A study of how outlier detectors can accurately authenticate multiple persons using the heart rate from consumer-grade wearables

Matei Chiriță[1]

Supervisors: David Tax[1]**, Arman Naseri Jahfari**[1]**, Ramin Ghorbani**[1]

[1]EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 25, 2023

Name of the student: Matei Chiriță
Final project course: CSE3000 Research Project
Thesis committee: David Tax, Arman Naseri Jahfari, Ramin Ghorbani, Guohao Lan

Abstract

The aim of this paper is to complete the gap in the knowledge and experiment using as little as only the heart rate of some subjects to manage to successfully authorise them in some supposed system. The focus will be on the Gaussian Mixture model and the One Class Support Vector Machine, both outlier detectors, because most of the past research was focused on supervised models. Using these two, this paper will experiment with recognising intruders with models trained to distinguish one authorised person and multiple authorised persons. In the first case, multiple data processing methods and hyperparameters will be tested together and compared. In the second case, the goal will be to use and modify the best-found parameters of the first case to train models that are able to detect multiple persons as authorised. This time, there are two methods that this paper is going to look into and compare their performance: training one single model to detect all the subjects and training multiple models, one per subject. The most notable results are with one authorised person, with a score of 0.936, with two authorised, 0.88, and with 12 authorised, 0.713, when using the area under the curve metric.

1 Introduction

Wearables, such as smartwatches, fitness trackers, and other health monitoring devices, are becoming increasingly popular, with a market that will increase to about 265.4 USD billion by 2026 [1]. They can gather a vast amount of data and give us powerful insights with their health and fitness tracking. Overall, the data that these devices gather can be split into physiological metrics such as heart rate, SpO_2 and calorie burn and behavioural metrics such as step counter [10].

Person identification is a broad area that can improve the quality of life in many ways by helping in automatically modifying the environment and options around people using their preferences or physical characteristics. This is useful in life-saving situations: for example at a hospital when unconscious and a fast and accurate identification can give the medics important information about possible allergies to some medications or some previous medical records, but also in day-to-day activities such as entering a car and expecting it to adjust the seat, steering wheel and mirrors even before you set foot in the car. The main problem is that the available identification methods are inconvenient and require time-consuming input such as a code or a password. Using wearables for this task could be the answer to this problem.

Previous work of Issam Hammad and Kamal El-Sankary [3] has shown that person identification is possible through closely monitoring the physical movements of a person's body. In their experiment, they placed five orientation trackers on each subject and, using these measurements, they managed to identify individual subjects with an accuracy of 0.95.

There are some other papers that show that person identification can be done using ECG signals such us [4] which achieved a 0.94 accuracy with a Support Vector Machine (SVM).

Vhaduri, Sudip and Poellabauer, Christian [10] also propose a method of user identification, but this time using both physiological and behavioural metrics, in contrast with the above papers that only use one of them. With the calorie burn, metabolic equivalent of task, and heart rate they managed to distinguish one person from the rest. When they used an SVM they achieved a mean accuracy of 0.92.

Another study published by Alexa Muratyan et al. [6] focuses more on authorizing some subjects while rejecting others, rather than recognising a person in a group. For this experiment, the data was collected from 25 subjects and it consists of SpO_2 and heart rate. The best results resulted from training a random forest to only use the heart rate and then to use both measures, giving 0.7 and 0.8 accuracies, respectively.

All these studies share a common problem, they use metrics that are not always available in consumer-grade wearables. So, there is a limited number of people that could take advantage of the above discoveries. Also, there are few papers that treat person identification as a classification problem in the general classes of authorised persons and unauthorised ones, and among them, there is a lack of using outlier detectors. The advantage of using outlier detectors is that these only need the data from the one known class to be able to detect data from the unknown class as outliers.

This paper will focus on the question: "Using only heart rate data from a consumer-grade wearable, how well can Gaussian Mixture and One Class Support Vector Machine outlier detectors accurately distinguish multiple authorised persons from multiple unauthorised persons?". To reach this final conclusion, these sub-questions will first be answered:

- What are some important parameters for the data processing and models such that the models give the best results when trained on one authorised person?
- Using the above-discovered parameters, how does training one model with data from multiple known persons compare to training multiple models, one per known person?

The paper has the following structure: In Chapter 2, the methodology in place will be discussed. Chapter 3 will give the experiments and their setup, followed by Chapter 4 which contains a discussion about some observations during the experiments and constructs a base for some feature research ideas. Chapter 5 gives a conclusion and answers the main research question. Finally, Chapter 6 will explain the importance of the responsible research methods in use.

2 Methodology

This chapter explains how the research will be conducted, the resources used and the steps for its completion.

2.1 Data preprocessing

In this research, a time series of heart rate measurements gathered from multiple people using their common wearables is used. The first step towards the goal of gathering more data

about the usage of outlier detectors for detecting new persons was data preprocessing.

Data preprocessing is an important part of the experiments because it could make a big difference in the final results. The scope of this process is making sure that the subject data is recognisable and distinguishable by models so that the authorised subjects can be told apart from the unauthorised ones.

Window segmentation

First, the data will need to be segmented in windows, so that some important features can be crafted using more than one data point. There is an exchange between the number of data points and the number of dimensions of one data point. The chosen window sizes are 1-hour intervals and 3-hour intervals. These were chosen so that the trained models have enough dimensions per window to work with and such that the activity of the subjects is expected to generate some patterns. The 1-hour intervals emphasise some repetitions in the subject's tight scheduled daily activities, but since these are rarely done precisely at the same hour, they cannot fully describe a day. The 3-hour intervals also capture day-to-day activities, but this time there is more room for variability in the starting hours of these tasks. These windows are describing a better overview of the day-to-day activities that do not have that tight of a schedule.

For a window to ensure statistical significance and to capture important features, the preprocessing only keeps the windows that contain all the measurements in their specific time interval. So, the windows that have incomplete data are left out. This raises some problems when it comes to the 3-hour windows since the data set has few subjects that have enough continuous 3-hour data to train or test a model. To solve this problem, these larger 3-hour segments were taken using a sliding window protocol with a stride of 20 minutes. The 1-hour segments are kept disjunctive since they generate enough data points per subject.

Feature extraction

Second, per each window, the plain heart rate will need transformations in order to highlight the most important features of that time frame and to capture the most out of the data point in as few dimensions as possible. Besides, emphasising the important components of the plain data also improves the speed at which the models are trained and tested. The chosen transformations are a combination of statistical features, Mel-frequency cepstral coefficients (MFCC) and principal component analysis (PCA).

The process can be easily seen in Figure 1. It clearly describes how the heart rate time series of one subject is first sectioned in windows, each window being placed in a dataset. Afterwards, each window is transformed using the statistical features, MFCC and PCA to get the final data points that the models will use.

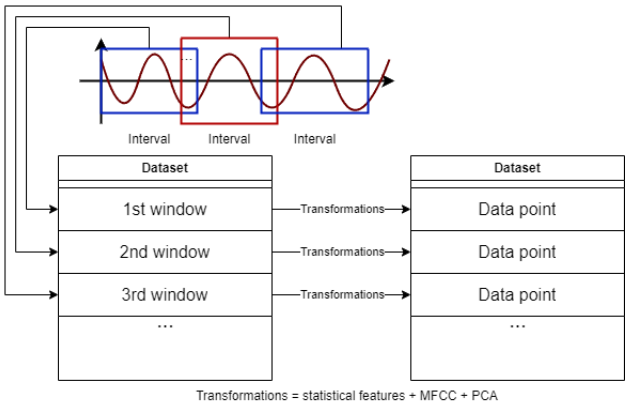

Figure 1: From time series to features and data points

The use of statistical features, besides offering some high dimensionality reduction, also extracts some meaningful properties of the data and captures its essential characteristics. They also help models distinguish different specific properties of each person's heart rate data. So, the experiments use a combination of the following statistical features: mean, median, standard deviation, 25^{th} / 75^{th} / 95^{th} percentile, interquartile range, range, mean absolute deviation, median absolute deviation, root mean square error.

Mel-frequency cepstral coefficients (MFCC) are an effective representation of the spectral characteristics of a time series. These are widely used in audio processing, but they are valuable in any time series application. The details of their calculation can be seen in [5]. The experiments will test how the usage of 0 to 20 components will influence the performance of the model, thus influencing the separability of data.

Principal component analysis (PCA) is the last transformation that will be applied to the raw data of a window, which again transforms the high-dimensionality data into low-dimensionality data, retaining its most important information. The chosen dimensions that will be tested are from 0 to 500.

2.2 Outlier detectors training

The next step in the process is to train some unsupervised models. To be more specific, unsupervised models refer to outlier detectors. These are trained on one or multiple subjects and then tested on a combination of multiple previously unseen subjects and known subjects. These models have the goal of differentiating the data points from known people, which can be called authorised, and unknown people, which can be called unauthorised.

Train and test sets separation

Using a dataset containing time series of the heart rates of multiple people, the window segmentation part of preprocessing is applied to each subject's data.

For each of the authorised persons, decided prior to each experiment, the first 0.8 of the total windows go towards the train set. This creates a data set in which one cannot differentiate between the authorised subjects. This behaviour is pos-

sible due to the fact that outlier detectors only need the known subjects to be trained. This data set is normalised and passed to be transformed in the feature extraction part and then used as training set by the models.

The train set is firstly composed of the last 0.2 of the total windows per authorised subject. Then the unauthorised persons, specific to each experiment, are added to the data set. After that, the test set is balanced such that the number of windows from known subjects is the same as the number of windows from unknown subjects by also making sure that all the authorised and unauthorised subjects have at least one window of data in the test set. The data set is then normalized using the mean and the standard deviation of the training set and passed to be transformed and tested against the trained models.

One versus many

The first experiments are looking into using a single subject as authorised and multiple as unauthorised, or in short one versus many. These will test various parameters for the feature extraction part and various model hyperparameters in order to push the model to give as good metrics as possible. Figure 2 describes how each test will be executed. First, the training set will contain data points from the authorised subject and the model will be trained using these. Second, the test set will contain data from both the authorised subject and multiple unauthorised subjects such as the number of known data points is equal to the number of unknown data points.

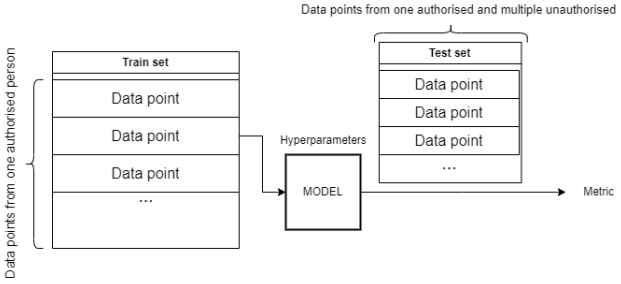

Figure 2: One versus many training steps

Many versus many

Using the feature extraction parameters and the model parameters discovered in the one versus many cases, the next experiments will cover training the model with multiple authorised persons and testing it against multiple unauthorised ones, namely many versus many cases.

This time the hyperparameters will be directly taken or deduced from the best-behaving model in one versus many. This was decided as it is impractical to calculate the best hyperparameters depending on the number of authorised subjects since a supposed system should be scalable in practice and be able to adapt to a growing number of known people.

For method comparison, when it comes to many versus many, the chosen outlier detectors will be first trained as a single model using data from all the known subjects, as in Figure 3, and as multiple models, each model identifying one of the known subjects and using their data as training as in Figure 4. In the second case, the number of models is the same as the number of subjects.

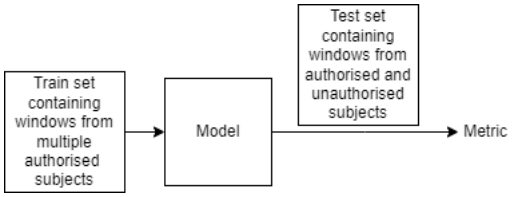

Figure 3: Multiple authorised, multiple unauthorised, one model training method

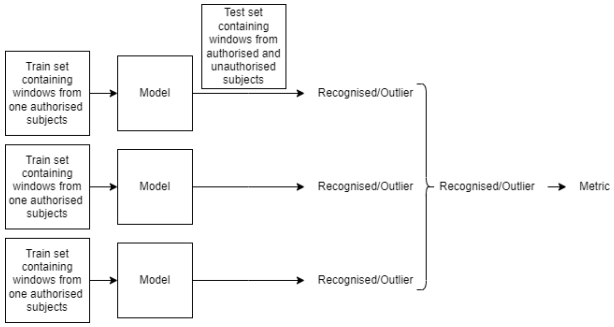

Figure 4: Multiple authorised, multiple unauthorised, multi-model training method

Models

The chosen models are the Gaussian Mixture model and the One Class Support Vector Machine (One Class SVM).

The Gaussian Mixture is best described in [7] and consists of a combination of Naive Gaussian models that are placed in such a way that they describe the data set in the best way possible.

The Support Vector Data Description model is a method that obtains a boundary around the train data with some flexibility. This method was first described in [8] by Schölkopf, which involves using a hyperplane (a plane in n-dimensions) for the decision boundary, and improved by David M.J. Tax and Robert P.W. Duin in [9] which involves a method of obtaining a spherical boundary.

Metric

During testing, for each data point, the log-likelihood of that point under the model is calculated as a score and used in the metric function.

Given the way that the results are presented, with a score per point and not a clear classification, the final classification has to be done by selecting a threshold such that the points scoring below it are classified as from an unauthorized person and the ones above are classified as from an authorized

person. This way of choosing a threshold can impact metrics such as accuracy, so the chosen metric is the area under the receiver operating characteristic curve (AUC). This metric goes through all possible thresholds and generates a score between 0 and 1 independent of the thresholds, 0 meaning that the model predicts 100% wrongly and 1 meaning that the model predicts 100% correctly.

2.3 Comparison

When it comes to one versus many, the main term of comparison will be between the best results of the chosen models. Besides this, the value that each data transformation brings to the model will also be analysed by comparing the behaviours of the models when excluding the MFCC, PCA and the statistical features one by one.

The next step will be to conduct a comparison between the approaches of training one model that detects multiple subjects or multiple models, one per subject.

The end task will be to give an overview of how well the outlier detectors behave in the task of identifying a group of persons and rejecting another one, taking into account the preliminary results and the above methods.

3 Experimental Setup and Results

This section will talk in detail about every step that was taken and the results that followed from each experiment. All experiments were implemented using Python 3.11.3 and the code is available in a git repository[1].

3.1 Dataset

The dataset on which the experiments were done is called ME-TIME [2], registered at ClinicalTrials.gov with ID: NCT05802563, and consists of heart rate and step counter data extracted from 54 persons using Fitbit Inspire 2, Charge 2 and Charge 5. The data was extracted as two time series, one for the heart rate and one for the step counter, with variable lengths, depending on the subject.

The given data was already a little processed, so this is the format of the data in the way it was presented: The heart rate has been resampled to 0.2 Hz (once every 5 seconds). The original sample rate was variable but 0.2 Hz most prevalent. If the heart rate is missing for a maximum of 12 samples, the signal is linearly interpolated. If the heart rate is constant for more than 12 samples, the constant sequence is removed. The step counter is sampled once every minute.

Since this research only analyses the results using the heart rate data, the step counter was left out.

3.2 Gaussian Mixture Model (GMM)

For the implementation[2], the GaussianMixture class from sklearn was used and the best hyperparameters were computed depending on the experiments.

One versus many

The goal of this experiment is to find the best parameters for a supposed model that can distinguish one subject and register any others as outliers. The main parameters that can be changed in this case are the window size, the statistical features, the number of MFCC components, the number of PCA dimensions, when it comes to the train and test data sets, and the number of Gaussian distributions of which the model is composed when it comes to the model itself.

With the window size set to 1 hour and 3 hours, the following experiment trains Gaussian mixture models with every combination of the parameters described in Table 1.

Parameter	Options
GMM Distributions	4, 10, 20, 40
MFCC components	0, 5, 10, 15, 20
PCA dimensions	0, 4, 10, 20, 50, 100, 250, 400, 450, 500
Statistical features	{}, {mean, median, standard deviation, 25^{th} / 75^{th} / 95^{th} percentile, interquartile range, range, mean absolute deviation, median absolute deviation, root mean square error}, {mean, median}, {mean, median, standard deviation}, {mean, median, 25^{th} / 75^{th} percentile}

Table 1: Parameters options taken for the one versus many experiments for Gaussian Mixture

Because of the fact that some people are more similar to others, the model can be more accurate at distinguishing some subjects than others. So it is not enough to take one combination of subjects and draw a conclusion on the performance of the model and the data preprocessing regarding their parameters. So, for each combination of parameters that define a model and a data format, the model was trained and tested using AUC against 100 sets of 12 people, 1 authorised and 11 unauthorised. For each model, the same 100 combinations of 12 people were used. Afterwards, the final score of a model and data format was calculated as a mean on the 100 AUC scores and it was combined with its standard deviation.

In Table 2 a ranking of the best parameters when taking 1-hour windows on data can be seen and in Table 3, also a ranking of the parameters but when taking 3-hour windows. The performance of the models was sorted by the mean AUC score. Both tables display only the first 12 best models although all over 1000 combinations of parameters resulted in a score of the model. The exhaustive search has been done using GridSearchCV[3] from sklearn because of its capability of parallelizing the executions, shortening the waiting times. Although the scores for some variants seem the same, they are clipped to 3 decimals, the scores still being in decreasing order.

[1] https://github.com/mateichirita/outlier_detection_for_hr_data

[2] https://scikit-learn.org/stable/modules/generated/sklearn.mixture.GaussianMixture.html#sklearn-mixture-gaussianmixture

[3] https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html

	GMM distributions	MFCC components	PCA dimensionality	statistical features	mean AUC	std AUC
1.	10	5	10	[]	0.830	0.072
2.	10	5	20	[]	0.830	0.072
3.	10	5	450	[]	0.830	0.072
4.	10	5	500	[]	0.830	0.072
5.	10	5	400	[]	0.830	0.072
6.	10	5	250	[]	0.829	0.072
7.	10	5	100	[]	0.829	0.072
8.	10	5	4	[]	0.829	0.073
9.	10	5	50	[]	0.829	0.072
10.	20	5	500	[]	0.828	0.072
11.	20	5	10	[]	0.828	0.073
12.	20	5	250	[]	0.828	0.072

Table 2: Top 12 Gaussian Mixture models with 1h windows in one versus many case

	GMM distributions	MFCC components	PCA dimensionality	statistical features	mean AUC	std AUC
1.	4	5	0	[mean, median]	0.936	0.044
2.	4	5	4	[]	0.935	0.042
3.	4	5	100	[]	0.935	0.043
4.	4	5	500	[]	0.935	0.043
5.	4	5	20	[]	0.935	0.043
6.	4	5	50	[]	0.934	0.044
7.	4	5	250	[]	0.934	0.044
8.	4	5	0	[mean, median, standard deviation]	0.934	0.048
9.	4	5	400	[]	0.934	0.044
10.	4	5	450	[]	0.934	0.045
11.	4	5	10	[]	0.934	0.045
12.	10	5	250	[]	0.933	0.043

Table 3: Top 12 Gaussian Mixture models with 3h windows in one versus many case

With the smaller window, the best results were given by having 10 Gaussian distributions, 5 MFCC components, 10 PCA dimensions and no statistical features, giving a mean AUC score of 0.83.

To assess the importance of each transformation on the dataset, the best model for the 1-hour windows is taken. The transformations applied before training the model are changed such that the MFCC distributions are set to 0, then the PCA dimensionality is set to 0 and finally, the number of statistical features is set to 0. When applying no MFCC, the AUC score drops to 0.634 and when applying no PCA, the score drops to 0.746. From this, the conclusion is that the MFCC has the most influence in the way the model behaves, with the drop of not using it being 0.196, while not using the PCA results in a drop of just 0.084. As for the statistical features, it is already known that there is no advantage in using them since the best model does not use them.

With the 3-hour windows, with a mean AUC score of 0.936, the best model uses 4 GMM distributions, 5 MFCC components, no PCA transformations, and mean and median. This is a significant improvement, of more than 0.1, over the best model trained with 1-hour windows. This means that the data from a 3-hour window gives more significant features that can distinguish one known person from a group of unknown persons.

To again assess the importance of each transformation, this time using the 3-hour window trained models, the same procedure applies. Taking no MFCC components for the best-behaving model decreases the mean AUC by 0.19, the score being 0.746. When no statistical features are used, the decrease is much lower, resulting in a score of 0.89. Again the MFCC gives the features with the highest significance for this model. This time, the PCA is already not used on the data, so it can be seen from the table that in the case of the best model, it has no value.

So, in both cases, the MFCC transformations are able to extract the most meaningful properties that separate the subjects enough so that their data can be distinguished by a Gaussian Mixture.

Many versus many

This section will provide a side-by-side comparison of the methods described in the methodology section, training one model with all authorised subjects and training multiple models, one per authorised subject subject

The model for single-model method is taken from the best model from the one versus many case but with a changed number of Gaussian distributions. So, the model is composed of 4 times the number of authorised people distributions, 5 MFCC components, no PCA transformations and mean and median, on the 3-hour windowed data. The number of distributions was multiplied by the number of people with the assumption that every 4 distributions will behave best at describing only one person.

The models for the multi-model method are exactly the best model on the 3-hour windowed data.

For testing and scoring, 100 unique combinations of 18 people were chosen. For each combination of subjects, the

first 2 to 12 persons are taken as authorised and the last 6 as unauthorised, each authorised having at least 1500 data points.

As a note on testing the second method of outlier detection, the test set was evaluated by all the models with the likelihood of each data point being in each model. From these evaluations, the implementation keeps the lowest likelihood among the models for it to be evaluated with the AUC. This way, if at least one model sees the data as from an authorised person, the data is considered from a known person.

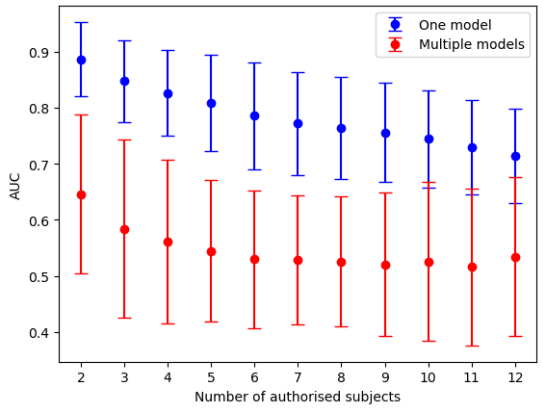

Figure 5: One model GMM performance versus multi-model GMM performance

As shown in Figure 5, the single model method has a slow drop as the number of authorised subjects increases. It starts with an average AUC of 0.88 and decreases until 0.71 when it was trained on 12 authorised people. The standard deviation, presented by the vertical bars, seems to remain constant at around 0.08.

Likewise, the multi-model method drops slowly, but this time the mean AUC seem to converge around 0.52. In this case, the maximum score is obtained again when trained with 2 authorised subjects and it is 0.64. This time the standard deviation stays around 0.14, remaining mostly constant.

The conclusion is that for 2 to 12 subjects, the one-model method behaves better and it has higher results. It also gives more stable results, with a lower standard deviation. Since the trend is towards a decreasing mean AUC for the first method, the more subjects the closer the scores of the two models.

3.3 One Class Support Vector Machine (One Class SVM)

The implementation used is taken from sklearn [4].

One versus many

The goal of this experiment is again, as with the Gaussian Mixture, to find the best parameters for a supposed model that can distinguish one subject and register any others as outliers.

[4]https://scikit-learn.org/stable/modules/generated/sklearn.svm.OneClassSVM.html

The main parameters that can be changed in this case are the window size, the statistical features, the number of MFCC components, the number of PCA dimensions, when it comes to the train and test data sets, and the nu parameter of the model which is an upper bound on the fraction of training errors and a lower bound of the fraction of support vectors. The kernel was set to a radial basis function kernel (rbf).

For the windows of 1 hour and 3 hours, the One Class SVM was trained to combine the parameters in Table 4.

Parameter	Options
nu	0.1, 0.3, 0.5, 0.8
MFCC components	0, 5, 10, 15, 20
PCA dimensions	0, 4, 10, 20, 50, 100, 250, 400, 450, 500
Statistical features	{}, {mean, median, standard deviation, 25th / 75th / 95th percentile, interquartile range, range, mean absolute deviation, median absolute deviation, root mean square error}, {mean, median}, {mean, median, standard deviation}, {mean, median, 25th / 75th percentile}

Table 4: Parameters options taken for the one versus many experiments for One Class SVM

With the same reasoning that 2 different people can have similar heart rate data, but they can also function in 2 completely different ways, it is again a good idea to test the hyperparameters of the model with different groups of people. This time again, I chose the same 100 combinations of 1 authorised person and 11 unauthorised per hyperparameter. The final score of a model with some hyperparameters is a mean of the AUC scores from the 100 tests.

Table 5 presents the best 12 models using 1-hour windows and Table 6 presents the top 12 models using 3-hour windows. These are the best 12 models of a total of 1000 trained with every combination of parameter values. Although the scores might look the same, only the most significant 3 decimals were left in the tables, the scores still being in decreasing order.

The best model when using 1-hour windows has a nu of 0.8, and the data is transformed into 20 MFCC components, 100 PCA dimensions and with no statistical features. This model has a mean AUC of 0.648.

To check the importance of each transformation, the best model for 1-hour windows is taken and the parameters are one by one set to 0. When setting the number of MFCC transformations to 0, the mean AUC score decreases from 0.648 to 0.6, so a decrease of 0.048. When doing the same thing with the PCA transformations, the value drops to 0.635, resulting in a smaller decrease of 0.013. This means that the MFCC processes the data in such a way that it gives the most information to the model. For the statistical features, it is already

	nu	MFCC components	PCA dimensionality	statistical features	mean AUC	std AUC
1.	0.8	20	100	[]	0.648	0.115
2.	0.8	20	250	[]	0.648	0.115
3.	0.8	20	4	[]	0.648	0.115
4.	0.8	20	50	[]	0.648	0.115
5.	0.8	20	400	[]	0.648	0.115
6.	0.8	20	500	[]	0.648	0.115
7.	0.8	20	10	[]	0.648	0.115
8.	0.8	20	20	[]	0.648	0.115
9.	0.8	20	450	[]	0.648	0.115
10.	0.8	20	250	[mean, median]	0.648	0.115
11.	0.8	20	20	[mean, median]	0.648	0.115
12.	0.8	20	400	[mean, median]	0.648	0.115

Table 5: Top 12 One Class SVM models with 1h windows in one versus many case

	nu	MFCC components	PCA dimensionality	statistical features	mean AUC	std AUC
1.	0.8	5	450	[]	0.785	0.112
2.	0.8	5	250	[]	0.785	0.112
3.	0.8	5	10	[]	0.785	0.112
4.	0.8	5	20	[]	0.785	0.112
5.	0.8	5	50	[]	0.785	0.112
6.	0.8	5	400	[]	0.785	0.112
7.	0.8	5	500	[]	0.785	0.112
8.	0.8	5	4	[]	0.785	0.112
9.	0.8	5	100	[]	0.785	0.112
10.	0.1	5	450	[]	0.785	0.108
11.	0.1	5	500	[]	0.785	0.108
12.	0.1	5	4	[]	0.785	0.108

Table 6: Top 12 One Class SVM models with 3h windows in one versus many case

shown in the table that they give no new information to the highest-rated model

Next, the best model using 3-hour windows has a nu of 0.8, and the data is transformed into 5 MFCC components, 450 PCA dimensions and with no statistical features. Compared to the best model resulting from 1-hour windows, this one has a mean AUC of 0.785, giving an increase of 0.137. This means, that in One Class SVM, again, the 3-hour window is more beneficial.

The importance of each transformation will be checked in the same way. Setting the number of MFCC transformations to 0 results in a score of 0.676 and using no PCA the score drops to only 0.758. This again results in MFCC being more important as the score decreases with 0.109 when not using it, compared to the decrease of 0.027 when not using PCA. The statistical features are the least important as they are not even used in the best combination.

So, overall, MFCC transformations are the ones that give the most relevant information to the best model and the least important ones are the statistical features.

Many versus many

This section will provide a side-by-side comparison of the methods described in the methodology section, training one model with all authorised subjects and training multiple models, one per authorised subject. This time One Class SVM model is used.

The models for single-model methods and multi-model methods are using the exact parameters of the best model in one versus many, a nu of 0.8, 5 MFCC components, 450 PCA dimensions and no statistical features.

The train and test sets are taken exactly as for many versus many methods in the Gaussian Mixture and the evaluation of the multi-model also follows the same reasoning.

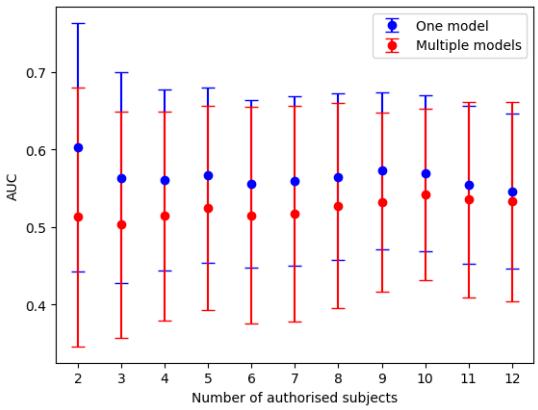

Figure 6: One model One Class SVM performance versus multi-model One Class SVM performance

As shown in Figure 6 the one-model method behaves better for the number of authorised subjects in the experiment, with

the highest mean AUC of 0.603 when only trained on 2 authorised subjects. The largest score for the multi-model method is 0.541 for 10 authorised subjects, although all the values are between 0.51 and 0.55, pointing out that the score has converged in that area. Also, going back to the first method again, the scores also decrease to a point where they converge around 0.55. Something else to note is that the difference between the AUC scores between the two methods decreases with each added authorised subject.

In conclusion, for 2 to 12 authorised subjects, the one model approach did better, but the trend indicates that with more subjects, the performance of the two methods might become almost identical.

4 Discussion

The first thing to talk about is the choice of using only the heart rate data. In some early experiments that were using very few data points, the step counter was also a valuable asset, improving the results from only using the heart rate. The problem with the step counter is the presence of numerous data gaps and the difficulty in finding windows with complete series of both step count and heart rate. So, while the models improved with more heart data windows, the improvement of adding a step counter could not be measured due to the lack of data. However, it is important to note that the step counter data has the potential of improving the models' performances.

Another choice note worthy taken in this paper is only taking into account 1-hour windows and 3-hour windows in one versus many measurements. This choice significantly impacts the models' performances and even more experiments on window sizes could have yielded some impressive results. The next step would have been taking 6-hour windows, equivalent to taking the 4 most important phases of a day, morning, noon, evening and night. This window size was left out due to the computational complexity of the calculations with 4320 measurements per window.

Moreover, the choice of using a sliding window approach with 20 minutes stride when it comes to the 3-hour windows was due to the limited availability of 3 hours of continuous data. While this approach allowed enough data to be generated, it may have resulted in some temporal loss of information. Meanwhile, to avoid the same problem, the 1-hour windows were sequential since they generated enough data.

Lastly, in the many versus many case, the convergence of the models' scores is not very obvious from training with a maximum of 12 authorised persons. Involving more people could bring a better overview of the behaviour of the score.

5 Conclusions

When training one model to distinguish one person from many others, the best scores were recorded when using the Gaussian Mixture model, with the highest being 0.936 AUC. The One Class SVM performed more poorly, having the largest score of 0.785, which is still significant. With both models, the most important transformation on the windowed data was the MFCC and the best window size was the largest one tested, 3 hours.

When training models to distinguish multiple persons from multiple persons, for 2 to 12 authorised subjects, the best approach was using one single model trained with all the authorised persons' data instead of one model per authorised subject. The trend of a decreasing gap between the two approaches while the number of authorised persons increases indicates that no matter the base model, at some point the two approaches might behave the same at around 0.5 AUC. Using the Gaussian Mixture model gave significantly better scores when using the one-model approach, with the highest mean AUC of 0.88 compared to the highest mean AUC of 0.603 of the One Class SVM. The scores were also overall higher with all numbers of authorised subjects. In the multi-model approach, the Gaussian Mixture again had the best results, with a high of 0.64 against a high of 0.541 of the One Class SVM.

In conclusion, using only the heart rate time series from a consumer-grade wearable with the Gaussian Mixture model and the One Class SVM model yields the best results of detecting unknown persons when trying to train the models with a single person. When training with multiple persons, the scores decrease with more persons the system knows. Besides, the approach of training one model per authorised person was less effective than training one model for all known people. Perhaps a different strategy of aggregating the results from multiple models could yield better results. So, to answer how far the selected outlier detectors could be pushed using the methods described in the paper to authorise multiple persons, the best result obtained with 2 known persons is 0.88 AUC and with 12 known persons is 0.71 AUC.

6 Responsible Research

This section ethically reflects on the data collection and other aspects that were taken into account while writing this paper and conducting the actual experiments.

6.1 Data set

The data set, ME-TIME, although it cannot be tied directly to a specific person because of the lack of personal details such as name and address, it still contains the heart rate and step counter of the subject, which is still categorised as personal information that could eventually lead to the origin of the data. Because of this, the procedure for working with such data was very strict. The data set was only saved on local devices, it was not uploaded anywhere online, this includes git repositories and personal cloud storage spaces. On the completion of the research, the data was erased from all the local devices.

6.2 Reproducibility

The results come from real experiments and they can be reproduced by any party with an availability of ME-TIME [2] data set containing a time series of heart rate samples and a time series of step counts. They can be reproduced using the description in Section 2 and the detailed setups in Section 3.

Prior to the research, an NDA was signed, that the data would not be published. So the code itself without the data has limited reproducibility.

References

[1] Rajendra Singh Bisht, Sourabh Jain, and Naveen Tewari. Study of wearable iot devices in 2021: Analysis & future prospects. In *2021 2nd International Conference on Intelligent Engineering and Management (ICIEM)*, pages 577–581. IEEE, 2021.

[2] ClinicalTrials.gov. Machine learning enabled time series analysis in medicine (me-time). https://clinicaltrials.gov/ct2/show/NCT05802563?term=Me-time&draw=2&rank=1. Accessed: 2023-06-05.

[3] Issam Hammad and Kamal El-Sankary. Using machine learning for person identification through physical activities. In *2020 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 1–5, 2020.

[4] Christoph Lipps, Lea Bergkemper, Jan Herbst, and Hans Dieter Schotten. I know you by heart: Biometric authentication based on electrocardiogram (ecg) signals. In *International Conference on Cyber Warfare and Security*, volume 17, pages 135–144, 2022.

[5] S. Molau, M. Pitz, R. Schluter, and H. Ney. Computing mel-frequency cepstral coefficients on the power spectrum. In *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221)*, volume 1, pages 73–76 vol.1, 2001.

[6] Alexa Muratyan, William Cheung, Sayanton V Dibbo, and Sudip Vhaduri. Opportunistic multi-modal user authentication for health-tracking iot wearables. In *The Fifth International Conference on Safety and Security with IoT: SaSeIoT 2021*, pages 1–18. Springer, 2022.

[7] Douglas A Reynolds et al. Gaussian mixture models. *Encyclopedia of biometrics*, 741(659-663), 2009.

[8] Bernhard Schölkopf, Robert C Williamson, Alex Smola, John Shawe-Taylor, and John Platt. Support vector method for novelty detection. In S. Solla, T. Leen, and K. Müller, editors, *Advances in Neural Information Processing Systems*, volume 12. MIT Press, 1999.

[9] David MJ Tax and Robert PW Duin. Support vector domain description. *Pattern recognition letters*, 20(11-13):1191–1199, 1999.

[10] Sudip Vhaduri and Christian Poellabauer. Wearable device user authentication using physiological and behavioral metrics. In *2017 IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, pages 1–6, 2017.