INTL JOURNAL OF ELECTRONICS AND TELECOMMUNICATIONS, 2025, VOL. 71, NO. 2, PP. 545-553
Manuscript received February 3, 2025; revised May, 2025. doi: 10.24425/ijet.2025.153603

# Biometrics gait system based on motion sensors embedded in a mobile phone: a case study for a two-day training set

Aleksander Sawicki, and Khalid Saeed

*Abstract*—**This paper presents the results of a study on developing a gait biometrics system based on motion sensors (an accelerometer and gyroscope), embedded in a smartphone. The experiments were conducted using a publicly available 13-person data corpus, with subjects participating in three data collection sessions. The study used CNN, CNN with attention and Multi-Input CNN neural networks. The training scenario from the first day resulted in an accuracy of 0.66 F1 score, 0.71 F1 score for training with the samples from the second day and 0.90 F1 score in the combined sets. It has been shown that it is more profitable to combine historical data than to update it with newer samples. Enriching the training set with a set of 30% synthetic samples produced by the LSTM-MDN generative models allowed to increase to accuracy to 0.94 F1-score. It was shown that synthetic samples can improve the generalization properties of the CNN network.**

***Keywords*—CNN; biometrics; IMU; deep learning; IMU**

## I. INTRODUCTION

The purpose of biometric systems is to identify subjects on the basis of physiological or behavioral characteristics or a combination of them. Currently, we are experiencing dynamic development of biometric solutions in both the academic and commercial sectors [1]. With the popularization and falling prices of microelectromechanical system (MEMS) based accelerometers, recent years have seen an increase in applications using motion analysis. In the field of behavioral biometrics, issues involving gait analysis are particularly exploited. The intensification of work in this area is motivated primarily by the possibility of data acquisition using mobile devices, in the form of smartphones or smartwatches [2]. An important advantage of such systems is the difficulty of intentionally imitating the gait of other subjects [3], as well as the lack of active interaction of the participant with other devices. Which is a requirement for, for example, handprint or iris data sampling [4].

Biometric motion systems can be evaluated in two basic variants that differ in the way they reflect the real life scenarios and therefore in the accuracy of the achieved results. The first type of validation is one in which both the training set (used in preparation of the decision model) and the test set were collected during the same data acquisition session. This type of validation, referred to as single-day (SD), is typically characterized by high identification efficiency, with low applicability and transferability. It is worth noting that the evaluation of biometric systems within a single day is criticized in the literature. The main objection is that repeatedly performing the same movement may be unreliable due to the phenomenon of muscle memory. The fact of the need for multiple intentional repetitions of a movement in a short interval may constitute priming [5]. Ultimately, taking into account the applicability of such a solution, it is difficult to imagine the operation of the authorization system in a scenario (i.e. one in which training and test data are taken on the same day).

The second evaluation option is cross-day (CD) validation, in which the acquisition of training and test collections takes place over two days. The described testing reflects potential real-world operating conditions, where the authorization system uses previously collected samples. Validation of this type typically achieves lower identification rates due to changing movement patterns. Manner of movement can affected by many factors among which [6]: fatigue/weakness/illness of the experiment participant; variations in the type of footwear and clothing; variations in the surface and slope of the ground on which the gait is performed; emotions.

In this study, we pay special attention to the dependability aspect of the biometric system by modifying the training sets of the decision model. In the first case, we make an attempt to increase the reliability of the prediction by increasing the training data in virtue of including samples from an additional acquisition session .

In the second case, we study a scenario in which the training set is created utilizing actual samples collected over two days as well as using synthetic samples. The second scenario aims to improve the reliability of the biometric system in a way that does not require new sampling, and therefore does not involve the sacrifice of significant additional costs. Finally, it should be noted that the experiments conducted indirectly touch on the aspect of dependability and identification accuracy.

## II. OBJECTIVES

The research work set three goals:

This work was supported by grant 2021/41/N/ST6/02505 and funded with resources for research by National Science Centre, Poland. For the purpose of Open Access, the author has applied a CC-BY public copyright license to any Author Accepted Manuscript (AAM) version arising from this submission.

All Authors are with Białystok University of Technology, Poland (e-mails: (a.sawicki@pb.edu.pl, https://orcid.org/0000-0002-8662-4484; k.saeed@pb.edu.pl, https://orcid.org/0000-0002-7741-7045 ).

- Verifying how the accuracy of a biometric system will be affected by adding supplementary samples collected during an additional acquisition day;
- Verifying whether it is more profitable to replace older historical teaching samples with newer ones, or rather to concatenate them.
- Investigate how the accuracy of the biometric system will be affected by synthetic samples generated by VAE or LSTM-MDN models

## III. Related works

In behavioral biometrics, the characteristics of the acquired samples are time-varying and dependent on the moment of acquisition. For example, biometric systems using electroencephalographic (EEG) signals use up to fifteen acquisition sessions [7]. In contrast, for systems based on wearable sensors such as an accelerometer or gyroscope, multi-session solutions are much less popular. Very often, the developed approaches learned and validated are using data from within a single day single-day validation (SD). This way of evaluating developed solutions has been criticized for lack of implementability in real-world scenarios [8]. Less common are disseminated solutions in which decision-making models are trained and tested over two days (CD validation). Studies that use data collected in the acquisition process over several days are least frequently used. It is also worth noting that the factor affecting the accuracy of the biometric system is not only the number of training sessions but also their timeline distance. In [9] it was shown that in gait biometrics, a significant decrease in the accuracy of the system is observed after a period of 9 month. The authors suggested updating/overwriting the reference samples.

For a multi-session scenario in the field of motion sensor-based behavioral biometrics, three publications are worth highlighting [8,10,11], In [8], accelerometer and gyroscope measurement signals embedded in smartwatch (smart watch) devices were used to build a biometric system. In the author's data corpus of 60 individuals, each person participated in six tracking sessions, and the entire acquisition was completed in three weeks. Participants in the experiment were asked to walk freely on a hard surface and along a set route. Th The study scenario can be considered semi-laboratory. Data were collected in blocks, during which, in addition to walking, participants were forced, for example, to stop to open doors or make several turns (this was not walking exclusively on a straight path). Within each session, data were recorded for a period of two minutes. Single samples were understood to be data processed using the using the sliding window technique. The results of feature engineering extracted vectors which were set to the input of the MLP network. In the case of SD analysis, the training set training set accounted for 60% of the available data, and the test set accounted for the remaining 40 %. In contrast, for CD validation, the data from the first day was selected as training data, and samples from the second day as test data. Despite the availability of four additional sessions, the idea of a system based on multi-day data was completely omitted. The described system was developed using only two sessions of motion tracking. Nevertheless, it should be noted that for the accelerometer Equal Error Rate (EER) of approximately 0.15 was obtained for the SD analysis and as high as 0.93 in the case of CD analysis. The published results indicate a small error for SD validation and a very large error for CD validation.

The second publication [10] used a prototype device based on a pair of Inertial Measurement Units (IMUs) in gait biometrics. Each sensor consisted of a three-axis accelerometer, gyroscope and magnetometer. The entire device consisted of two sensors placed around the left wrist and right thigh. Twenty participants took part in the study. Gait patterns were recorded over an invariant distance of 50 meters within three gait trials. The data acquisition was carried out over a period of seven days. In this case, instead of sampling using the sliding window technique, an advanced preprocessing pipeline was involved. It included gait cycle segmenting algorithm and a method for of minimizing the effect of sensor mounting on the measurement values. The approach presented here again uses feature engineering and uses the Support Vector Machine (SVM) model as a classifier. A shortcoming of the work is the unclear manner in which the validation of the developed system was performed. In the body of the publication one can read that a 10-fold cross-validation was applied. Another passage says that for CD validation, 70% of the samples from the first day and 30% of the samples from the other days were used. In addition, despite the presence of three sessions of tracking, no experiments were conducted to verify the effect of the number of training sessions on the identification results. However, the study was carried out on training covering the first day and validation using individual data from the other days. Ultimately, due to the small time interval between sessions #2 and #3, the results between CD1 and CD2 are similar to each other. For the SD validation scenario an efficiency of 0.976 was achieved, whilst for validation over two days CD1 0.896, and for CD2 0.869.

Finally, it should be emphasized that the presented approaches do not use the full capabilities of the available datasets. In the case of work [8], it could be possible to use a set of six, and in the case of work [10] of three motion tracking sessions. However, both papers indicate a decrease in accuracy for cross-day validation in comparison to SD validation.

In contrast, in our previous work [11], we conducted a proper multi-session study. A biometric system was developed using three motion tracking sessions. The open gait corpus of the Signet research group was used as the database. The dataset consisted of 14 (not 13 as is the case in the current study) subjects who performed three gait attempts over three days. Their gait was recorded using a cell phone located in the front pocket of their pants. The study also used two types of neural network architecture-CNN with attention mechanism and CNN with multiple inputs. The former decision model achieved a performance above 0.8 F1-score, while the latter achieved a above 0.85 F1-score.

However, the conducted experiments indicated a significant problem with the quality of the dataset. The performance measure for participant “6” was exactly 0, and all of its samples were labelled by the classifier as samples of participant “7” Manual inspection revealed that the samples of test subject “7” were identical to the training samples of “6”. Although the dataset was made public by a higher university, the data corpus contained a serious flaw in the form of data leakage and samples duplication. The current research is an extension of previously presented work by omitting the defective participant from the training set and increasing the number of used classifiers.

## IV. METHODOLOGY

### A. Dataset

The study used the publicly available data corpus of the SIGNET group [12]. The database contains gait recordings captured with a cell phone equipped with a three-axis accelerometer and a three-axis gyroscope. During the acquisition it was placed in the front pocket of the pants. Participants were asked to walk as close as possible to their natural gait for a period of about 5 minutes. The data corpus includes subsets of recordings of participants who took part in a single data acquisition session, as well as subsets where gait samples were collected multiple times. In the second case, it was possible to develop biometric systems that were validated in a cross-day scenario (i.e., in which the prediction of test data ran for a collection on a different day than the learning of the classifiers ), the presented approach is closer to training scenarios and is more applicable. In such conditions, the way the cell phone was placed in the pocket between days could have changed, for example, by placing the phone upside down. In addition, between sessions, participants were allowed to change clothes as well as footwear (which significantly changes the way they move). In addition, the acquisition process itself was conducted over a period of six months, the conditions for testing the biometric system were also demanding.

The corps had the unique feature of having recordings available for 13/14 participants who attended three acquisition sessions. This feature made it possible to conduct three basic tests, for which the test set of the classifier remained invariably created from day three samples. The first test involved training with data from the first day, the second scenario involved training from the second day, and the last involved training the combined data from two possible days.

The presented corpus has several significant drawbacks:

- First of all, as shown in our previous studies, the corpus contains data leakage between the participant. The test collection (day III) of participant "6" represents the training data of participant "7" (day II). This resulted in the complete identification of all test samples of "6" as "7". Therefore, participant 6 was completely omitted from the current study, limiting the number of labels to 13 participants.
- The second drawback is that there is no information on what time period between acquisition sessions.
- The data was in the form of unsegmented block recordings. Previously conducted work indicated that segmenting the data into the form of so-called gait cycles allows to obtain higher efficiency of the biometric system. A gait cycle is defined as the time from the moment the right foot touches the ground to the moment it touches the ground again [13]. As a result of the segmentation, samples were obtained which formed the input of the classifiers. In the case of day one, 4231 samples were collected (minimum 116, maximum 402), during the second day 3814 (minimum 115, maximum 516), and during the third day of acquisition 4254 (minimum 198, maximum 680).

Figure 1 shows the already segmented signals for one of the participants. The drawing has 3 rows and 2 columns. The first column presents data from the triaxial accelerometer and the next column from the triaxial gyroscope. The rows show the data collected during each session/day. Each window has three graphs in red, green and blue, which is a result of the use of triaxial sensors.

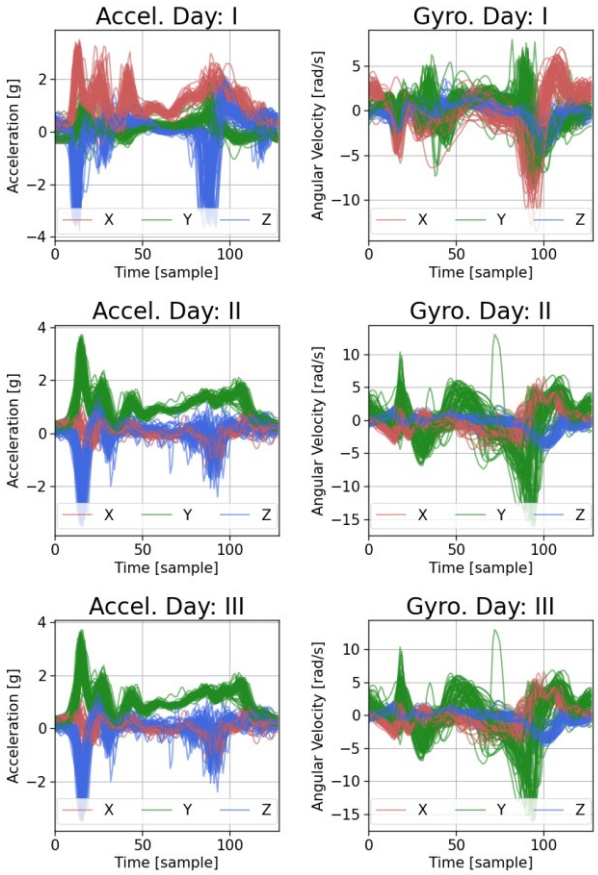


Fig. 1. Example of measurement data for triaxial accelerometer and gyroscope in raw form for three days of acquisition.

Analyzing the data from the accelerometer, it can be seen that the data collected during Day II and Day III of the acquisition are close to each other, while the data from Day I are quite far from them. For the latter, positive values are observed for the X-axis of the sensor, and for the other days for the Y-axis. The existing difference is due to the fact that in the case of Day II and Day III the phone was placed in a similar manner in the pocket, while on Day I its orientation was different. The motion sensors measure in a local frame of reference, so their positioning in the pocket affects the recorded values.

### B. Data processing

Pre-processing included detecting gait cycles, (the segmentation algorithm was described in our previous work [11]) and applying additional processing to minimize the impact of how the phone is placed in the pocket. In the case study (cell phone acquisition), it was very important to minimize the negative impact of sensor orientation on the measurement value. As part of this work, the so-called "Orientation Independent Transformation" algorithm described in detail in [12] was implemented. This method was based on the use of triaxial accelerometer signals to create a new artificial reference system. An accelerometer is a sensor that measures the sum of the acceleration caused by the motion of an object ($\widetilde{a_s}$) and the

components resulting from the gravitational acceleration ($g_s$) (which is approximately 1 g). The measured values depend on the sensor orientation modeled by the $R$ matrix of equation (1).

$$a = R(\widetilde{a_s} + g_s) \quad (1)$$

where:
$a$ – value read by the accelerometer;
$R$– rotation matrix from global to sensor coordinates;
$\widetilde{a_s}$ – acceleration resulting from the motion of an object, in the global reference system;
$\widetilde{g_s}$ – gravitational acceleration, in the global reference system ⟨0,0,1g⟩.

The method of minimizing the impact of the mounting method was based on the creation of three orthogonal axes (vectors in 3D space) $\langle \xi, \zeta, \psi \rangle$ which were used to create a new reference system. Due to the fact that the accelerometer always measures values resulting from gravitational acceleration, analysis of the accelerometer's measurement values as vectors in 3D space makes it possible to determine the average value, which will represent the coarse direction of gravity. Therefore, vector $\xi$ (i.e., the new axis of the coordinate system) was determined roughly as the average direction of acceleration. Then, for the second axis of the coordinate system, analysis of variance was applied. In the case of walking, there will be significant changes in acceleration in the direction of movement, so the vector $\zeta$ was determined using principal component analysis as the direction with maximum variance. The last axis of the coordinate system $\psi$ was determined as the vector product of the previous two axes in order to maintain orthogonality. Figure 2 presents the measurement values transformed to the new reference system.

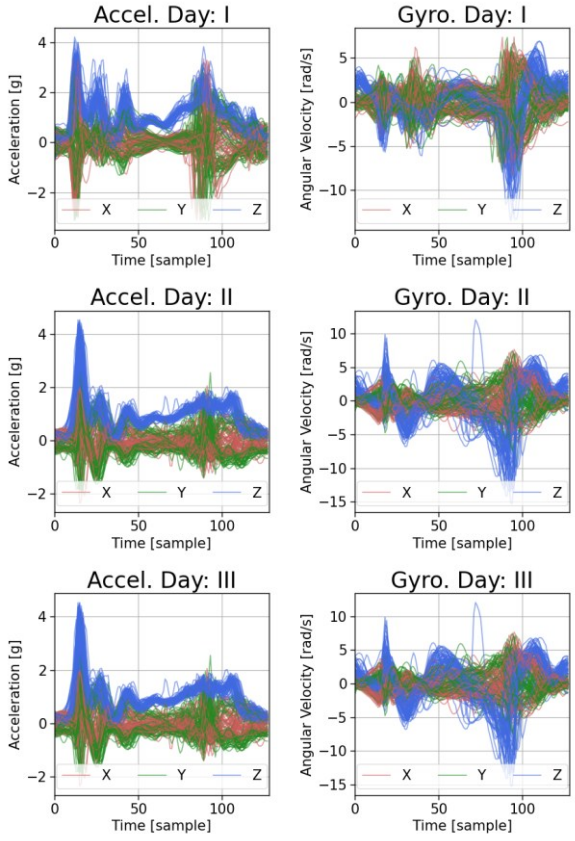


Fig.2. Example of measurement data of triaxial accelerometer and gyroscope in processed form for three days of acquisition.

From the illustration shown, it can be observed that:

- The signals have a similar shape over the three days. For the accelerometer signal for each of the three rows, the Z-axis signal is dominant and similar to each other.
- For both the X and Y axis signals of the accelerometer, untypical symmetries can be observed. These signals are often reflected relative to the "0" axis.

The symmetry with respect to the "0" value for the X and Y signals of the accelerometer is due to the fact that the analysis of variance was used in determining the coordinate system $\zeta$ axis. If the participant in the experiment alternately accelerates or decelerates, direction negation will occur. From the fact that the last $\psi$-axis is based on the vector product (including the $\zeta$-axis) negation may also occur. The "Orientation Independent Transformation" [12] method is able to minimize the influence of the way the sensor is mounted, while it is not immune to changes in gait speed.

### C. *Synthetic data generation*

The approaches presented in this section, based on gait cycle extraction [12] and the use of a decision model in the form of a CNN network [1], are related in the literature. In our subjective opinion, modifications to the architecture of the CNN network will not lead to significant changes in efficiency due to the limited number of learning samples (in the data corpus used, a minimum of about 100 samples for one day and one participant).

In the field of human activity recognition, similar to the present research, the approaches based on the generation of synthetic samples [19] have recently been noted to allow the improvement of identification rates. This paper describes the results of basic research that has tried to apply artificially generated samples under very specific conditions: the gait patterns observed for specific individuals have two main trends (as can be seen for the X,Y axis of the sensors in Figure 2), and the data are much less numerous than in the case of HAR applications.

The present study examines the utility of synthetic samples generated by generative models and their influence on the precision of biometric systems. Two distinct architectural approaches were explored, each based on a different model: a variational autoencoder and LSTM-MDN model. For all the types of studied generative models, the collection was initially divided into 13 subsets according to the participant's label (Figure 3). This was followed by leading to the training of generative models so that each instance could creates a ordered number of synthetic samples. Subsequently, the original and synthetic samples were merged to create a new data set, which was then employed to train the classifiers.

The study employed a two-group approach to sample generation, utilizing variational autoencoder-based models as the primary method. In this type of architecture, due to the presence of a bottleneck, there is a compression of data in the feature space. Thus, after the training process, the model is able to reproduce the main data trends. In the study, two variants of autoencoders were examined, i.e. dedicated to working with time series timeVAE[14] and PyRaug multidimensional data[15]. Conversely, the potential of utilizing LSTM-MDN models, which are capable of modeling data distribution parameters based on training data, was explored. The paper employs the first author's implementation of the LSTM-MDN models, as detailed in reference [16].

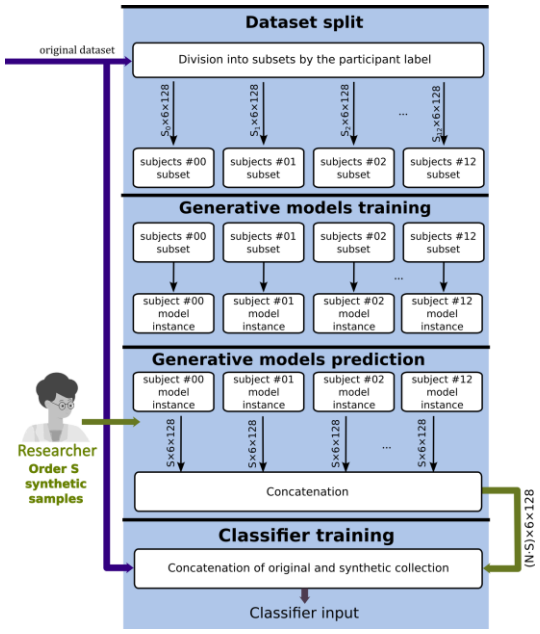


Fig.3. Idea diagram for ordering synthetic samples by a scientist

The study was conducted when {15, 60, 240} samples were generated for each participant. With 13 participants, there were {195, 780 ,3120} artificial samples, respectively. In the case of the combined datasets from the two days of acquisition (8045 samples), the artificial data were respectively {2.4%, 8.8%, 27.9%} of the training dataset.

### *D. Classifiers*

The use of data processing pipelines that include manual segmentation of gait cycles, minimizing the impact of sensor assembly and classification by a CNN-type network is based on the solution shown in [12]. In the case of unprocessed data, the signals will always be "smooth," whilst with additional transformations the signals will have discontinuities at the boundary of individual gait cycles. For architectures based on recursive models, this can cause problems with learning classifiers. There are several other reasons why-the authors did not choose to use recurrent networks.

Typically, CNN networks have lower learning data requirements than recursive networks. Convolutional networks allow the solution to be extended with advanced inference analysis from the Explainable AI stream based on gradient methods, e.g. the SHAP package or LIME. It is also important to note that analysis by a convolutional neural network allows for a more critical evaluation of all available gait samples. When using a recursive architecture, such as a long short-term memory (LSTM) network, a prediction of one label is made based on a selected number of historical samples. This raises additional questions about how many previous gait cycles to consider, which affects the number of test data points.

In the present study, experiments were conducted using three neural network architectures. The CNN [1] and CNN networks with the attentional mechanism (based on[17]) and Multi-Input CNN [18]. The details of the different layers of the network are presented in Table I.

Table I
CNN ARCHITECURE

| CNN | | |
|---|---|---|
| **Layer** | **Type** | **Details** |
| 1 | convolution_1 | in=1, out=32,ks=[1,9],stride=[2] |
| 2 | max_pooling_1 | ks=[1,2], stride=[2] |
| 3 | convolution_2 | in=32, out=64,ks=[1,3],stride=[1] |
| 4 | convolution_3 | in=64, out=128,ks=[1,3], stride=[1] |
| 5 | max_pooling_3 | ks=[1,2], stride=[2] |
| 6 | convolution_4 | in=128, out=128,ks=[6,1], stride=[1] |
| 7 | dense | in=2048, out=13 |
| 8 | softmax | |
| **CNN with Attention Mechanism** | | |
| **Layer** | **Type** | **Details** |
| 1 | convolution_1 | in=1, out=32,ks=[1,9], p=[0,4], stride=[1,2] |
| 2 | max_pooling_1 | ks=[1,2], stride=[1,2] |
| 3 | batch_normalisation_1 | n_features=32, |
| 4 | convolution_2 | in=32, out=64, ks=[1,3], p=[0,1], stride=[1,1] |
| 5 | convolution_3 | in=64, out=128,ks=[1,3],p=[0,1], stride=[1,3] |
| 6 | max_pooling_3 | ks=[1,2], stride=[1,2] |
| 7 | batch_normalisation_3 | n_features=128, |
| 8 | convolution_4 | in=128, out=128,ks=[6,1],p=valid |
| 9 | batch_normalisation_4 | n_features=128, |
| 10 | attention | channel=128,reduction=8 |
| 11 | dense | in=2048, out=13 |
| 12 | softmax | |
| **Multi-Input CNN** | | |
| **Layer** | **Type** | **Details** |
| 1 | convolution_1 | in=1, out=240,ks=[1,10],p='VALID' |
| 2 | batch_normalisation_1 | n_features=240, |
| 3 | max_pooling_1 | ks=[1,2], stride=[1,2] |
| 4 | convolution_2 | in=240, out=300,ks=[1,7], p='VALID' |
| 5 | batch_normalisation_2 | n_features=300, |
| 6 | max_pooling_2 | ks=[1,2], stride=[1,2] |
| 7 | convolution_3 | in=300, out=360,ks=[1,5], p='VALID' |
| 8 | batch_normalisation_3 | n_features=360, |
| 9 | max_pooling_3 | ks=[1,2], stride=[1,2] |
| 10 | convolution_4 | in=360, out=420,ks=[1,3], p='VALID' |
| 11 | batch_normalisation_4 | n_features=420, |
| 12 | average_pooling_4 | ks=[1,5], stride=[1,2] |
| 13 | dropout | P=0.5 |
| 14 | dense | in=2520, out=13 |
| 15 | softmax | |

The first two architectures [1,17] accepted data block with a dimension of 6 × 128. The first dimension resulted from the use of a triaxial accelerometer and a gyroscope. Last of the used classifiers[18] accepted two data arrays of 3 × 100 dimension as input, where data of two modalities were given on separate branches. Each network was trained for a period of 300 epochs, using cross entropy cost function and the ADAM algorithm as the optimization method. It should also be noted that the number of network parameters varied greatly, with 158, 350 for the CNN, 160, 973 for the CNN with the attentional mechanism, and 1, 539, 853 for the Multi-Input CNN.

In the context of this research, a significant focus has been placed on the usage of convolutional neural networks (CNNs) with an attentional mechanism. With regard to the implementation specifics, the Squeeze-and-Excitation (SE) attentional mechanism has been employed. Of particular interest is the second excitation component, which introduces additional channel-wise parameters representing multipliers of individual channels. In this approach, attention provides supplementary parameters that model the channel dimension, with the weighting factors indicating its relevance and ultimately influencing the prediction outcome.

## V. RESULTS

### E. *Baseline results*

Figure 4 shows the subject identification score as box plots for raw (a) processed (b) signal. The Y-axis shows the value of the F1-score measure of 10-fold repeated simple validation. The X axis shows the tested scenarios - the number of motion tracking sessions in the training set. Cases were analyzed when: the training set contains samples taken during session I, session II and when they are combined. In addition, three architectures of tested classifiers were marked with color. CNN in blue, CNN with Attention mechanism in orange, and Multi-Input CNN in green.

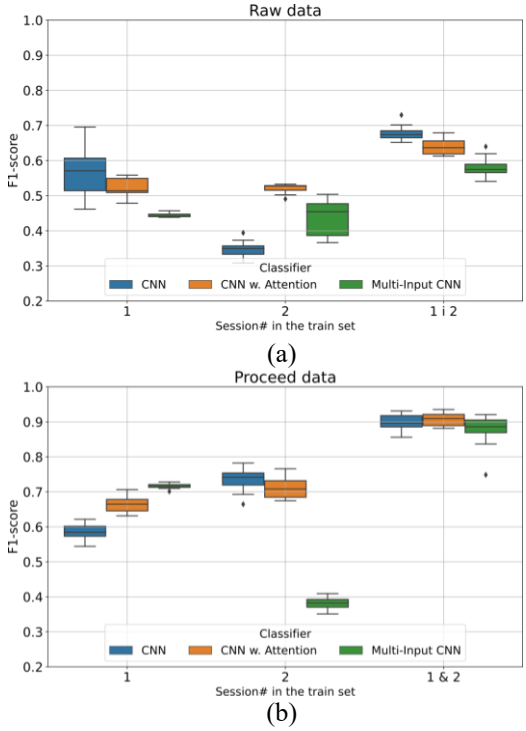


Fig.4. identification results raw data (a) proceed data (b)

The figure shows that both when using raw data (Figure 4 a) and processed data (Figure 4 b), maximum efficiency is achieved when using the combined Day I and Day II training sets. This is about 0.65 F1-score for the raw data and 0.9 F1-score for the processed data. In the analyzed case study, updating the training samples - changing the training session from I to II (where the time interval until the acquisition of test samples is shorter) in some cases reduced the classification metrics. For raw data, a decrease in effectiveness is observed for the CNN classifier(blue color), and for processed data for the Multi-Input CNN classifier (green color). On the other hand, the scenario of combining samples from days I and II in each of the analyzed cases allowed to increase the accuracy of the biometric system. This is our recommended scenario for biometrics system building.

### F. *Synthetic data results*

The highest classification scores were observed for data processed in the scenario of combining two days of data acquisition. In the case of using a CNN classifier with an attentional mechanism average 0.903 F1-score was observed.

For such running conditions, an experiment was conducted in which synthetic samples were added to the original data set.

Samples was created using variational autoencoder-based models as well as LSTM-MDN models. The last of these has achieved efficacy gains in our other studies [16]. In which a different 100-person dataset was analyzed (which did not require minimizing the impact of sensor montage using the method described in [12], due to professional motion tracking system usage). Figure 5 presents the results in the form of heatplot graphs, where blue color indicates the baseline without synthetic samples, red color indicates the average result lower than the baseline, and green color higher. Figure contains a total of 6 subgraphs arranged in 2 rows and 3 columns. The top row presents the baseline, and the results achieved for two autoencoders variants timeVAE and RHVAE/PyRaug. For the former, only results worse than the baseline were achieved. In the second architecture type, the highest result was 0.920 and was observed with the generation of 60 samples.

The lowest row presents the results observed for data generation using LSTM-MDN models with different numbers of tested normal distributions. High results were observed for modeling two data distributions and 240 samples. In this architectures models, there is the possibility of additional variance amplification (what was presented as additional rows). However, in the analyzed cases, this amplification was not essential.

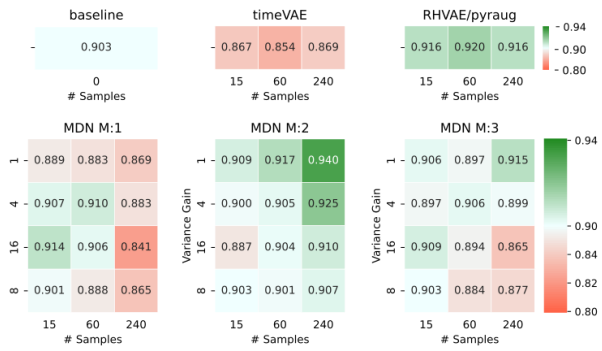


Fig.5. Median results from 20 iterations of learning and evaluation heatplot charts

The study shows that the metrics of the biometric system can be improved by using synthetic samples. For generative models based on variational autoencoders, the highest increase to 0.920 (about 2%) was possible for the RHVAE architecture. For LSTM-MDN models, the highest increase in biometric system performance to 0.940 (about 4%) occurred with the generation of 240 samples for modeling two data distributions.

The research in this paper employs the F1-score metric as a measure that accurately reflects the performance of a biometric system in an engineering context. It should be noted that in practical applications, particularly cross-day validation, the generation of samples that are highly similar to the original will not necessarily result in an increase in identification metrics. In the publication [20], the t-distributed stochastic neighbor embedding (t-SNE) method was used with dimensionality reduction to 2 in order to demonstrate the similarity between the real and synthetic data. In the present work, this approach was repeated. Additionally, an overview figure was prepared for the data of participant 11, which exhibited particular issues.

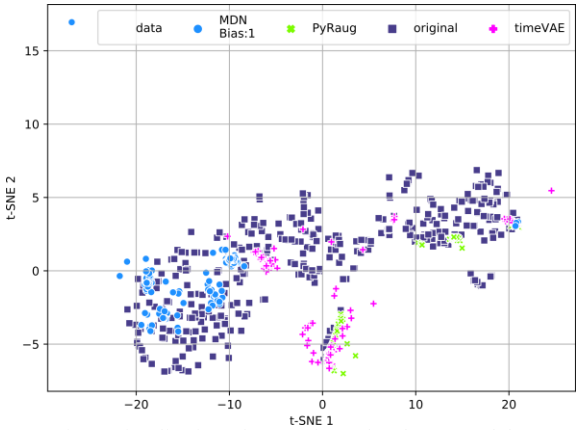


Fig.6. Visualization of t-SNE for real and generated data

In Figure 6, the empirical data are presented using rectangular markers with a dark blue color. This set is the most numerous and encompasses the largest numerical ranges within the visualized space. The data generated by the LSTM-MDN method with a BIAS parameter equal to 1 is presented using blue circle-shaped markers. In this case, the generated samples only partially coincide with the real data, and outliers also appear. The same is true for the TimeVAE and PyRaug methods, which also have markers, only partially complete the reduced feature space. The presented figure shows that the generated data are differentiated, and each method allows the creation of samples of a different character.

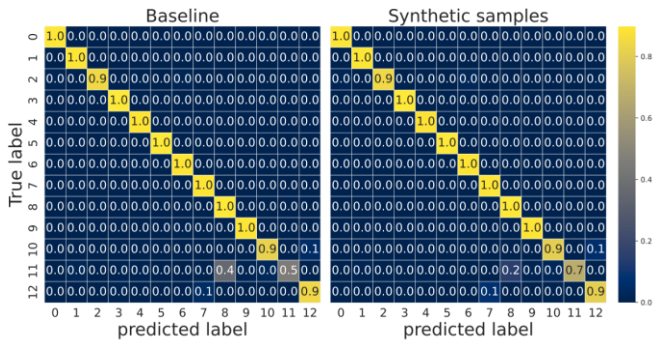


Fig.7. Averaged confusion matrix for 20 repetitions

Figure 7 presents the confusion matrices for the baseline case and for the synthetic data generation scenario (LSTM-MDN M:2, AUG_NUM:240, BIAS:1). The heat plot figures show the averaged values from 20 iterations, and it can be concluded from these that the increase in efficiency presented in Figure 5 results primarily from the improved identification of a participant with ID 11.

In further experiments, we used SHAP (SHapley Additive exPlanations) methodology to explain what the reason was for the poor identification of samples 11 by the base model. Figure 8 shows the result of the analysis in the case of the base model (a) and when synthetic samples are included (b). For each subgraph, there are 13 columns (number of labels), and two rows (a demonstration example for two samples for which the base model indicated the label incorrectly, and the model learned from the synthetic samples correctly). Each heatplot has a dimension of 6×128, where the first dimension corresponds to the number of measurement channels and the second to the length of the sample. Areas colored blue indicate data that cause the label to be disregarded (negative values of SHAP coefficients), and pink color positively affecting the prediction of the label (positive values of SHAP coefficient), light blue color indicates data not affecting the prediction.

For the baseline model, the test samples are only strongly associated with labels 11 and 8, showing a negative impact with the former (shade in blue). In addition, in the case of label 11, a lack of interaction with the third row (Z-axis of the accelerometer,) which takes on a neutral hue, is evident. A strong negative interaction with the 4th row (X axis of the gyroscope) is also visible.

The model demonstrated a notable impact on the labels assigned to 11, 8, and 9 when it was trained on both real and synthetic data. The graphical representation demonstrates that the artificially generated data exhibited a notable degree of proximity to the original data set of participant 9. Furthermore, in this instance, a more pronounced positive impact on participant 11's label was discernible (evident in the elevated pink hue), particularly in rows 4 and 5 (represented by the *X* and *Y* axes of the gyroscope).

By applying the Explainable AI method, we observed that the use of synthetic data of participant 11 had little effect on the activation of participant 9's label. This may indicate the similarity of participant 11's artificial data and participant 9's original data. This is a very important observation that indicates the need for validation of synthetic data.

Additional materials including experimental results and trained models for the scope of Exaplainable AI are made available in a private repository [https://github.com/ asawicki-pb/Biometrics-gait-system-based-on-motion-sensors-embedded-in-a-mobile-phone](https://github.com/ asawicki-pb/Biometrics-gait-system-based-on-motion-sensors-embedded-in-a-mobile-phone) to which access will be granted after email contact.

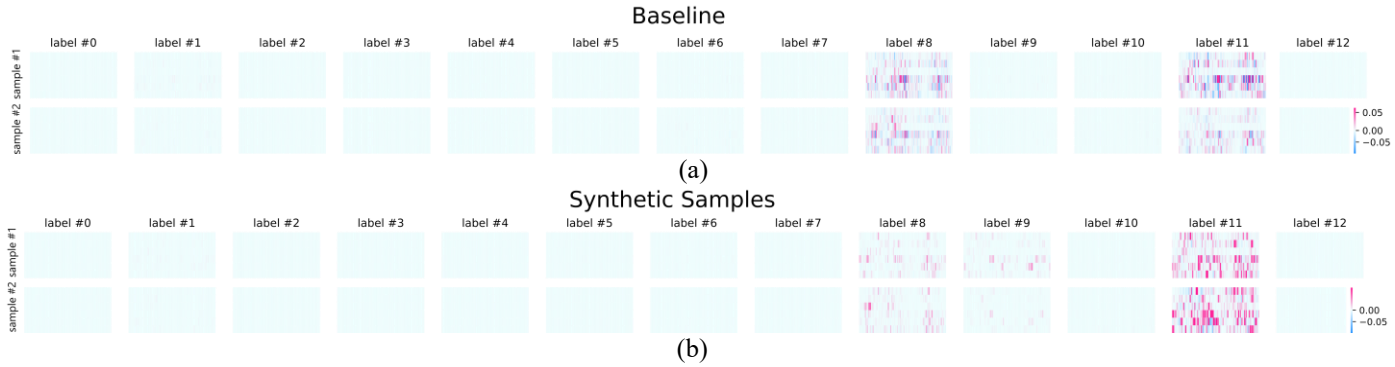


Fig.8. SHAP (SHapley Additive exPlanations) analysis results for the baseline CNN model (a) and the synthetic samples scenario (b)

## VI. LIMITATIONS

The current research is a case study developed on a publicly available dataset with the unique attribute of the availability of three acquisition sessions. The data made available should be approached with a certain amount of distrust. In our earlier studies, we pointed out a significant problem in data leakage between the training set (day II) and the validation set (day III). The test data (day III) of participant "6" represents the training data of participant "7" (day II). Thus, in the research presented here, the size of the collection was reduced from 14 participants to 13. This made the already modest dataset limited.

As is the case in other scientific studies in which a segmentation process is performed from block recordings, the result of the final identification is indirectly dependent on the quality of the implementation of the detection issue. Moreover, as a consequence of the segmentation process, the number of training samples (segmented gait samples) for each participant differed. This results in an imbalanced number of labels.

The combined training set from Days I and II yielded a minimum of 372 samples and a maximum of 1,233 samples, representing a ratio of over threefold. However, the unbalancing process was partially resolved when synthetic samples were added. In such a scenario, the minimum number of labels is 612, whilst the maximum is 1473 (more than twofold the minimum). The degree of data imbalance was thus diminished.

It is also important to acknowledge that, in the context of behavioral biometrics (including gait analysis), a multitude of factors can potentially compromise the accuracy and reliability of individual identification. It is important to note that aspects such as changing footwear between tracking sessions, changing the surface or slope of the gait path, or even the act of walking on different surfaces can affect the patterns of gait. In the case of cell phones, an additional factor is the manner in which the device is positioned within the pants pocket. With regard to the accelerometer sensor, the measurement values are also susceptible to rapid fluctuations. These fundamental disturbances can be mitigated through frequency filtering–that has already been implemented at the stage of data release. Furthermore, the impact of the mounting method can be offset through data transformation to an artificially created coordinate system (Methodologist's section, data processing).

The change in gait pattern due to a change in footwear or gait surface is beyond our control. However, using generative models to create synthetic samples allows to increase the generalization properties of decision models

## VII. CONCLUSIONS

In this work, we have successfully developed a biometric system based on accelerometer and gyroscope readings. These sensors were embedded in a cell phone located in the right front pocket of the pants. The work was carried out on the basis of a publicly available data corpus containing the unique feature of availability of three motion tracking sessions within separate days. As part of the ongoing work, three scenarios were verified in which a constant test set was built from the data collected during Day 3, whilst changing the method of creating the training set. When the training set consisted of Day I samples, the accuracy was 0.65 F1-score, whereas when it consisted of Day II samples, it was 0.7 F1-score. Finally, when the Day I and Day II data were combined, the accuracy was 0.9 F1-score. Classification was carried out for three variants of the CNN network, classical CNN, CNN with attentional mechanism and Multi-Input CNN.

Next, for the case of a set of combined two acquisition days and a CNN with attentional mechanism, the effect of the generation of synthetic samples by LSTM-MDN networks was examined. In the case of the combined dataset, in which synthetic samples accounted for about 30% (240 samples for each participant), an increase from 0.903 to 0.940 F1-score (about 4%) was observed. The study indicated that when more than one teaching session is available, it is more profitable to concatenate the data than to update it. In addition, by using synthetic samples at the learning stage, there is potential for a slight improvement in performance. The process of generating additional synthetic samples should be treated as a final step and should not be expected to have as far-reaching an impact on identification performance as a change in preprocessing

## REFERENCES

[1] Q. Zou, Y. Wang, Q. Wang, Y. Zhao, and Q. Li, 'Deep Learning-Based Gait Recognition Using Smartphones in the Wild', arXiv [cs.LG]. 2020. https://doi.org/10.48550/arXiv.1811.00338

[2] S. Sprager and M. B. Juric, 'Inertial Sensor-Based Gait Recognition: A Review', Sensors, vol. 15, no. 9, pp. 22089–22127, 2015. https://doi.org/10.3390/s150922089

[3] G. Giorgi, F. Martinelli, A. Saracino, and M. Alishahi, 'Try Walking in My Shoes, if You Can: Accurate Gait Recognition Through Deep Learning', 09 2017, pp. 384–395. https://doi.org/10.1007/978-3-319-66284-8_

[4] C. Wan, L. Wang, and V. V. Phoha, 'A survey on gait recognition', ACM Computing Surveys, vol. 51, no. 5, Aug. 2018. https://doi.org/10.1145/3230633

[5] A. Ajit, N. K. Banerjee, and S. Banerjee, 'Combining Pairwise Feature Matches from Device Trajectories for Biometric Authentication in Virtual Reality Environments', in 2019 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR), 2019, pp. 9–97. https://doi.org/10.1109/AIVR46125.2019.00012

[6] J. E. Boyd and J. J. Little, 'Biometric Gait Recognition', in Advanced Studies in Biometrics: Summer School on Biometrics, Alghero, Italy, June 2-6, 2003. Revised Selected Lectures and Papers, M. Tistarelli, J. Bigun, and E. Grosso, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, pp. 19–42. https://doi.org/10.1007/11493648_2

[7] R. Plucińska, K. Jędrzejewski, U. Malinowska, and J. Rogala, 'Influence of Feature Scaling and Number of Training Sessions on EEG Spectral-based Person Verification with Artificial Neural Networks', in 2023 Signal Processing Symposium (SPSympo), 2023, pp. 139–143. https://doi.org/10.23919/SPSympo57300.2023.10302695

[8] N. Al-Naffakh, N. Clarke, and F. Li, 'Continuous User Authentication Using Smartwatch Motion Sensor Data', in Trust Management XII, 2018, pp. 15–28. https://doi.org/10.1007/978-3-319-95276-5_2

[9] D. S. Matovski, M. S. Nixon, S. Mahmoodi, and J. N. Carter, 'The Effect of Time on Gait Recognition Performance', IEEE Transactions on Information Forensics and Security, vol. 7, no. 2, pp. 543–552, 2012. https://doi.org/10.1109/TIFS.2011.2176118

[10] S. Lee, S. Lee, E. Park, J. Lee, and I. Y. Kim, 'Gait-Based Continuous Authentication Using a Novel Sensor Compensation Algorithm and Geometric Features Extracted From Wearable Sensors', IEEE Access, vol. 10, pp. 120122–120135, 2022.
https://doi.org/10.1109/ACCESS.2022.3221813

[11] A. Sawicki and K. Saeed, 'Smartphone-Based Biometric System Involving Multiple Data Acquisition Sessions', System Dependability - Theory and Applications. DepCoS-RELCOMEX 2024. Lecture Notes in Networks and Systems, vol 1026, 2024, pp 252–260.
https://doi.org/10.1007/978-3-031-61857-4_25

[12] M. Gadaleta and M. Rossi, 'IDNet: Smartphone-based gait recognition with convolutional neural networks', Pattern Recognition, vol. 74, pp. 25–37, 2018. https://doi.org/10.1016/j.patcog.2017.09.005

[13] M. W. Whittle, 'Chapter 2 - Normal gait', in Gait Analysis (Fourth Edition), Fourth Edition., M. W. Whittle, Ed. Edinburgh: Butterworth-Heinemann, 2007, pp. 47–100. https://doi.org/10.1016/B978-0-7506-8883-3.X5001-6

[14] A. Desai, C. Freeman, Z. Wang, and I. Beaver, 'TimeVAE: A Variational Auto-Encoder for Multivariate Time Series Generation', arXiv [cs.LG]. 2021. https://doi.org/10.48550/arXiv.2111.08095

[15] C. Chadebec, E. Thibeau-Sutre, N. Burgos, and S. Allassonnière, 'Data Augmentation in High Dimensional Low Sample Size Setting Using a Geometry-Based Variational Autoencoder', IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 45, no. 3, pp. 2879–2896, Mar. 2023. https://doi.org/10.48550/arXiv.2105.00026

[16] A. Sawicki and D. Grabowski 'Application of Mixture Density Network for Sample Generation in Behavioral Biometrics', Computer Information Systems and Industrial Management. CISIM 2024. Lecture Notes in Computer Science, vol 14902 (2024), pp 30–43. https://doi.org/10.1007/978-3-031-71115-2_3

[17] H. Huang, P. Zhou, Y. Li, and F. Sun, 'A Lightweight Attention-Based CNN Model for Efficient Gait Recognition with Wearable IMU Sensors', *Sensors*, vol. 21, no. 8, 2021. https://doi.org/10.3390/s21082866

[18] R. Delgado-Escaño, F. M. Castro, J. R. Cózar, M. J. Marín-Jiménez, and N. Guil, 'An End-to-End Multi-Task and Fusion CNN for Inertial-Based Gait Recognition', *IEEE Access*, vol. 7, pp. 1897–1908, 2019. https://doi.org/10.1109/ACCESS.2018.2886899

[19] Li, X., Luo, J., Younes, R.: ActivityGAN: generative adversarial networks for data augmentation in sensor-based human activity recognition. In Adjunct Proc. of the ACM Int. Joint Conf. on Pervasive and Ubiquitous Computing and Proc. of the ACM Int. Sym. on Wearable Computers Association for Computing Machinery, 249–254 (2020), https://doi.org/10.1145/3410530.3414367

[20] D. Carneros-Prado, C. C. Dobrescu, L. Cabañero, L. Villa, Y. V. Altamirano-Flores et al.' Synthetic 3D full-body skeletal motion from 2D paths using RNN with LSTM cells and linear networks', Computers in Biology and Medicine, Volume 180, 2024, 108943, ISSN 0010-4825 https://doi.org/10.1016/j.compbiomed.2024.108943.