

## Article

# The Walk of Guilt: Multimodal Deception Detection from Nonverbal Motion Behaviour

Sharifa Alghowinem <sup>1,2,3,\*</sup> , Sabrina Caldwell <sup>2</sup> , Ibrahim Radwan <sup>4</sup> , Michael Wagner <sup>2</sup> and Tom Gedeon <sup>2,5,6</sup> 

<sup>1</sup> Media Lab, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

<sup>2</sup> Research School of Computer Science, Australian National University, Canberra 2600, Australia; sabrina.caldwell@anu.edu.au (S.C.); michael.wagner@canberra.edu.au (M.W.); tom@cs.anu.edu.au (T.G.)

<sup>3</sup> College of Computer and Information Science, Prince Sultan University, Riyadh 11586, Saudi Arabia

<sup>4</sup> Human-Centred Computing Laboratory (HCC Lab), University of Canberra, Canberra 2617, Australia; ibrahim.radwan@canberra.edu.au

<sup>5</sup> Faculty of Science and Engineering, Curtin University, Perth 6102, Australia

<sup>6</sup> John von Neumann Faculty of Informatics, Obuda University, 1034 Budapest, Hungary

\* Correspondence: sharifah@mit.edu

**Abstract:** Detecting deceptive behaviour for surveillance and border protection is critical for a country's security. With the advancement of technology in relation to sensors and artificial intelligence, recognising deceptive behaviour could be performed automatically. Following the success of affective computing in emotion recognition from verbal and nonverbal cues, we aim to apply a similar concept for deception detection. Recognising deceptive behaviour has been attempted; however, only a few studies have analysed this behaviour from gait and body movement. This research involves a multimodal approach for deception detection from gait, where we fuse features extracted from body movement behaviours from a video signal, acoustic features from walking steps from an audio signal, and the dynamics of walking movement using an accelerometer sensor. Using the video recording of walking from the Whodunnit deception dataset, which contains 49 subjects performing scenarios that elicit deceptive behaviour, we conduct multimodal two-category (guilty/not guilty) subject-independent classification. The classification results obtained reached an accuracy of up to 88% through feature fusion, with an average of 60% from both single and multimodal signals. Analysing body movement using single modality showed that the visual signal had the highest performance followed by the accelerometer and acoustic signals. Several fusion techniques were explored, including early, late, and hybrid fusion, where hybrid fusion not only achieved the highest classification results, but also increased the confidence of the results. Moreover, using a systematic framework for selecting the most distinguishing features of guilty gait behaviour, we were able to interpret the performance of our models. From these baseline results, we can conclude that pattern recognition techniques could help in characterising deceptive behaviour, where future work will focus on exploring the tuning and enhancement of the results and techniques.

**Keywords:** deception detection; body pose; nonverbal behaviour; motion analysis; multimodal fusion



Academic Editor: António Joaquim da Silva Teixeira, Nuno Almeida and Samuel Silva

Received: 22 August 2024

Revised: 13 December 2024

Accepted: 20 December 2024

Published: 26 December 2024

**Citation:** Alghowinem, S.; Caldwell, S.; Radwan, I.; Wagner, M.; Gedeon, T. The Walk of Guilt: Multimodal Deception Detection from Nonverbal Motion Behaviour. *Information* **2025**, *16*, 6. <https://doi.org/10.3390/info16010006>

**Copyright:** © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

Deception is a set of actions and behaviours that aim to mislead or conceal the truth from others. When deception is detected, consequences can occur that raise the stakes for the deceptive person, which influences their behaviour [1]. Deception detection has been a concern for several entities besides the police. For example, detecting deception

for border control officers could protect a country from terrorists, non-genuine tourists, and drug smugglers. Moreover, shoplifted stolen products amount to \$100 billion U.S. dollars worldwide [2], where detecting such acts could help reduce the burden on retailers, police, and the economy.

Several studies have attempted to automatically detect deception from different channels including physiological, verbal, and nonverbal cues. These studies detect different types of deception, such as overt deception in verbal statements of mistruth. These techniques have had some success rate for lie detection because lying is associated with increased stress, anxiety, and cognitive load. However, very few studies have investigated motion behaviour such as walking as a nonverbal cue of deception. Given the importance of detecting deception from a distance using non-invasive sensors (e.g., surveillance cameras), research in automatic deception recognition should explore this area. Even with the limited number of studies analysing body movement, they showed success in recognising general emotion as surveyed in [3].

This research is motivated to follow these successful examples in the context of deception detection from body movement, namely, gait. We investigate automatic modelling of gait patterns associated with deception using non-invasive sensors (audio and video), as well as an accelerometer sensor, by extracting and analysing a wide range of behavioural gait features. The novelty of this research is as follows:

- Studies on deception recognition from verbal and nonverbal cues, in general, are limited. Given the importance of this analysis in customs control and surveillance, this work is an attempt to enrich the literature in this field.
- This study extracts and analyses a novel and comprehensive feature set of body movement and gait inspired by psychology and behavioural expression through the literature on dancing.
- We are the first to analyse acoustic features from gait step sounds for deception.
- A comprehensive set of features from an accelerometer sensor in the context of deception is analysed.
- We also investigate a multimodal/fusion approach of the gait signals (audio, video, and accelerometer).
- Finally, we provide a detailed interpretation of the model and the behavioural gait features that are strongly associated with deception behaviour for future analysis and confirmations.

In our approach, we draw on the principles outlined in our previous work [4], where we explored the interpretation of depression detection models through feature selection techniques. Our framework emphasizes the power of feature selection methods in enhancing not only the model's performance, but also its interpretability. Specifically, it highlights how these techniques can distill vast datasets into the most informative features, thereby supporting more transparent and understandable models. We utilized a framework that employed stability measures in different levels to ensure that only the most relevant features were selected.

In its first level, the framework employs multiple feature selection algorithms to isolate and retain the most informative features from high-dimensional data. By applying an ensemble of feature selection methodologies, the framework ensures a comprehensive evaluation of each feature's significance.

Then, to address the challenge of feature selection stability, the framework includes stability measures that evaluate the consistency of selected features across different randomized runs of the algorithms. Two key stability metrics are utilized: Jaccard Similarity Index (JI), where a higher JI indicates greater consistency and reliability in feature selection, and Between Thresholds Stability (BTS), which assesses the stability of feature

importance rankings across thresholds. By analyzing how the importance of features varies with different selection criteria, BTS provides insights into the robustness of feature relevance determinations.

The integration of these measures into the feature selection process ensures that selected features not only contribute to model accuracy but also enhance the interpretability of the model's outputs. This is particularly valuable for affective computing applications where understanding the underlying mechanisms of prediction is as important as the predictions themselves.

By adapting similar methodologies, our current study not only focuses on the detection of deception from gait but also emphasizes a systematic interpretation of the selected features to elucidate the underlying behaviors indicative of deceptive motion. This approach aligns with our goal of achieving both accuracy and explainability in the context of the Whodunnit deception dataset.

## 2. Related Work

Based on decades of research, channels used in deception and lie detection can be divided into three main areas of investigation. Physiological signals, such as heart rate, respiratory rate, skin conductivity, and blood pressure, are one of the main investigated channels in the last decade to detect deception [5]. Since such measures need contact sensors, systems that use them cannot be deployed in a public setting. Speech analysis has also been investigated for this purpose, where researchers hypothesise that there are certain behaviours in speech that differentiate liars from truth-tellers. Even though modelling for lie detection from speech showed interesting results [6], the option for speech signals might not be present in certain situations (e.g., shoplifting), or not acquired from the relevant sensors (e.g., video-only surveillance cameras). Nonverbal cues, from facial, hand, and body gestures, have been investigated recently for deception detection, where nervous behaviour is reported as one of the main indicators [7]. For example, eye blinking, trembling of the leg or foot, more trunk movements, and more position shifts were found to be signs of lying in [8]. Most of the work carried out on lie and deception detection from verbal and nonverbal cues focuses on interview settings, where speech prosody, facial expression, and body gestures are present, and where only a handful of research is available for overall body movement such as gait [7]. Given its non-invasive nature, and its suitability for public settings, this work focuses on analysing and modelling the behavioural pattern of overall body movement to detect deception.

### 2.1. Automatic Deception Detection from Body Movement

To model deception behaviour, a labelled dataset for deception versus truthful behaviours is used to train the models. Such datasets are very limited and have a variety of contexts. One of the widely used datasets for modelling deception is a real trial dataset that was collected from television and public videos and then labelled based on the conviction [9]. Another realistic (not scenario-based) dataset was recently collected, where participants could voluntarily choose to describe an actual image they saw on the screen or anything else [10]. However, these datasets show a frontal view of the upper body of the participants, which is not suitable for full body analysis. Other datasets include a scenario where participants are asked to steal \$20 and then discuss selected topics [11], a werewolf game play [12], and lying about a video content task [13]. Studies using these datasets focused on facial expression, speech prosody, and physiological signals (see our survey paper for full review).

To the best of our knowledge, there are only two studies that analysed body motion to detect deception, which are [14,15]. In [14], a full body motion capture suit was used to

record the location, velocity, and orientation of 23 body points. A total of 90 subjects were involved in a scenario, where they either told the truth or a lie about playing a game and stealing currency, then completed a self-report on their emotional experience. Absolute movement was measured as a feature, which is the mean value of pose differences over time. This was measured over the whole body, arms, legs, and head, after a normalisation procedure. The statistical analysis results showed that the movements were guilt-related, which occurred independently of other emotions. Automatic binary classification using logistic regression of being truthful or deceptive reached a performance of 82%.

In [15], a new dataset was collected, where participants were asked to hand an object (a roll of money for the deception condition) to a person. Overall, 589 walking videos labelled with deception and natural walks were used for modelling deception. Several features were extracted such as body posture and body movement, as well as manually annotated gestures such as fidgeting, along with the deep features extracted from a deep learning network (namely LSTM) using the 3D pose of the body. These features were fed into convolutional and fully connected layers to classify deception, where they achieved an accuracy of 88%.

## 2.2. Automatic Affect Detection from Body Movement

Emotions are also expressed nonverbally through body movement as surveyed in [3,16–19]. Affect expressive movement is categorised into four types: communicative (e.g., gestures), functional (e.g., walking), artistic (e.g., choreography), and abstract (e.g., arm lifting), where a single or a combination of these types represents affect [3]. For example, anxiety is linked to expanded limbs and torso, fear is linked to bent elbows, and shame is linked to bowed trunk and head [16]. In computational movement detection, actions are detected through body models, image models, and spatial statistics, where body grammar, template, and temporal statistics are extracted [20]. In [17], a focus on low-level and high-level features, as well as features from coding systems for body movement emotion recognition was given. Efforts have been made to create a consensus-based reliable coding system, where two of the main body coding systems are Body Action and Posture (BAP) and Labal Movement Analysis (LMA).

BAP is a micro-description of body movement proposed by [21], where the body movements are described on an anatomical level, a form level, and a functional level. Body behaviours and actions are described through body parts and joint movement, as well as location and orientation. Automatic coding and annotation of BAP were proposed in [22], where a full body motion tracking suit is used. However, such a sensor suit is not suitable for remote body behaviour analysis. To the best of our knowledge, there is no automatic remote recognition of BAP coding.

Dancing motion was coded by Laban, and used for illustrating and recognising performers' emotion from their body motion [23]. In their analysis, eight cameras were used to capture the acted dance movements. Laban divides human motion into four components: body, effort, shape, and space. In their study, some of the Laban features that are automatically calculable were selected for emotion classification, where the results reached up to 98% in multi-class (discrete) emotional states. Continuous emotion recognition in theatre performance using LMA was proposed in [24], where a Microsoft Kinect device was used for motion capture. The initial results were promising, and showed potential for advanced emotion recognition using LMA.

Besides the body coding systems described above, static and dynamic features and their combinations were used for affect recognition from body behaviour [19]. These features could also be represented as geometrical and appearance features. Using dynamic features, body posture extracted features proved to be valuable in recognising acted

emotions both as uni-modal and when combined with other modalities [25]. Statistical measures were calculated from the dynamic computer vision features such as Quantity of Motion (QoM), Contraction Index (CI), and fluidity, velocity, and acceleration, where the uni-modal classification results were higher than speech and facial expression classification results. A 3D joint Euler rotation was recorded in [26] for affective posture recognition, where the automatic recognition achieved comparable results to the human observers.

Gait patterns, in particular, were analysed for recognising certain emotion expressions [27–29]. Raw data and processed features from motion capture, video cameras, and Kinect were deployed for automatic emotion recognition from gait analysis as surveyed in [18], where the results vary between studies given the variations of the studies' goals and the emotions recognised. However, to the best of our knowledge, no study analysed gait and body movement for the automatic detection of deception or guilt. Since guilt could be considered an affect, we aim to follow the above literature to analyse the video recordings of the Whodunnit dataset (see Section 3.1) to extract body movement features for detecting it.

Gait acoustics were analysed for the purpose of strike detection, and are therefore used in recognising human actions (e.g., walking, running). Such recognition could be used for diagnosing gait- and posture-related pathologies, treatments, and corrections. Moreover, it could be used for person identification through their gait patterns [30]. The use of gait acoustics is preferable due to its non-invasive approach, inexpensive sensor, and the ability to analyse these acoustics remotely for security and surveillance. The acoustics gait profile proposed in [31] showed reliable results in characterising gait patterns to be used for clinical and biometric applications. Since acoustics from gait can be sensitive to variations in clothes, footwear, and floor surfaces, ref. [32] developed robust time difference features to account for such variations. A multimodal gait dataset with several footwear variations was collected and analysed in [33,34], where microphone signals were processed, performing an average recall of 99%.

Wearable sensors that measure motion using accelerometers have been used not only in movement recognition as surveyed in [35], but also in recognising emotion from movement (e.g., [36]). In [37], body movements were categorised based on their general characteristics (e.g., light, strong, or sustained walking). Several statistical features were extracted from three sensors attached in different body locations, where the wrist location performed the highest in the movement recognition, with a classification average of 85%. Using the accelerometer embedded in mobile phones, ref. [38] attempted to recognise daily activities, where a comprehensive list of statistical features were extracted and the classification accuracy reached 86% using ANN. On the other hand, ref. [36] focused on emotion recognition based on accelerometer movement. They showed emotion elicitation clips to participants, then asked them to walk for one minute while wearing a smart bracelet with a built-in accelerometer. Features from temporal, frequency, and temporal–frequency domains were extracted, then fed to an SVM, where the classification accuracy reached up to 91% in two-category classification. Inspired by their findings, we aim to recognise guilt from gait accelerometer signals.

### 3. Method

The data collected and used for this work and the process for guilty behaviour analysis are described in the following sections. Moreover, the process of guilt detection is summarised in Figure 1.

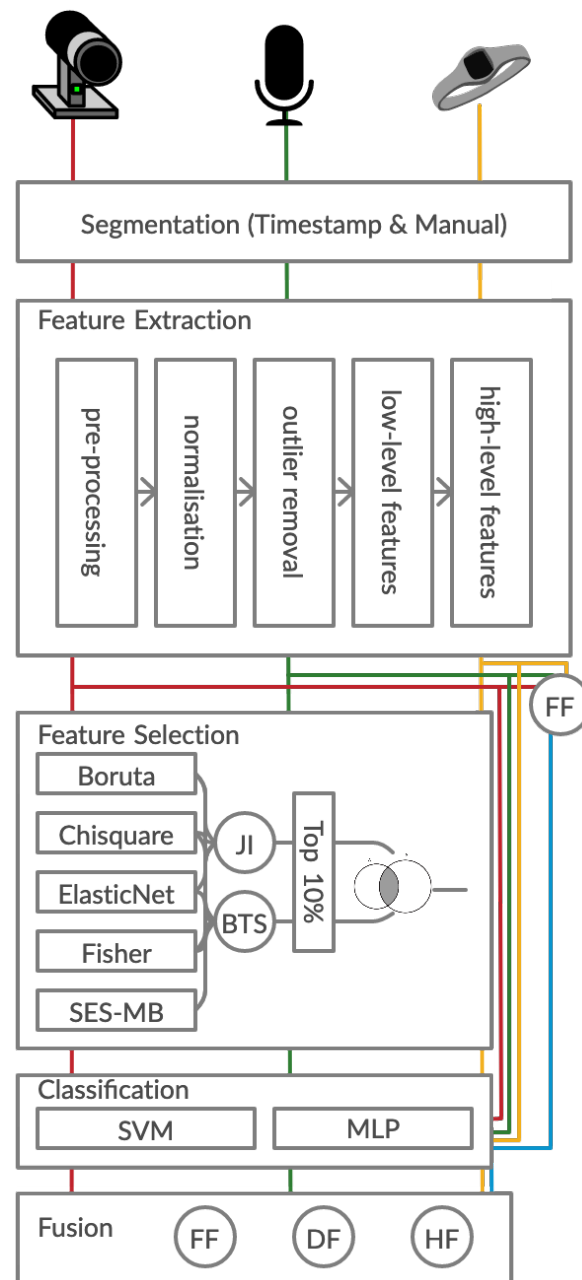


**Figure 1.** Summary of the guilty behaviour detection from walking.

### 3.1. Dataset Collection Procedure

In this work, we use a segment of a larger collected dataset named “Whodunnit”, which aimed to induce deceptive behaviour. The dataset collected responses in terms of physiological, facial, vocal, etc., from different sensors. Subjects were led to believe that they were participating in a job interview experiment, where bio-signals were recorded for analysis. Subjects were asked to follow instructions that were provided verbally from the research team or written in paper-form cards. The data collection procedure can be divided into 7 steps, where each step is executed in a different location and involves different sensor devices for recording. Deception scenarios include different levels of guilt and innocence in a mock crime situation, where subjects found written instructions to either steal, consider stealing, or not steal a phone (Step #3).

One of the dataset collection steps is to walk down and up the stairs to sign an interview attendance sheet (Step #5). During this step, the participants’ body action and

movement are captured using four webcam cameras located on several corners of the stairs for this purpose (see Figure 2).

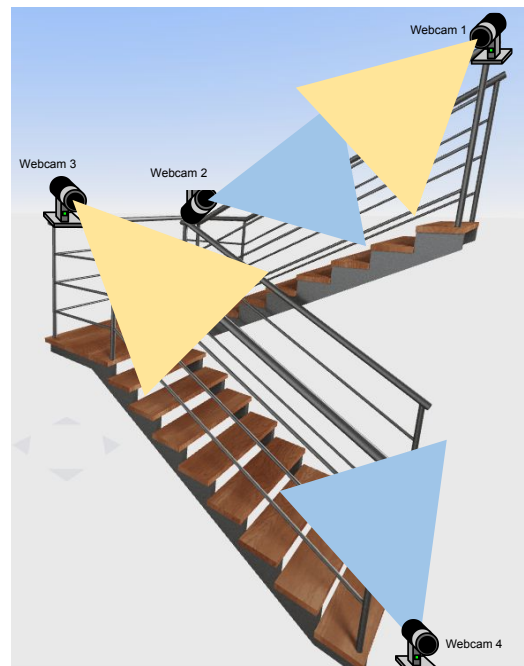


**Figure 2.** Camera positions during participant movement (blue triangle indicates angle direction upward of camera view; yellow indicates angle direction downward of camera view).

Microsoft LifeCam Cinema HD cameras, Microsoft Corporation, Washington, DC, USA (<https://www.microsoft.com/accessories/en-au/products/webcams/lifecam-cinema/h5d-00016>, accessed on 1 January 2017) were used for this purpose to provide high-quality recordings. The resolution of the recording was  $640 \times 360$  pixels with a frequency of 30 frames per second with three RGB channels. Moreover, physiological signals were recorded using an E4 wristband, Empatica Inc., Boston, MA, USA (<https://www.empatica.com/research/e4/>, accessed on 1 January 2017), and audio was recorded using a RODE smartLav+ lapel microphone, RØDE Microphones, Sydney, Australia (<http://www.ode.com/microphones/lavalier>, accessed on 1 January 2017). The E4 wristband was equipped with sensors to acquire different biosignals, such as Blood Volume Pulse, a 3-axis accelerometer, electrodermal activity, skin temperature, and electromyographic muscle activity, etc., with a sampling rate of 32 kHz. Since this work focuses on motion, only the accelerometer signal is analysed. Audio was recorded during the entire interaction with the subject using the RODE smartLav+, RØDE Microphones, Sydney, Australia, a lapel microphone with a connection to an iPad, for portability and recording storage. The sampling rate for the audio recording was 11 kHz. The microphone was attached to the participant's clothes around the sternum, and the iPad was placed in a pouch bag placed on the participant's waist. In this work, we analyse the acoustics of the steps produced by the participant during walking through the stairs to investigate any acoustic characteristics that might be associated with deception behaviour.

The dataset contains recordings from 49 subjects, 26 males and 23 females, with ages ranging between 19 and 43 ( $\mu = 26.4, \sigma = 6.1$ ). The scenario distribution was as follows: 17 subjects instructed not to steal the phone (innocence), 16 instructed to steal the phone (guilty), and 16 instructed to consider stealing but deciding not to steal the phone. In the current work, we only consider the innocent and guilty subjects for analysis and classification, which results in 33 subjects included in this study.

### 3.2. Feature Extraction

Since the devices used for the data collection were activated for recording at different times, and since they were not linked with each other, the time for each step in each session was logged manually for later synchronisation with the other devices. For the physiological sensor, each frame of the recorded signal was time-stamped, and the time-stamp was mapped with the manual time logged for the start and the end of the walking step. Webcam video files were also segmented for the four cameras and checked for alignments using the recording time-stamp with the manual one. For audio recording, initial segmentation was performed using the time-stamps. However, due to audio noise caused by the paper flipping and signing the attendance-sheet, a finer manual segmentation was performed. Only walking acoustics from going down and up the stairs were kept, so that the part of signing the paper was removed. As with any real-life data collection, a few files were missing or corrupted. Table 1 shows the available files after segmentation for each step. Once the segmentation of each signal was complete, signal preparation for feature extraction and analysis began. Since the signals were different, these procedures differed for each of them, which is detailed in the below sections and summarised in Table 2.

**Table 1.** Number of valid and invalid segmented files of each device used in this work.

| Modality               | Device    | Valid Files | Invalid Files |
|------------------------|-----------|-------------|---------------|
| Physiological Response | E4        | 46          | 3             |
| Acoustics              | Lapel Mic | 31          | 18            |
|                        | WebCam 1  | 38          | 11            |
| Body Actions           | WebCam 2  | 44          | 5             |
|                        | WebCam 3  | 43          | 6             |
|                        | WebCam 4  | 37          | 12            |

**Table 2.** Summary of extracted features from each modality and their descriptions.

| Feature Group              | Descriptions and Calculation                                                                                                                      |
|----------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>Body Movement</b>       |                                                                                                                                                   |
| Trunk lean angle           | The angle between the sternum and the collarbone (clavicle) line and the origin. The angle should indicate trunk leaning left and right.          |
| Elbows horizontal movement | The distance between the elbow and the body side, which is calculated from the cross point of the elbow and the shoulder (left and right elbows). |
| Elbows vertical movement   | The distance between the elbow and the shoulder line (left and right elbows).                                                                     |
| Hands to hip distance      | The distance between the hand and the hip points (left and right hands).                                                                          |
| Face touching              | The distance between the hand and the head points (left and right hands).                                                                         |
| Neck touching              | The distance between the hand and the neck points (left and right hands).                                                                         |
| Holding arm                | The line cross between one forearm to the other arm (left and right arms).                                                                        |
| Crossed arms               | Indicating if both arms are holding each other.                                                                                                   |
| Shoulders angle            | The angle between the left and right shoulder line and the origin.                                                                                |
| Arms symmetric             | Using the sternum points as the centre, the symmetric measure of the left and right elbows is calculated.                                         |
| Elbow articulation         | The angle between the elbow, the shoulder, and the body side (left and right arms).                                                               |
| Knee bend                  | The angle between the hip, knee, and ankle points (left & right knees).                                                                           |
| Leg movement               | The distance between the sternum and the knee points (left & right legs).                                                                         |
| Foot to hip distance       | The distance between the ankle and the hip points (left and right feet).                                                                          |
| Hands to shoulder distance | The distance between the hand and the shoulder points (left and right hands).                                                                     |
| Hands distance             | The distance between the left and right hands.                                                                                                    |
| Gait size                  | The distance between the left and right ankles.                                                                                                   |
| Body volume                | The area of a polygon from the outer body points (i.e., sternum, collarbone, and neck points are not included).                                   |
| Upper body volume          | The area of a polygon from the outer upper body points (i.e., arms and head points).                                                              |
| Lower body volume          | The area of a polygon from the outer lower body points (i.e., legs and sternum points).                                                           |
| Left body volume           | The area of a polygon from the outer left body points (i.e., left leg, left hand, head, and sternum points).                                      |
| Right body volume          | The area of a polygon from the outer right body points.                                                                                           |

Table 2. Cont.

| Feature Group                   | Descriptions and Calculation                                                                                  |
|---------------------------------|---------------------------------------------------------------------------------------------------------------|
| <b>Step Acoustics</b>           |                                                                                                               |
| Discrete Wavelet Transform      | For audio signal transformation to analyse the temporal and spectral properties from non-speech signals.      |
| Teager energy operator          | The non-linear transform of the time-domain signal that measures the harmonics produced from the sound wave.  |
| Energy (power)                  | The signal power using root-mean-square and log functions of the wave signal.                                 |
| MFCC                            | Signal Cepstral analysis using mel-frequency cepstral coefficients.                                           |
| Frequency variability           | Using jitter, which measures the interference with normal signal.                                             |
| Amplitude variability           | Using shimmer, which measures variability in the signal amplitude in comparison to the fundamental frequency. |
| Magnitude of the signal         | measuring sound level using intensity and loudness.                                                           |
| <b>Accelerometer Sensor</b>     |                                                                                                               |
| Accelerometer movement          | The X, Y, and Z of the sensor location in the space.                                                          |
| Magnitude of the accelerometer  | The square root of the X, Y, and Z squares.                                                                   |
| Velocity of movement            | The change in accelerometer movement and its magnitude from one frame to another.                             |
| Acceleration of movement        | The change in velocity from one frame to another.                                                             |
| Magnitude frequency             | The fast fourier transform of the magnitude signal.                                                           |
| Magnitude logarithmic frequency | The logarithmic function of the fast fourier transform of the magnitude signal.                               |
| Magnitude amplitude             | Shifting the magnitude frequency component to the centre of spectrum.                                         |

### 3.2.1. Body Movement Features

As a preparation for the analysis, the video recordings from the webcams located at the staircase were segmented based on the start and end time of step #5. Each camera contains segments of both frontal and back views of the participants' walk, as each participant is asked to go down the stairs and back up to continue the study, which was also segmented for analysis. Pose and body joint detection from the back view is tricky, since most algorithms use face detection to align the body joints. To tackle this issue, Radwan et al. proposed a generative adversarial-based framework [39]. This method estimates the body part positions in two steps based on the deformation level of these body parts. Parts with less deformation are estimated first (parents) and then the parts with large deformations (children). Their proposed method implicitly learns the hierarchy between these two levels using a hierarchy-aware objective function. We have employed this approach to estimate the body pose of the participants, with both frontal and back poses being estimated accurately (see Figure 3).

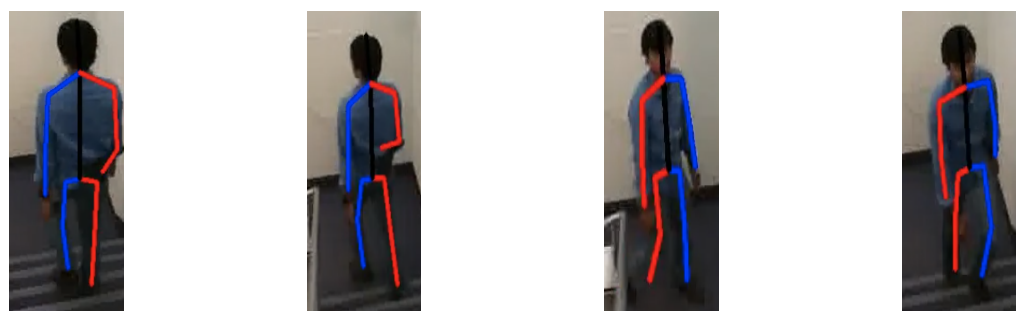


**Figure 3.** Sample of body joints' localisation while walking the stairs (red lines relate to the right side of the body and the blue ones to the left side).

The raw data of the located 16 body joints are used to extract features that are then used for motion behaviour analysis. However, before the features are extracted, normalisation techniques should be considered to account for both features within and between participants' variability. That is, the distance between the participant's body and the camera while moving is variable (during approaching and moving away from the camera), as well as the differences between participants in body size (height, shoulder size, etc.). Normalisation ensures reliable measures of the extracted features with comparability for

analysis and classification. In this work, we use the distance between the sternum and the collarbone (clavicle) points to normalise the distance between the other points (their distance is divided by the distance between the two given points). We select these two points for the normalisation since they are rigid, which make them robust from continuous, sudden, and skewed movements. Normalisation is performed in each frame, then outlier detection using Grubbs' test for outliers [40] is used to remove the frames with skewed measures (e.g., erroneous joint location).

As mentioned earlier, BAP and LMA are two of the main body coding systems, which suggest specific features that can be used for body motion analysis. Inspired by their approach, we extracted several features that aligned with our objectives and were feasible to extract from our data as listed in Table 2. Some of the features suggested by BAP need the 3D location of joints to be extracted; however, we focus the extracted features on the 2D features since our joint locations are in 2D. Laban body analysis aimed at illustrating and recognising emotions from dancing movements; we extracted some of the LMA features to investigate the feasibility of detecting deception from body movement. We extracted static and dynamic body features as listed in Table 2 from each frame (low level). Then, we extracted high-level functional (statistical) features to measure the temporal change over time.

### 3.2.2. Steps Acoustics Features

Inspired by the acoustic gait analysis literature, we aimed to analyse the gait acoustic features as an additional modality for guilt detection. Audio signals from the walking scenario part of the Whodunnit dataset were segmented to separate the other scenarios of the dataset. Within the walking segments, the part where the subject stops to sign the attendance sheet was removed from the acoustic analysis in order to evaluate gait only and remove any acoustic noise from pen and paper flipping.

The segmented audio signal then is analysed in three different forms: normal, Discrete Wavelet Transform (DWT), and Teager energy operator (TEO). DWT has been successful in non-speech signal analysis (e.g., music beat), since it retains the temporal information in addition to the frequency [41]. TEO, proposed by [42] to extract energy from a signal, is advantageous for its ability in suppressing signal noise; it has been utilised in many applications including audio denoising, heart pulse detection, and music beat analysis. We believe that DWT and TEO are applicable in our gait analysis, since finding gait strokes and phases while illuminating noise from clothing is critical. From these three different forms, we extract the step acoustic features as described in Table 2, where both time-based and frequency-based features are analysed. These features are extracted at a low level (per frame), where the frame size is set to 25 ms at a shift of 10 ms while using a Hamming window. Similar to the body movement feature extraction, high-level functional (statistical) features were extracted from the low-level features for temporal analysis. Moreover, first and second derivatives were calculated from the low-level features, and functional features were also applied to these derivative features.

### 3.2.3. Accelerometer Sensor Features

Hand position and movement patterns vary in different emotional contexts [43]; therefore, we aimed to analyse the speed and motion pattern of the hand while walking to investigate these features' ability in detecting deception and guilt behaviour expression from our dataset. The accelerometer sensor, as mentioned earlier, records timestamps for each frame, where this timestamp is used to segment the recording to the start and end time of the walking step. In this work, we only use the location point of the accelerometer sensor, since we focus on the motion pattern of the hand. The location points are normalised

using z-normalisation for each participant's segment, to reduce the variability between participants heights.

$$\mathbf{a}_{norm} = \frac{\mathbf{a} - \mu}{\sigma}$$

where  $\mathbf{a}$  represents the accelerometer data,  $\mu$  is the mean, and  $\sigma$  is the standard deviation. Then, similar to body movement features, outlier values are detected using Grubbs' test [40] and then removed. As detailed in Table 2, we extract low-level features from each frame, and then extract temporal features over time.

The high-level temporal features from all the low-level features are statistical functions over a window of time. Several window sizes were empirically tested with a step of 5 frames from 15 frames to all frames. A full segment is 20 s long on average for each direction (going up or down). However, the results from these tests were under the chance level and did not significantly contribute to our findings. Therefore, we focused on other parameters that showed a more meaningful impact on the final outcomes. Given the differences in each modality signal, different high-level temporal features were extracted. For body movement, we extracted 10 temporal features from each of the per-frame features, which were the minimum, maximum, average, standard deviation, variance, range, skewness, kurtosis, number of peaks, and number of valleys. In addition, we extracted two more temporal features, which are (1) the overall body volume for each of the body volume features (left, right, etc.), and (2) body movement jerks, which is the variance of body movement regarding the overall body movement variance. From the acoustic features and their derivatives, we extracted six functional features from each of the low-level ones, which were the minimum, maximum, average, standard deviation, variance, and range. For the accelerometer low-level features, we extracted 18 functional features, which were the following: the minimum, maximum, average, median, standard deviation, variance, range, skewness, kurtosis, number of peaks, number of valleys, root mean square, entropy, interquartile range, zero-crossing, mean absolute deviation, median absolute deviation, and trapezoidal numerical integration.

### 3.3. Dimensionality Reduction

Given the high dimensionality of the feature space from all modalities, we tested feature selection methods to reduce the feature space. In this work, we followed our previously developed framework for feature selection, as described in [4]. In that study, we conducted an extensive ablation analysis to demonstrate the individual contribution of each feature selection algorithm compared to the combined framework outcome. Given that the scope of this paper is to apply the framework for deception detection from gait, we focused on reporting the overall final results obtained using the complete framework. For detailed contributions of each feature selection algorithm and the mathematical formulation using this framework in a different context, please refer to [4]. Using the feature selection framework, we applied five different feature selection algorithms with different categories on our feature space. We focus on supervised feature selection techniques, which aim at finding a small and representative feature set that differentiates classes from each other by removing redundant and non-discriminative features of the classes. There are several categorisations for feature selection methods, such as filter, wrapper, embedded, and data-structure methods. The feature selection categories and algorithms used in this work are the following: (1) Boruta [44] is one of the wrapper methods; (2) Chi square [45] is a statistical-based filter method; (3) Elastic Nets [46] is an embedded algorithm; (4) Fisher score [47] is a similarity-based filter method; and (5) state-of-the-art statistically equivalent signature Markov Blanket (SES-MB) that learns a network structure from input features [48]. While the first four methods score the feature based on their relevance in identifying the class, the SES-MB finds the best representative set for a target class.

To aggregate the features from these five methods, we follow the approach described in [4], which weighs each feature using two stability measures (Jaccard similarity index (JI) and Between Thresholds Stability (BTS)) before aggregating the top features from several ensembles, and then intersect these top features for the final selection. Even though dimensionality reduction techniques focus on filtering redundant and weak features (in relation to the classification task), our framework utilises these techniques for listing the selected features for interpretation. Therefore, we not only use these selected features for the classification, but also we inspect and analyse them for interpretation. In this work, we use 100 ensembles in each of the feature selection methods and select the top 10% features from each modality.

### 3.4. Classification

To detect deception from walking behaviour, we utilise two classification algorithms in a binary (i.e., guilty/non-guilty) subject-independent scenario. To mitigate the effect of the limited amount of data, leave-one-subject-out cross-validation was used in all the classifiers without any overlap between training and testing data. To measure the performance, we used the weighted accuracy and Matthews correlation coefficient (MCC). MCC showed reliable results for classification performance, as it produces high scores only if all elements in the confusion matrix have good results [49].

Even though deep learning approaches show a high performance, they require a huge dataset for training if trained from scratch, which we did not have. Transfer learning from a relevant task could solve the small dataset issue; however, gait analysis in general is understudied and we could not acquire enough data for a transfer learning approach, which we are working on for future work. To get a performance baseline of deception detection from gait, we used two traditional classifiers: Support Vector Machines (SVMs) and a Multilayer Perceptron Neural Network (MLP).

Support Vector Machines (SVMs) have only a few parameters to be tuned, balancing between simplicity and accuracy. The SVM kernel selected in this work is the Radial Basis Function (RBF). Optimization for the cost parameter  $C$  and the gamma parameter  $\gamma$  is performed using a wide-range grid search.

**Non-linear SVM hypothesis:**

$$h_{\theta}(x) = \begin{cases} 1 & \text{if } \theta^T f \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

**Cost Function:**

$$J(\theta) = C \left[ \sum_{i=1}^m y^{(i)} \text{Cost}_1(\theta^T f^{(i)}) + (1 - y^{(i)}) \text{Cost}_0(\theta^T f^{(i)}) \right]$$

**Gaussian RBF:**

$$\exp\left\{-\gamma\|\mathbf{x} - \mathbf{x}'\|^2\right\}$$

**Parameter Grid Search:**

$$C \in \{-80, -79, \dots, 79, 80\}$$

$$\gamma \in \{-80, -79, \dots, 79, 80\}$$

The optimization process involves a wide to narrow search steps, ensuring that the optimal values for  $C$  and  $\gamma$  are accurately determined.

For MLP implementation, we used two hidden layers, and we searched for the number of perceptrons that produce the highest performance on the training set. For the first layer, the search range was from 5 to 50 perceptrons, while the second layer was from 2 to 10 perceptrons. Moreover, we used stochastic gradient descent as the solver function, a ReLU as activation function for the hidden layers, and the MCC as the cost function.

### Multi-Layer Perceptron Network

$$f(x) = \left( \sum_{i=1}^m w_i \cdot x_i \right) + b$$

where  $m$  is the number of neurons in the previous layer,  $w$  is a random weight,  $x$  is the input value, and  $b$  is a random bias.

### Search Range for Perceptrons:

For the first hidden layer:

$$\text{Number of Perceptrons} \in \{5, 6, \dots, 49, 50\}$$

For the second hidden layer:

$$\text{Number of Perceptrons} \in \{2, 3, \dots, 9, 10\}$$

### Activation Function for Hidden Layers:

$$\text{ReLU}(x) = \max(0, x)$$

The input to these classifiers is either the full feature set in the modality in question (or the fused ones), or the top 10% of features selected by the feature selection framework described above.

### 3.5. Multimodal Fusion

Since there are three different sensors for analysing guilty gait behaviour, we follow the framework proposed in [50] to investigate multimodal fusion. We compare feature fusion (FF), decision fusion (DF), and hybrid fusion (HF). Feature fusion stacks all features from different modalities into one huge vector for each sample, therefore requiring that features from all different modalities exist. Features from multiple modalities are concatenated into a single vector  $\mathbf{z}$ :

$$\mathbf{z} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_m]$$

where  $\mathbf{x}_i$  are feature vectors from each modality  $i$ . However, since our real-world data have missing samples from some modalities, not all subjects have all modalities available in their sample. To overcome the missing samples from some sensors, we compare the performance of subjects who have full samples only, with the performance of subjects with missing samples evaluated by filling the missing samples with zeros. Decision fusion, on the other hand, fuses the results at the classifier decision level. This makes it easier to fuse subjects with missing modality samples. We use majority voting for the final fusion decision in a parallel and hierarchical way. That is, the parallel decision fusion aggregates the classification results from each individual modality in one level, treating the four webcams as individual modalities, while the hierarchical approach aggregates the results from the webcams first on one level, and then the results are fused on a second level with the acoustic and accelerometer modalities.

We also investigate hybrid fusion in two methods. One method of hybrid fusion is to combine the advantages of feature and decision fusions, while reducing their disadvantages. In this method, the individual modalities are fused in feature-level fusion then the new

feature-fused modality is treated as an individual modality for decision fusion. The second way of carrying out hybrid fusion used in this work is to combine the results from the different classifiers (SVMs and MLP) in a decision-level manner. This method not only leverages the advantages of the two classifiers, but also increases the reliability of the final results.

While the fusion process involves some computational overhead, the statistical nature of the features extracted and used in this work significantly reduced the resources needed for training. This made the inference phase feasible for real-time scenarios, although there were demands in extracting and processing these features initially. We utilized the NCI national Australian GPU cluster for these experiments, with training taking an average of 30 min per experiment.

### 3.6. Statistical Analysis

Since our original feature space was high-dimensional, statistical tests on the features could not be conducted with reliable correction. However, our feature selection approach, while reducing the dimensional to the top 10%, could allow for reliable statistical analysis to give insights into the characteristics of the behavioural patterns of guilty walks. Since our analysis was carried out in a binary manner, a two-sample two-tailed *t*-test was used. The two-tailed *t*-tests assume unequal variances with significance  $p = 0.05$ . The direction of the *t*-test was also analyzed.

## 4. Results

As illustrated above (see Figure 1), the raw signals from different devices underwent several processes (e.g., normalisation) to extract the final set of features. These extracted features, with and without the selection process, which aimed to filter the most representative features for this task, were used for the classification task, and the results are shown in Table 3. Furthermore, the selected features are analysed for the interpretation of the characteristics of guilty walk behaviour; the filtered features are listed in Figure 4.

**Table 3.** Classification results using different modalities—single or fused.

|                                            |        | # Samples | SVM           |      |          |      | MLP           |      |             |             |
|--------------------------------------------|--------|-----------|---------------|------|----------|------|---------------|------|-------------|-------------|
|                                            |        |           | Full Features |      | SF (10%) |      | Full Features |      | SF (10%)    |             |
|                                            |        |           | Acc           | MCC  | Acc      | MCC  | Acc           | MCC  | Acc         | MCC         |
| Single Modality                            |        |           |               |      |          |      |               |      |             |             |
| Modality                                   | Device |           |               |      |          |      |               |      |             |             |
| Vision                                     | Cam 1  | 24        | 62.5          | 0.38 | 58.3     | 0.30 | 54.2          | 0.21 | 58.3        | 0.30        |
|                                            | Cam 2  | 28        | 53.6          | 0.19 | 53.6     | 0.19 | 71.4          | 0.43 | 57.1        | 0.16        |
|                                            | Cam 3  | 28        | 53.6          | 0.12 | 53.6     | 0.19 | 64.3          | 0.29 | 57.1        | 0.28        |
|                                            | Cam 4  | 23        | 52.2          | 0.02 | 60.9     | 0.32 | 56.5          | 0.21 | 52.2        | 0.20        |
| Acoustics                                  | Mic    | 20        | 55.0          | 0.23 | 55.0     | 0.23 | 55.0          | 0.23 | 60.0        | 0.33        |
| Acceleration                               | E4     | 32        | 53.1          | 0.18 | 53.1     | 0.18 | 65.6          | 0.35 | 59.4        | 0.26        |
| Feature Fusion                             |        |           |               |      |          |      |               |      |             |             |
| Subjects with full modalities              |        | 8         | 75.0          | 0.49 | 75.0     | 0.49 | 75.0          | 0.47 | <b>87.5</b> | <b>0.75</b> |
| Subjects with at least 1 cam and all other |        | 17        | 58.8          | 0.18 | 58.8     | 0.27 | 70.6          | 0.45 | 58.8        | 0.34        |
| Subjects with at least 1 modality          |        | 33        | 51.5          | 0.17 | 51.5     | 0.17 | 54.5          | 0.18 | 54.5        | 0.17        |
| Decision Fusion                            |        |           |               |      |          |      |               |      |             |             |
| All modalities (raw)                       |        | 33        | 51.5          | 0.17 | 51.5     | 0.17 | 72.7          | 0.46 | 57.6        | 0.26        |
| All cams only                              |        | 32        | 56.7          | 0.27 | 56.7     | 0.27 | 53.3          | 0.11 | 50.0        | 0.00        |
| All modalities (hierarchical)              |        | 33        | 54.5          | 0.25 | 54.5     | 0.25 | 63.6          | 0.27 | 51.5        | 0.00        |

Table 3. Cont.

|                                                       | # Samples | SVM           |      |          |      | MLP           |      |          |      |
|-------------------------------------------------------|-----------|---------------|------|----------|------|---------------|------|----------|------|
|                                                       |           | Full Features |      | SF (10%) |      | Full Features |      | SF (10%) |      |
|                                                       |           | Acc           | MCC  | Acc      | MCC  | Acc           | MCC  | Acc      | MCC  |
| Hybrid Fusion—Feature Fusion + Single Modalities      |           |               |      |          |      |               |      |          |      |
| Subjects with full modalities                         | 33        | 51.5          | 0.17 | 51.5     | 0.17 | 78.8          | 0.58 | 57.6     | 0.26 |
| Subjects with at least 1 cam and all other            | 33        | 54.5          | 0.25 | 54.5     | 0.25 | 69.7          | 0.44 | 51.5     | 0.00 |
| Subjects with at least 1 modality                     | 33        | 51.5          | 0.17 | 51.5     | 0.17 | 60.6          | 0.33 | 63.6     | 0.38 |
| Hybrid Fusion—Classifiers Decision Fusion (SVM + MLP) |           |               |      |          |      |               |      |          |      |
| All modalities (raw)                                  | 33        | 60.6          | 0.36 | 72.7     | 0.55 | -             | -    | -        | -    |
| All cams only                                         | 32        | 73.3          | 0.47 | 66.7     | 0.39 | -             | -    | -        | -    |
| All modalities (hierarchical)                         | 33        | 57.6          | 0.31 | 72.7     | 0.55 | -             | -    | -        | -    |

**SF**: subset of selected features; **Acc**: Balanced Weighted Accuracy; **MCC**: Matthews correlation coefficient. The **bold** values indicate highest performance.

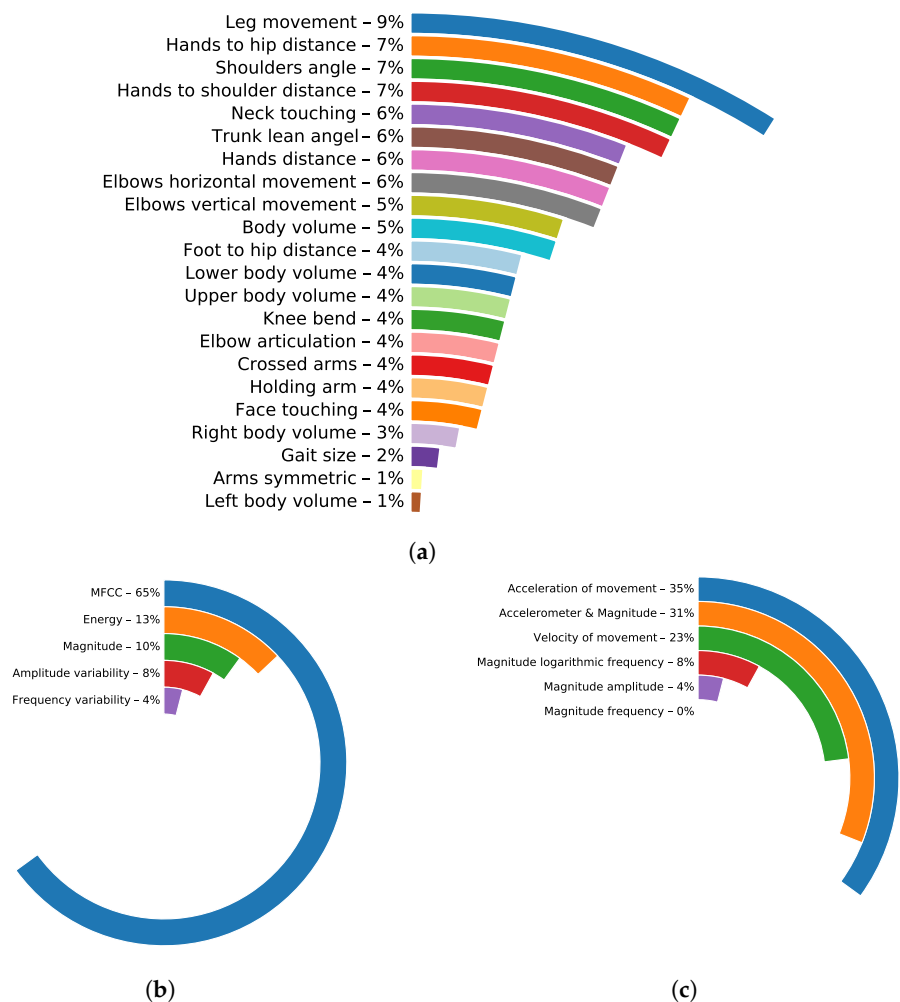


**Figure 4.** Interpretation of the selected features from each modality. (a) Top body movement features. (b) Top step acoustics features. (c) Top accelerometer sensor features.

#### 4.1. Classification of Guilty Walks

In this classification problem, we investigate detecting guilty/deceptive behaviour from walking (see Table 3). To this end, we compare the classification results from walking behaviour extracted from videos, step acoustics, and accelerometers for single modality classification, as well as combined modalities fused at the feature level, decision level, and

hybrid level. We also compare two different classification methods (SVM and MLP) with and without feature selection.

Comparing single modality classification (vision, acoustics, and acceleration), the highest classification result is from the second webcam (Cam 2), followed by the acceleration, while the acoustics generated the lowest classification results. While each camera records the participant while both going up and down, in a preliminary experiment, we classified the behaviour of escalating and de-escalating the stairs separately. The results showed no differences in classifying behaviour based on the direction of movement, which gives an indication that our pose estimation is robust to the direction of movement. We conducted the same preliminary experiment on both acoustics and acceleration signals, and the classification results were similar. The results presented here are the aggregation of the features extracted from both the escalating and de-escalating behaviours. Looking at the body movement features, Cam 2 had the highest classification results (from MLP—full features). Given that all camera signals are processed similarly, with similar lighting condition, and the inclusion of both escalating and de-escalating walk behaviour, it is not clear why Cam 2 has an advantage over the other cameras for its high classification results. However, since Cam 2 captured the participant's full front body just after they potentially felt deceptive, the footage might have captured the onset of the participant's feeling. This strategic placement allowed Cam 2 to effectively capture the onset of guilty emotions. This insight may help in optimizing camera placement and orientation for similar future experiments or deployments.

To ensure that the classification results were consistent, where randomness was eliminated, we conducted several fusion approaches. Feature-level fusion extends the feature vector for each sample to include all features; however, it expects no missing samples. Since some of the samples are missing, we deal with this in different ways. First, we fuse and classify the features from the subjects where all sensors signals exist (eight subjects only), which achieved the highest results. Then, to increase the sample size, we fused the features of at least one camera with the acoustic and acceleration features (17 subjects only), where the results from MLP—full features were similar to the 8 subjects' classification. This indicates the importance and the contribution of each modality to the final results. Finally, we fuse the features from all three modalities, while filling in the missing samples with some values. For an exploratory phase, we experimented with replacing the missing samples with zeros, the average of other subject's features, and random values. The classification results from this exploratory phase showed higher classification results when using zero values, which is presented in this table. Nonetheless, the classification results from such filling in this fusion approach did not perform beyond the chance level.

Decision-level fusion is not as sensitive as feature-level fusion to missing samples when we conducted several decision-level fusion experiments. First, we fused the classification decision (using majority voting of existing decisions) of all modalities, while treating each camera decision as an individual modality, which resulted in the best classification results from the decision level. Then, we conduct a two-step (hierarchical) decision fusion, by first fusing the vision decisions then fusing the results with the acoustics and acceleration decisions. The results from fusing the vision decisions in the MLP classification were at random chance (based on MCC results) and significantly reduced from the individual modality classification results. On the other hand, the same fusion approach on the SVM showed a stable average improvement. This could indicate an overfitting in the individual modalities with MLP, which the decision fusion exposed. Nonetheless, such instability could be eliminated with advanced fusion techniques, as explored below. The two-step (hierarchical) decision fusion improved the SVM results but not the MLP results compared

to the one-step (raw) decision fusion, which was expected as a result of the first-step results discussed earlier.

We also investigated two approaches to hybrid-level fusion. First, we explored hybrid fusion where feature-level fusion is treated as a modality to be fused in a decision-level with the single modalities (before the feature-level fusion). Since this approach relies on feature-level fusion, we follow the same experiments as in the feature fusion. First, using the feature fusion where all samples exist as an extra modality, where only eight subjects benefited from this approach, showed the highest classification compared to the results under this fusion group. Similarly, in the other two methods, even though the results did not improve from the feature-level and decision-level fusions, the results were more robust to overfitting and randomness.

Finally, we explored a second hybrid fusion approach, which fused the decision results from the two classifiers. Even though the MLP results were superior to the SVM results in most cases with and without the feature selection, SVM showed stability and robustness to the randomness in the classification results. Moreover, SVM classification results from using only 10% of the full feature set were similar to when using the full feature set. This is significant, given the huge reduction in the feature space, and shows the stability of the SVM classifier. MLP classification results dropped significantly in most cases when using the selected features, while performing higher classification compared to SVM. These observations encourage hybrid-classifier-level fusion, not only to get the benefit of both classifiers, but also to improve the confidence of the final results. We follow the same path as with the decision fusion, where modalities are fused in one-step (raw) and two-step (hierarchical) levels. At this level, improvement of the classification results is not the end goal, while stable results are; improvement can occur. Since more decision results are included in this approach, it helps the majority voting to make an affirmative decision. For example, instead of having only four decisions in single classifier results, in this hybrid-fusion, we would have eight decisions from which to obtain the majority voting. As can be seen in Table 3, fusing classifiers' results when using the full feature set stabilised (averaged) between the SVM and MLP at both the raw and hierarchical levels. On the other hand, the results from the top 10% features improved significantly compared to the decision-level fusion, where such a pattern was observed in several instances with the selected features' classification.

Overall, we can see that the classification results for guilty walk behaviour are promising. When all modalities are present, the classification results are the highest. However, in real-world applications, this might not be sustainable, though we showed different approaches overcoming missing signals and modalities. Hybrid-classifier-level fusion showed not only the most stable classification results, but also above-average results, especially with the top selected features. Moreover, the results from feature selection performed similarly to when using the full feature space, allowing for the generalizability and explainability of the model performance, as discussed in the following section.

#### 4.2. Characteristics of Guilty Walks

As mentioned in Section 3.3, the top 10% features with high distinctive power for guilty walks are systematically selected using five different feature selection methods. The feature groups listed in Table 2 consist of several high-level features extracted from the signal and its derivatives, where we illustrate the percentage of features selected from each group in Figure 4 for interpretation. In other words, Figure 4 illustrates the percentage of contribution of each feature group into the final dimensionality reduction results. The contribution percentages are derived using our previously developed feature selection framework, which employs two stability measures: Jaccard similarity index (JI)

and Between Thresholds Stability (BTS). This robust process ensures that only the most stable and significant features contribute to the final model. The fact that none of the feature groups were totally eliminated from the selection process indicates the relevance of these groups to the problem in question to detect a guilty walk, and therefore supports the choice of their extraction.

For body movement modality, as shown in Figure 4a, all feature groups contributed to the final selection, where several feature groups contributed more than the others. As expected, leg movement showed the highest strength, followed by the distance of the hands and hips and shoulders. This is expected because the leg and arm movements are the most dominant features during a walk. Moreover, for a guilty walk, we would also expect subtle movements that would differentiate it from a normal walk. This was observed from touching the neck and the other hand, as well as touching the face, and arm holding and crossing, which is in line with Ekman and Friesen's work in [51].

Features extracted from the audio signal (see Figure 4b), from the unprocessed normal signal and the processed signal using the wavelet and Teager energy, demonstrated a balanced contribution from each feature group. Clearly, MFCC dominated with 65% of the final selected features, since it contains the largest number of features compared to the other groups. Even with only 10% selected features, features from other feature groups managed to penetrate the final selection. We believe that this, as with the body feature groups, is a good sign indicating the strength of these features and their interaction with each other in detecting deceit. Moreover, features selected from both processed signals and unprocessed ones contributed equally in the final selection in each feature group, which also indicates that some patterns of the step acoustic features can be found in the processed signals as well.

Finally, the percentage of the final selected features from each feature group of the accelerometer signal is illustrated in Figure 4c. The acceleration of the movement, followed by the location and its magnitude, as well as the velocity, are the most leading feature groups in distinguishing guilty movement, which is in line with [52]. Unlike body and acoustic modalities, the features under the Fast Fourier Transform of the magnitude (magnitude frequency) group were excluded from the final selected features. Nonetheless, its logarithmic feature group survived the selection process, which could indicate that processing the signal revealed patterns that are more relevant to the task.

The feature selection process was performed on each modality separately to show the strongest features in the modality in identifying guilty walks. However, we believe that executing the process on fused modalities will show the interaction between features from different modalities, as described in [4]. In this work, we could not perform this analysis because of the missing samples from some of the modalities where the current feature selection methods were sensitive to missing information, but this should be investigated in the future.

## 5. Conclusions

As an attempt to detect deception behaviour from a distance and to fill the gap in this research area, we investigated the recognition of guilty walks using non-invasive sensors (e.g., camera) that could be used in applications such as surveillance. We used a subset of the Whodunnit deception dataset, where a full body view was recorded at different angles using four cameras, step sound was recorded by a microphone, and movement dynamics were recorded using an accelerometer. From these three modalities, we extracted and analysed a wide range of behavioural gait features, which were used for multimodal classification and also were interpreted using a systematic framework for feature selection. The classification results showed high potential in detecting deceit from gait using different methods of

fusion. Even though each of the individual modality performances was moderate, using multi-modal fusion not only boosted the accuracy, but also the confidence and reliability of the final results. We believe that the fusion methods, especially the hybrid fusion method, reduced the effect of overfitting of some of the individual models. Moreover, we performed a feature selection framework to select 10% of the features systematically. Even with the huge reduction of the feature space using this method, the classification results using selected features were similar if not better than when using the full feature space. This is interesting since it also allows for interpretation of the features and the modelling results. Our interpretation using the features selected using the framework showed a strong correlation between not only the body movement, but also neck, face, and hand touching, which are associated with deception. Using teager energy and wavelet of the audio signal showed an equal contribution to detecting guilty walks compared to using the normal signal.

Future work should investigate modelling from the 3D estimation of multi-view cameras, where advanced body movement and posture features can be analysed. Moreover, a comparison of handcrafted features and deep features in modelling deception could also give insights into such modelling.

In this study, certain parameter configurations, such as varying window sizes, did not yield successful results and therefore were not included in the main findings. Understanding why specific parameter adjustments do not enhance model performance is crucial for both refining existing methodologies and informing future research endeavors. Therefore, we advocate for further investigation into diverse parameter settings and their impact on model effectiveness.

In conclusion, our findings hold significant potential for enhancing multimodal interaction systems by providing deeper insights into user behavior through implicit cues. For instance, in interactive security systems, understanding deception-related behaviors like guilty walks could augment user context data, leading to more nuanced responses to potential threats. This framework's ability to distill implicit interactions into actionable insights allows systems to adapt dynamically based on user behavior patterns detected in real time. Integrating these capabilities with existing interaction modalities can lead to more seamless and intuitive user experiences, broadening the scope of human–computer interactions beyond explicit commands.

**Author Contributions:** Conceptualization, S.A. and T.G.; data curation, S.C.; formal analysis, S.A.; funding acquisition, T.G.; investigation, S.A.; methodology, S.A., I.R. and M.W.; project administration, S.A.; resources, T.G.; software, S.A. and I.R.; supervision, T.G.; validation, S.A.; visualization, S.A.; writing—original draft, S.A.; writing—review and editing, S.C. and I.R. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** The study was conducted in accordance with the Declaration of Helsinki, and approved by the Institutional Review Board of Australian National University.

**Informed Consent Statement:** Informed consent was obtained from all subjects involved in the study.

**Data Availability Statement:** The data presented in this study are available on request from the corresponding author.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## References

- Mann, S.; Vrij, A.; Bull, R. Suspects, lies, and videotape: An analysis of authentic high-stake liars. *Law Hum. Behav.* **2002**, *26*, 365–376. [CrossRef] [PubMed]
- National Association for Shoplifting Prevention (NASP). Available online: <http://www.shopliftingprevention.org/> (accessed on 9 April 2019).
- Karg, M.; Samadani, A.A.; Gorbet, R.; Kühnlenz, K.; Hoey, J.; Kulić, D. Body movements for affective expression: A survey of automatic recognition and generation. *IEEE Trans. Affect. Comput.* **2013**, *4*, 341–359. [CrossRef]
- Alghowinem, S.; Gedeon, T.; Goecke, R.; Cohn, J.; Parker, G. Interpretation of Depression Detection Models via Feature Selection Methods. *IEEE Trans. Affect. Comput.* **2020**, *14*, 133–152. [CrossRef] [PubMed]
- Cardona, P.A.N. A compendium of pattern recognition techniques in face, speech and lie detection. *Int. J. Res. Rev. Appl. Sci.* **2015**, *24*, 108.
- Nasri, H.; Ouarda, W.; Alimi, A.M. ReLiDSS: Novel lie detection system from speech signal. In Proceedings of the 2016 IEEE/ACS 13th International Conference of Computer Systems and Applications (AICCSA), Agadir, Morocco, 29 November–2 December 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 1–8.
- Vrij, A.; Fisher, R.P. Lying and nervous behaviours Unravelling the misconception about deception and nervous behaviour. *Front. Psychol.* **2020**, *11*, 1377. [CrossRef]
- Vrij, A.; Mann, S.A.; Fisher, R.P.; Leal, S.; Milne, R.; Bull, R. Increasing cognitive load to facilitate lie detection: The benefit of recalling an event in reverse order. *Law Hum. Behav.* **2008**, *32*, 253–265. [CrossRef]
- Pérez-Rosas, V.; Abouelenien, M.; Mihalcea, R.; Burzo, M. Deception Detection Using Real-Life Trial Data. In Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, ICMI '15, Seattle, WA, USA, 9–13 November 2015; pp. 59–66. [CrossRef]
- Gupta, V.; Agarwal, M.; Arora, M.; Chakraborty, T.; Singh, R.; Vatsa, M. Bag-of-lies: A multimodal dataset for deception detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Long Beach, CA, USA, 16–17 June 2019.
- Pérez-Rosas, V.; Mihalcea, R.; Narvaez, A.; Burzo, M. A Multimodal Dataset for Deception Detection. In Proceedings of the LREC, Reykjavik, Iceland, 26–31 May 2014; pp. 3118–3122.
- Raiman, N.; Hung, H.; Englebienné, G. Move, and i will tell you who you are: Detecting deceptive roles in low-quality data. In Proceedings of the 13th International Conference on Multimodal Interfaces, Alicante, Spain, 14–18 November 2011; pp. 201–204.
- Diana, B.; Elia, M.; Zurloni, V.; Elia, A.; Maisto, A.; Pelosi, S. Multimodal Deception Detection: A t-pattern Approach. In Proceedings of the 2015 ACM on Workshop on Multimodal Deception Detection, Seattle, WA, USA, 13 November 2015; pp. 21–28.
- Van Der Zee, S.; Poppe, R.; Taylor, P.; Anderson, R. To freeze or not to freeze: A motion-capture approach to detecting deceit. In Proceedings of the Hawaii International Conference on System Sciences, Kauai, HI, USA, 5–8 January 2015.
- Randhavane, T.; Bhattacharya, U.; Kapsaskis, K.; Gray, K.; Bera, A.; Manocha, D. The Liar's Walk: Detecting Deception with Gait and Gesture. *arXiv* **2019**, arXiv:1912.06874.
- Kleinsmith, A.; Bianchi-Berthouze, N. Affective body expression perception and recognition: A survey. *IEEE Trans. Affect. Comput.* **2013**, *4*, 15–33. [CrossRef]
- Zacharatos, H.; Gatzoulis, C.; Chrysanthou, Y.L. Automatic emotion recognition based on body movement analysis: A survey. *IEEE Comput. Graph. Appl.* **2014**, *34*, 35–45. [CrossRef]
- Stephens-Fripp, B.; Naghdy, F.; Stirling, D.; Naghdy, G. Automatic affect perception based on body gait and posture: A survey. *Int. J. Soc. Robot.* **2017**, *9*, 617–641. [CrossRef]
- Noroozi, F.; Kaminska, D.; Corneanu, C.; Sapinski, T.; Escalera, S.; Anbarjafari, G. Survey on emotional body gesture recognition. *IEEE Trans. Affect. Comput.* **2018**, *12*, 505–523. [CrossRef]
- Weinland, D.; Ronfard, R.; Boyer, E. A survey of vision-based methods for action representation, segmentation and recognition. *Comput. Vis. Image Underst.* **2011**, *115*, 224–241. [CrossRef]
- Dael, N.; Mortillaro, M.; Scherer, K.R. The body action and posture coding system (BAP): Development and reliability. *J. Nonverbal Behav.* **2012**, *36*, 97–121. [CrossRef]
- Velloso, E.; Bulling, A.; Gellersen, H. AutoBAP: Automatic coding of body action and posture units from wearable sensors. In Proceedings of the 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction, Geneva, Switzerland, 2–5 September 2013; IEEE: Piscataway, NJ, USA, 2013; pp. 135–140.
- Aristidou, A.; Charalambous, P.; Chrysanthou, Y. Emotion analysis and classification: Understanding the performers' emotions using the LMA entities. *Comput. Graph. Forum* **2015**, *34*, 262–276. [CrossRef]
- Senecal, S.; Cuel, L.; Aristidou, A.; Magnenat-Thalmann, N. Continuous body emotion recognition system during theater performances. *Comput. Animat. Virtual Worlds* **2016**, *27*, 311–320. [CrossRef]
- Kessous, L.; Castellano, G.; Caridakis, G. Multimodal emotion recognition in speech-based interaction using facial expression, body gesture and acoustic analysis. *J. Multimodal User Interfaces* **2010**, *3*, 33–48. [CrossRef]

26. Kleinsmith, A.; Bianchi-Berthouze, N.; Steed, A. Automatic recognition of non-acted affective postures. *IEEE Trans. Syst. Man Cybern. Part B (Cybern.)* **2011**, *41*, 1027–1038. [[CrossRef](#)]
27. Montepare, J.; Koff, E.; Zaitchik, D.; Albert, M. The use of body movements and gestures as cues to emotions in younger and older adults. *J. Nonverbal Behav.* **1999**, *23*, 133–152. [[CrossRef](#)]
28. Karg, M.; Kuhnlenz, K.; Buss, M. Recognition of affect based on gait patterns. *IEEE Trans. Syst. Man Cybern. Part B (Cybern.)* **2010**, *40*, 1050–1061. [[CrossRef](#)]
29. Michalak, J.; Troje, N.F.; Fischer, J.; Vollmar, P.; Heidenreich, T.; Schulte, D. Embodiment of sadness and depression—Gait patterns associated with dysphoric mood. *Psychosom. Med.* **2009**, *71*, 580–587. [[CrossRef](#)]
30. Wan, C.; Wang, L.; Phoha, V.V. A survey on gait recognition. *ACM Comput. Surv. (CSUR)* **2018**, *51*, 89. [[CrossRef](#)]
31. Altaf, M.U.B.; Butko, T.; Juang, B.H.F. Acoustic gaits: Gait analysis with footstep sounds. *IEEE Trans. Biomed. Eng.* **2015**, *62*, 2001–2011. [[CrossRef](#)] [[PubMed](#)]
32. Huang, J.; Di Troia, F.; Stamp, M. Acoustic Gait Analysis using Support Vector Machines. In Proceedings of the ICISSP, Funchal, Portugal, 22–24 January 2018; pp. 545–552.
33. Wang, C.; Wang, X.; Long, Z.; Yuan, J.; Qian, Y.; Li, J. Estimation of temporal gait parameters using a wearable microphone-sensor-based system. *Sensors* **2016**, *16*, 2167. [[CrossRef](#)] [[PubMed](#)]
34. Wang, C.; Wang, X.; Long, Z.; Yuan, J.; Qian, Y.; Li, J. Multimodal gait analysis based on wearable inertial and microphone sensors. In Proceedings of the 2017 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (Smart-World/SCALCOM/UIC/ATC/CBDCom/IOP/SCI), San Francisco, CA, USA, 4–8 August 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 1–8.
35. Lara, O.D.; Labrador, M.A. A survey on human activity recognition using wearable sensors. *IEEE Commun. Surv. Tutor.* **2013**, *15*, 1192–1209. [[CrossRef](#)]
36. Zhang, Z.; Song, Y.; Cui, L.; Liu, X.; Zhu, T. Emotion recognition based on customized smart bracelet with built-in accelerometer. *PeerJ* **2016**, *4*, e2258. [[CrossRef](#)]
37. Kikhia, B.; Gomez, M.; Jiménez, L.; Hallberg, J.; Karvonen, N.; Synnes, K. Analyzing body movements within the laban effort framework using a single accelerometer. *Sensors* **2014**, *14*, 5725–5741. [[CrossRef](#)]
38. Pires, I.M.; Garcia, N.M.; Pombo, N.; Flórez-Revuelta, F.; Spinsante, S. Pattern recognition techniques for the identification of Activities of Daily Living using mobile device accelerometer. *arXiv* **2017**, arXiv:1711.00096. [[CrossRef](#)]
39. Radwan, I.; Moustafa, N.; Keating, B.; Choo, K.K.R.; Goecke, R. Hierarchical Adversarial Network for Human Pose Estimation. *IEEE Access* **2019**, *7*, 103619–103628. [[CrossRef](#)]
40. Grubbs, F.E. Procedures for detecting outlying observations in samples. *Technometrics* **1969**, *11*, 1–21. [[CrossRef](#)]
41. Tzanetakis, G.; Essl, G.; Cook, P. Audio analysis using the discrete wavelet transform. In Proceedings of the WSES International Conference Acoustics and Music: Theory and Applications (AMTA 2001), Skiathos, Greece, 26–30 September 2001; Volume 66.
42. Kaiser, J.F. On a simple algorithm to calculate the ‘energy’ of a signal. In Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, Albuquerque, NM, USA, 3–6 April 1990; IEEE: Piscataway, NJ, USA, 1990; pp. 381–384.
43. AlKadhi, B.; Alghowinem, S. An Interacting Decision Support System to Determine a Group-Member’s Role Using Automatic Behaviour Analysis. In Proceedings of the Intelligent Computing, Chongqing, China, 6–8 December 2019; Arai, K., Kapoor, S., Bhatia, R., Eds.; Springer: Cham, Switzerland, 2019; pp. 455–470.
44. Kursa, M.; Rudnicki, W. Feature selection with the boruta package. *J. Stat. Softw.* **2010**, *36*, 1–13. [[CrossRef](#)]
45. Liu, H.; Setiono, R. Chi2: Feature selection and discretization of numeric attributes. In Proceedings of the International Conference on Tools with Artificial Intelligence, Herndon, VA, USA, 5–8 November 1995; pp. 388–391.
46. Zou, H.; Hastie, T. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* **2005**, *67*, 301–320. [[CrossRef](#)]
47. Duda, R.O.; Hart, P.E.; Stork, D.G. *Pattern Classification*, 2nd ed.; Wiley-Interscience: New York, NY, USA, 2000.
48. Aliferis, C.F.; Statnikov, A.; Tsamardinos, I.; Mani, S.; Koutsoukos, X.D. Local Causal and Markov Blanket Induction for Causal Discovery and Feature Selection for Classification Part I: Algorithms and Empirical Evaluation. *J. Mach. Learn. Res.* **2010**, *11*, 171–234.
49. Zou, H.; Jakovlić, I.; Chen, R.; Zhang, D.; Zhang, J.; Li, W.X.; Wang, G.T. The complete mitochondrial genome of parasitic nematode *Camallanus cotti*: Extreme discontinuity in the rate of mitogenomic architecture evolution within the Chromadorea class. *BMC Genom.* **2017**, *18*, 840. [[CrossRef](#)]
50. Alghowinem, S.; Goecke, R.; Wagner, M.; Epps, J.; Hyett, M.; Parker, G.; Breakspear, M. Multimodal Depression Detection: Fusion Analysis of Paralinguistic, Head Pose and Eye Gaze Behaviors. *IEEE Trans. Affect. Comput.* **2018**, *9*, 478–490. [[CrossRef](#)]

51. Ekman, P.; Friesen, W.V. Nonverbal leakage and clues to deception. *Psychiatry* **1969**, *32*, 88–106. [[CrossRef](#)]
52. Meservy, T.O.; Jensen, M.L.; Kruse, J.; Burgoon, J.K.; Nunamaker, J.F.; Twitchell, D.P.; Tsechpenakis, G.; Metaxas, D.N. Deception detection through automatic, unobtrusive analysis of nonverbal behavior. *IEEE Intell. Syst.* **2005**, *20*, 36–43. [[CrossRef](#)]

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.