, I. M. Yassin, and Z. I. Rizman, "Automated classification of autism spectrum disorders gait patterns using discriminant analysis based on kinematic and kinetic gait features," *J. Appl. Environ. Biol. Sci.*, vol. 7, no. 1, pp. 150–156, 2017. [Online]. Available: https://www.textroad.com

[14] C. Z. Hasan, R. Jailani, N. M. Tahir, and H. M. Desaa, "Vertical ground reaction force gait patterns during walking in children with autism spectrum disorders," *Int. J. Eng., Trans. B, Appl.*, vol. 31, no. 5, pp. 705–711, 2018.

[15] C. Z. C. Hasan, R. Jailani, and N. Tahir, "Automated classification of gait abnormalities in children with autism spectrum disorders based on kinematic data faculty of electrical engineering," *Int. J. Psychiatry Psychotherapy*, vol. 2, no. August, pp. 10–15, 2017.

[16] V. L. Chester and M. Calhoun, "Gait symmetry in children with autism," *Autism Res. Treatment*, vol. 2012, pp. 1–5, Jan. 2012.

[17] S. Chakraborty, A. Nandy, T. Yamaguchi, V. Bonnet, and G. Venture, "Accuracy of image data stream of a markerless motion capture system in determining the local dynamic stability and joint kinematics of human gait," *J. Biomech.*, vol. 104, May 2020, Art. no. 109718, doi: 10.1016/j.jbiomech.2020.109718.

[18] A. Pfister, A. M. West, S. Bronner, and J. A. Noah, "Comparative abilities of Microsoft Kinect and vicon 3D motion capture for gait analysis," *J. Med. Eng. Technol.*, vol. 38, no. 5, pp. 274–280, Jul. 2014.

[19] A. Ardalan, A. H. Assadi, O. J. Surgent, and B. G. Travers, "Whole-body movement during videogame play distinguishes youth with autism from youth with typical development," *Sci. Rep.*, vol. 9, no. 1, pp. 1–11, Dec. 2019, doi: 10.1038/s41598-019-56362-6.

[20] S. Rahman, S. F. Ahmed, O. Shahid, M. A. Arrafi, and M. A. R. Ahad, "Automated detection approaches to autism spectrum disorder based on human activity analysis: A review," *Cognit. Comput.*, vol. 14, no. 5, pp. 1773–1800, Sep. 2022. [Online]. Available: https://link.springer.com/article/10.1007/s12559-021-09895-w

[21] W. Feng, G. Liu, K. Zeng, M. Zeng, and Y. Liu, "A review of methods for classification and recognition of ASD using fMRI data," *J. Neurosci. Methods*, vol. 368, Feb. 2022, Art. no. 109456.

[22] F. Zhao, Z. Chen, I. Rekik, S.-W. Lee, and D. Shen. (2020). *Diagnosis of Autism Spectrum Disorder Using Central-Moment Features From Low- and High-Order Dynamic Resting-State Functional Connectivity Networks*. [Online]. Available: https://www.cdc.gov/ncbddd/autism/data.html

[23] N. K. Zakaria, "Experimental approach in gait analysis and classification methods for autism spectrum disorder: A review," *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 9, no. 3, pp. 3995–4005, Jun. 2020.

[24] N. K. Zakaria, "ASD children gait classification based on principal component analysis and linear discriminant analysis," *Int. J. Emerg. Trends Eng. Res.*, vol. 8, no. 6, pp. 2438–2445, Jun. 2020.

[25] B. Henderson, P. Yogarajah, B. Gardiner, M. McGinnity, K. Forster, B. Nicholas, D. Wimpory, and J. Wanigasinghe, "Effects of intra-subject variation in gait analysis on ASD classification performance in machine learning models," in *Proc. 31st Irish Signals Syst. Conf. (ISSC)*, Jun. 2020, pp. 1–6.

[26] S. Pandya, S. Jain, and J. P. Verma, "AI based classification for autism spectrum disorder detection using video analysis," in *Proc. IEEE Bombay Sect. Signature Conf. (IBSSC)*, Dec. 2022, pp. 1–6.

[27] N. Zhang, M. Ruan, S. Wang, L. Paul, and X. Li, "Discriminative few shot learning of facial dynamics in interview videos for autism trait classification," *IEEE Trans. Affect. Comput.*, vol. 14, no. 2, pp. 1110–1124, Apr./Jun. 2022.

[28] J. Han and B. Bhanu, "Individual recognition using gait energy image," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 2, pp. 316–322, Feb. 2006.

[29] S. Albawi, T. A. Mohammed, and S. Al-Zawi, "Understanding of a convolutional neural network," in *Proc. Int. Conf. Eng. Technol. (ICET)*, Aug. 2017, pp. 1–6.

[30] A. Elkholy, Y. Makihara, W. Gomaa, M. A. Rahman Ahad, and Y. Yagi, "Unsupervised GEI-based gait disorders detection from different views," in *Proc. 41st Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2019, pp. 5423–5426.

[31] M. Babaee, L. Li, and G. Rigoll, "Person identification from partial gait cycle using fully convolutional neural networks," *Neurocomputing*, vol. 338, pp. 116–125, Apr. 2019.

[32] T. T. Verlekar, P. Lobato Correia, and L. D. Soares, "Using transfer learning for classification of gait pathologies," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Dec. 2018, pp. 2376–2381.

[33] T. Verlekar, L. Soares, and P. Correia, "Automatic classification of gait impairments using a markerless 2D video-based system," *Sensors*, vol. 18, no. 9, pp. 1–16, 2018.

[34] S. Sivapalan, D. Chen, S. Denman, S. Sridharan, and C. Fookes, "Gait energy volumes and frontal gait recognition using depth images," in *Proc. Int. Joint Conf. Biometrics (IJCB)*, Oct. 2011, pp. 1–6.

[35] C. Wang, J. Zhang, J. Pu, X. Yuan, and L. Wang, "Chrono-gait image: A novel temporal template for gait recognition," in *Proc. Eur. Conf. Comput. Vis.*, in Lecture Notes in Computer Science: Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics, vol. 6311, 2010, pp. 257–270.

[36] A. AbdulRahman, I. Hadi, and Y. Rajihy, "Generating 3D dataset of gait and full body movement of children with Autism spectrum disorders collected by Kinect v2 camera," *COMPUSOFT, Int. J. Adv. Comput. Technol.*, vol. 9, no. 8, pp. 3791–3797, 2020. [Online]. Available: https://ijact.in/index.php/ijact/article/view/1193

[37] C. Z. C. Hasan, R. Jailani, and N. M. Tahir, "Use of statistical approaches and artificial neural networks to identify gait deviations in children with autism spectrum disorder," *Int. J. Biol. Biomed. Eng.*, vol. 11, pp. 74–79, 2017.

[38] R. Rivest, *The MD5 Message-Digest Algorithm*, document RFC 1321, RFC Editor, 1992. [Online]. Available: https://www.rfc-editor.org/info/rfc1321

[39] S. Ilias, N. M. Tahir, and R. Jailani, "Feature extraction of autism gait data using principal component analysis and linear discriminant analysis," in *Proc. IEEE Ind. Electron. Appl. Conf. (IEACon)*, Nov. 2016, pp. 275–279.
[40] S. Ilias, N. M. Tahir, R. Jailani, and C. Z. C. Hasan, "Classification of autism children gait patterns using neural network and support vector machine," in *Proc. IEEE Symp. Comput. Appl. Ind. Electron. (ISCAIE)*, May 2016, pp. 52–56.
[41] L. Sadouk, T. Gadi, and E. H. Essoufi, "A novel deep learning approach for recognizing stereotypical motor movements within and across subjects on the autism spectrum disorder," *Comput. Intell. Neurosci.*, vol. 2018, pp. 1–16, Jul. 2018, Art. no. 7186762, doi: 10.1155/2018/7186762.
[42] Y. Tian, X. Min, G. Zhai, and Z. Gao, "Video-based early ASD detection via temporal pyramid networks," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2019, pp. 272–277.
[43] S. Rahman, S. Rahman, O. Shahid, M. T. Abdullah, and J. A. Sourov, "Classifying eye-tracking data using saliency maps; classifying eye-tracking data using saliency maps," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Jan. 2021, pp. 9288–9295.
[44] Y. Tang, G. Tong, X. Xiong, C. Zhang, H. Zhang, and Y. Yang, "Multi-site diagnostic classification of autism spectrum disorder using adversarial deep learning on resting-state fMRI," *Biomed. Signal Process. Control*, vol. 85, Aug. 2023, Art. no. 104892. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1746809423003257
[45] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. 3rd Int. Conf. Learn. Represent. (ICLR)*, 2015, pp. 1–14.

B. HENDERSON received the B.Eng. degree in computer science from Queen's University, Belfast, U.K., in 2019. He is currently pursuing the Ph.D. degree with Ulster University, Londonderry, U.K. His research interests include computer vision, artificial intelligence, and robotics.

PRATHEEPAN YOGARAJAH received the degree (Hons.) in computer science from the University of Jaffna, Jaffna, Sri Lanka, in 2001, the M.Phil. degree in computer vision from Oxford Brookes University, Oxford, U.K., in 2006, and the Ph.D. degree in computing and engineering from Ulster University, Londonderry, U.K., in 2015. Currently, he is a Lecturer in computer science with the School of Computing and Intelligent Systems, Ulster University. His research interests include biometrics, computer vision, image processing, steganography and digital watermarking, robotics, and machine learning. He received the Oxford Brookes University HMGCC Scholarship Award, in 2005, co-received the Proof of Principle Award from Ulster University, in 2012, and the Proof of Concept from Invest Northern Ireland, in 2013.

BRYAN GARDINER (Member, IEEE) received the degree (Hons.) in electronics and computer systems, in 2006, and the Ph.D. degree from Ulster University, Northern Ireland, in 2010. He is currently a Senior Lecturer and the Associate Head of the School of Computing, Engineering and Intelligent Systems, Ulster University. An active member of the Intelligent Systems Research Centre, Cognitive Robotics Team. His research interests include computer vision, data analytics, and mobile robotics. His research is currently supported by funding from MRC, HSCNI, and Interreg NWE. He has undertaken a number of technology commercialization and knowledge exchange projects funded via Innovate U.K. and Intertrade Ireland. He has served as a reviewer for numerous international conferences and journals. He is a member of the IEEE UKRI Society and IEEE Signal Processing Society (SPS) and an affiliate member of Computational Imaging SIG, Irish Pattern Recognition & Classification Society, International Association of Pattern Recognition, and British Machine Vision Association.

T. MARTIN MCGINNITY (Senior Member, IEEE) received the B.Sc. degree (Hons.) in physics from the New University of Ulster, in 1975, and the Ph.D. degree from the University of Durham, U.K., in 1979. Currently, he is an emeritus Professor with the School of Computing, Engineering and Intelligent Systems, Ulster University (UU). Before taking semi-retirement in 2018, he was a formerly Pro Vice Chancellor and the Head of the College of Science and Technology, Nottingham Trent University (NTU), the Dean of Science and Technology with NTU, and the founding Director of the Intelligent Systems Research Centre, UU, where he was also the Head of the School of Computing and Intelligent Systems. He is the author or coauthor of over 300 research papers, has supervised over 30 Ph.D. student to successful completion and attracted over £40 million in research funding. His research interests include computational intelligence, computational neuroscience, modeling of biological information processing in FPGA reconfigurable hardware, and sensory systems in cognitive robotics.