Using IMU data from earable sensors in the process of user authentication

MATTHIJS VAN DER BENT, University of Twente, The Netherlands

Earables are a relatively new tool in mobile computing. They are more and more used wirelessly to connect to phones and other devices. Listening to music comes to mind. But they can also be used as sensors to collect data. This data can be used for interesting purposes. User experience can be more seamless when earables become smarter and can make decisions based on the data they gather. But could the information that they collect also be used to make assumptions about the users' identity? That is the focus of this research. It explores methods to analyse movement data from earables and machine learning is used in this process. The research gives insight into to what extent the IMU data can be useful in determining the identity of the user and thus providing necessary information for the authentication process. Different machine learning algorithms are be implemented and tested for their suitability.

Additional Key Words and Phrases: User authentication, earable, sensing, machine learning

1 INTRODUCTION

Mobile devices are used more than ever. More and more people and systems have become reliant on them. As a result, many devices contain sensitive information. And as devices become more adapted to users, the demand for new methods of authentication sparks interest. Rather than traditional methods like passwords, authentication based on inertial measurement units may be able to provide a more seamless user experience. A master thesis written by Guse concluded [2] that further research and work had to be done, but that their research showed promise and potential for the use of gestures in mobile devices as means to authenticate a user. Also, other biometric features have proven useful in authenticating users. Research has shown [5] that using data generated from walking can be a way to accurately identify people.

Sensor data is often used to determine which activities or gestures are being performed by users. In those scenarios the question by whom this gesture is performed is not asked as it is often already known. But this brings the question if the data that is collected as a result of performing a gesture can be used to make a decision on the identity of the performer. Research has already been done into how static bodily characteristics can be used to identify people [4]. But in this paper we use data obtained from people who are performing head gestures while wearing earable sensors.

1.1 Objective and research questions

The objective of this research is to discover if, given a set of performed gestures, a user can be accurately authenticated. This will be done by answering the following research question and sub research questions:

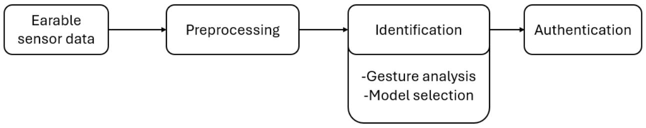

Fig. 1. Schematic overview of this research.

- Main RQ: How accurately can people be authenticated by machine learning models using IMU data collected from earable sensors?
- Sub RQ1: Which gestures are the most useful in determining the identity of the user?
- Sub RQ2: Which machine learning algorithm yields the highest accuracy in authenticating a user?

By answering these questions, we can find out how well the data generated by gestures can be used to authenticate a person. If it proves to work out successfully, it may mean a smaller need for extra data besides what the IMU provides. Other research projects often use extra or different sensors, and merely measuring the extent to which only IMU data provides accurate authentication results is not entirely clear yet.

The way these research questions are answered is by firstly extracting useful data from the large dataset by creating and selecting features and grouping these into a combined dataset. Also, different settings in window size and overlap are explored to find a good way of doing this. With the conclusions that are drawn from the initial results, we can continue to perform analysis. When the data is ready for analysis, different classification algorithms are used to make predictions on the identity of people and which gestures are most suitable to predict this. The best performing algorithm in terms of the chosen performance metrics is used for the user authentication part where a model checks if the person is actually who they claim to be. The structure can be seen in figure 1. It will be shown firstly that the Random Forest and xgBoost algorithm perform best in the identification process. Then, they are used in authentication, scoring error rates of less than 0.05 and 0.1 respectively.

2 RELATED WORK

Research on authentication using facial signals has been done [8]. In this research project, they used earable sensors. Instead of IMU data from the earable itself, they capture a signal coming from the mouth called ToothSonic. The objective of this project is to authenticate a user based on facial signals. Though the location of the sensors and the observed data is different, the results show that it is definitely

TScIT 41, July 5, 2024, Enschede, The Netherlands

possible to use teeth movements in the process of user authentication while making use of earables to collect the data.

A project that classified gestures [3] presented a system that collects and classifies IMU data that is collected through earables. Classification algorithms were applied to determine the activity that the person was performing. Multiple classification algorithms delivered an accuracy of more than 90 percent. This clearly shows that a relation between a person and activity can be predicted by using sensor data. Though it may not necessarily be the case that this process works the other way around with similar results.

In another research project (2024) [7], earbuds were also used to collect IMU data. However, the vibrations are induced with facial touching interaction with the earbuds which is slightly different from performing gestures. Still, the study showed very good results with an error rate of 0.0003 percent. This shows great potential for the analysis of gesture data.

A system has been developed that also makes use of IMU data [5]. This project shows that it is possible to make use of IMU data in the process of user authentication. However, the system is more advanced by again having sensors in the mandible of the observed users. The resulting vibrations are measured and used in functions that authenticate the user with low error rates. Models are constructed from the sensor data and a personal profile is created and stored. Then, a so-called MandiblePrint is generated and that print is compared with the input data. Based on the similarity between those, the decision was made to authenticate the user or not. Also the possibilities of defending possible attacks are explored in this project.

3 METHODOLOGY

3.1 Tools and dataset

For this research, an existing dataset will be used [1]. This dataset contains observed accelerometer and gyroscope sensor data. Both sensors capture data at 100Hz. Each movement was performed for 2 minutes. The data is collected on 30 people. For this research, the following gestures are analyzed: chewing, chin raise, nod, shake, speaking and tilt. Python scikit-learn will be used for the data analysis and machine learning tasks. The dataset description provides us with the information that first, the gesture is demoed to the person. Which is then followed by the gestures being performed continuously for another 2 minutes. All while being in a seated upright position.

3.2 Data preprocessing

For the preprocessing of the data, features had to be extracted. The dataset contains recordings for a total of 2 minutes per gesture for each individual person. A movement could take more than a few seconds and has pauses in between them. Therefore the window size can be important in the process. Different window sizes are briefly compared. In the final dataset that is used for analysis, each person is labeled with a unique id.

3.3 Features

A new dataset is generated that contains all the features with their corresponding values. Each of the individual data subjects is accessed, the values for the intended activities are read. Then, a window is constructed by taking a specified amount of sensor readings and each window is filled with the calculated values. Different window sizes were briefly compared to confirm that the chosen one second was

In the dataset, along with the numerical features, there is a textual column that contains information about the movement of that specific data entry. Since the classification algorithms do not work well with text, they were converted using one-hot encoding. This means that to each gesture, a 1 is assigned if it is the gesture currently being performed and a 0 if it is not. To get more insight in which or how many features are useful, a brief analysis is done using the best performing classifiers in the identification phase. For this test, a set of features consisting of the mean, maximum and minimum values was considered to present the standard configuration. Then, other features namely the standard deviation, skewness and median were added to see if they improved the scoring metrics.

3.4 Models and analysis

For the development of the models, the dataset was split into a training and testing part. The ratio that was used was 75 percent training and 25 percent testing data across all models. The classification algorithms used are Decision Tree, Random Forest (RF), Support Vector Machine(SVM), Naive Bayes (NB), K-Nearest Neighbors (KNN), and xgBoost.

3.4.1 *Identification.*

One way to test the capabilities of the models is by using them to identify people. In this situation, each model is trained with all data and is then presented with an instance of it. The performance metrics accuracy, precision, recall and f1-score then determine how well the algorithm performs at identifying a user. For this task, it is relevant to say that the dataset is balanced. Each of the 30 people in the dataset is known by a unique id, the algorithm is trained and then tested to correctly identify a person. This was then done for each of the models and then we compared the best of them based on their accuracy, precision, recall and their f1-score.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$F1 = \frac{2 * TP}{2 * TP + FP + FN} \tag{4}$$

Where TP (True Positive) and TN (True Negative) mean that a prediction made by the model are in line with the true value. Which would mean a correct authentication. Where FN (False Negative) means that a model actually thinks the user is not who they claim to be and will be rejected, while they were actually correct and falsely rejected. The opposite FP (False Positive) occurs when a user is authorized while this should not happen. Since the latter can be the most harmful in the process of authentication, we would look at the FP rate as an important measure of possible vulnerability. To find an answer to which gesture is most informative when it comes to classifying who is doing it, the dataset is separated into subsets. Each subset contains then the data for a specific gesture on which the model will be trained. This gives information about which gesture is most useful in determining the person performing it.

3.4.2 Authentication. Once it was known which algorithms were best suited to determine identities based on the sensor data, we used those models to test the authentication capabilities of the models. The way it works is that for each of the people in the dataset, a model will be trained. This model basically learns the features of the movements of that specific person during its training. This model can be seen like a password. All data gets fed to each model, while in the dataset being labeled as person that should be authenticated or person that should be rejected (1 or 0 respectively). The classes in the dataset in this scenario are divided in a binary manner. For example, when we look at user with id number 4, this id gets labeled as 1 (being the true user) and all other id's are labeled with 0 (not being the true user). The metrics that are measured in this scenario are the false rejection rate (FRR) and the false acceptance rate (FAR). Where false rejection means that a model actually thinks the user is not who they claim to be and will be rejected, while they should have been authenticated and thus were falsely rejected. The opposite, false acceptance, occurs when a user is authorized while this should not happen. In addition, we will make use of Synthetic Oversampling Minority Technique (SMOTE), which is used for imbalanced datasets. Since the dataset is imbalanced when the labels are converted to binary ones, this technique creates more samples to bring balance in the training data which the model will learn on. The testing data stays unaffected. The class distribution before applying this technique was all samples divided by 30, which means the false instances outnumbered the true instances with a 30 to one ratio in the training data. Afterwards, it was more evenly distributed with a 50/50 ratio between the binary values. In addition, K-fold cross validation was applied with 10 splits.

$$FRR = \frac{FN}{FN + TP} \tag{5}$$

$$FAR = \frac{FP}{FP + TN} \tag{6}$$

4 RESULTS

4.1 Dataset formation

Tables 1-6 provide an overview of the different combinations used regarding the amount of measurements in a window and the overlap of those windows.

Table 1. Decision tree accuracy results

	50 measurements	100 measurements	150 measurements	200 measurements
0 overlap	0.85	0.86	0.87	0.87
0.20 overlap	0.86	0.85	0.88	0.88
0.50 overlap	0.87	0.88	0.9	0.89

Table 2. Random forest accuracy results

	50 measurements	100 measurements	150 measurements	200 measurements
0 overlap	0.95	0.95	0.95	0.94
0.20 overlap	0.96	0.95	0.96	0.95
0.50 overlap	0.97	0.97	0.97	0.97

Table 3. SVM accuracy results

	50 measurements	100 measurements	150 measurements	200 measurements
0 overlap	0.73	0.73	0.73	0.72
0.20 overlap	0.74	0.73	0.73	0.74
0.50 overlap	0.75	0.75	0.75	0.76

Table 4. Naive Bayes accuracy results

	50 measurements	100 measurements	150 measurements	200 measurements
0 overlap	0.56	0.57	0.57	0.61
0.20 overlap	0.56	0.57	0.59	0.61
0.50 overlap	0.55	0.58	0.59	0.60

The metrics in table 1 to 6 show the accuracy for each classifier algorithm when they identify people in the dataset using different window sizes and overlaps. A review of classifiers [6] is used to help understand these results. The reason that the other metrics are not mentioned in these tables is that they are very close to the accuracy in each table and would take up much space without adding much. Two outliers in terms of low scores can be noticed, Naive Bayes and SVM. Naive Bayes is the lowest performing algorithm in terms of the performance metrics. This is likely explained by the fact that this algorithm assigns similar values to a class and then selects a class by the values it consists of. The issue in this case may be that there are a lot of values in the dataset that are somewhat similar and therefore may confuse the algorithm. The Support Vector Machine is the second lowest scoring algorithm. SVM is a difficult algorithm to understand, but it generally performs less with noisy data that can be tough to categorize. This could be an explanation for the relatively poor performance of the algorithm.

Given these results, we move on to use xgBoost and Random Forest for the rest of the classification tasks to see how they perform there. Given that we can also choose how the dataset is shaped, we use the approach with a 1 second window (100 readings as the sensors run on 100Hz) and having them overlap each other by 50 percent. It is clear to see that using the overlap improves the accuracy in nearly all situations. That means that the identification can be performed with a maximum accuracy of 98 percent by the xgBoost classifier.

4.2 Features

In figure 1 can be seen that the adding of more complex calculations of window features does not necessarily mean performance improvements. Especially when all of the extra features are added to the standard set of minimum, average and maximum. The data may

Table 5. K-nearest Neighbors accuracy results

	50 measurements	100 measurements	150 measurements	200 measurements
0 overlap	0.93	0.92	0.92	0.91
0.20 overlap	0.93	0.92	0.93	0.92
0.50 overlap	0.96	0.95	0.96	0.95

Table 6. xgBoost accuracy results

	50 measurements	100 measurements	150 measurements	200 measurements
0 overlap	0.96	0.95	0.94	0.94
0.20 overlap	0.97	0.96	0.96	0.95
0.50 overlap	0.98	0.98	0.98	0.97

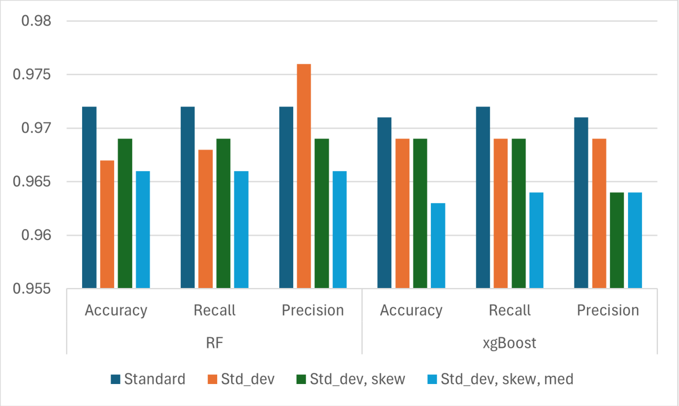

Fig. 2. Overview of the effect of adding specific features to the data.

become too complex and give the algorithms too much to decide on. This could mean that the model prefers a lower number of features for this dataset, or that these specific features are not informative enough.

4.3 Identification results

In table 7, we can see that both algorithms perform more or less similar. It seems that chin raise is the most informative gesture out of the selected ones.

The final identification results including the performance metrics can be seen in Table 8. Here, the xgBoost scores a little higher than Random Forest on all metrics. The accuracy, precision, recall and f1-score are determined to be around 97 percent for Random forest and 98 percent for xgBoost.

4.4 Authentication results

Initial results showed that the FRR was relatively high, over 0.2 for some classes. In that situation, the FAR was very low, less than 0.01. This meant that the model had to be slightly adjusted. A fair balance had to be found between both error rates. A decision parameter was introduced to adjust on which certainty the model should act. As it was deemed too strict in with the standard configuration, its value was lowered to 0.2.

The intention of the setup of this algorithm was to keep the FAR less than 0.01. The FRR knows large differences between the classes. This might be explained by some persons making certain gestures in similar ways. It seems that other persons have very unique ways of performing the gestures which makes them very recognizable.

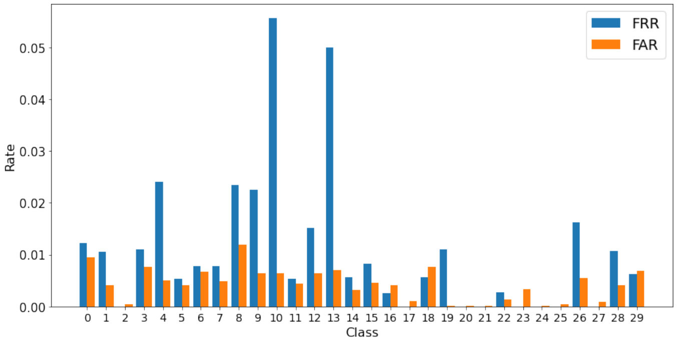

Fig. 3. Random Forest authentication performance.

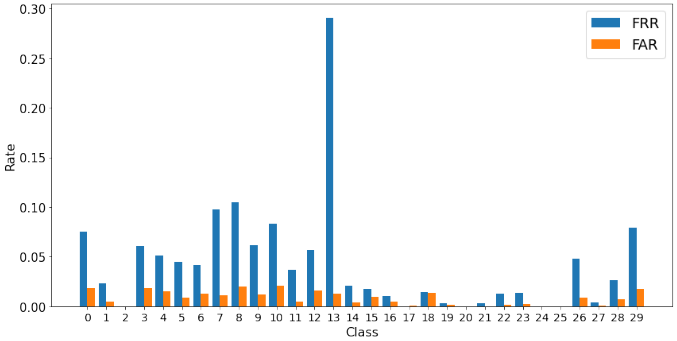

Fig. 4. xgBoost authentication performance.

The xgBoost algorithm shows a more stable FRR with one exception. Person 13 seems to be difficult to authenticate for both models. The xgBoost algorithm can be concluded to be more error prone, despite being more stable. Also, most FRR values stay within the 0.1 boundary, but the Random Forest shows no values above 0.05.

Figures 5 and 6 show the average results of the binary classification as authentication method. Due to the setup of the models, the rate at which negative instances are classified as positive is 1.57 and 3.93 for Random Forest and xgBoost respectively. This is because the false rejections were seen as less problematic as opposed to false acceptances. The latter would mean a breach of security. Thus, the results show good potential for using the algorithms in the process of user authentication.

5 DISCUSSION

As mentioned in the related work section, one research project performed a similar task but instead of classifying persons, they used classification tools for determining movements. An accuracy of over 90 percent was achieved by multiple models. Where classifiers in this research scored up to 8 percent higher than that in the identification part.

In this research, no further testing on real people was done. The data from the used dataset is therefore the only reference material

	Chewing	Chin raise	Nod	Shake	Speaking	Tilt
xgBoost	0.9918	0.9941	0.9861	0.9893	0.9242	0.9648
RF	0.9912	0.992	0.984	0.9883	0.9123	0.968

Table 7. Each gesture analyzed individually.

	Accuracy	Precision	Recall	F1-score
xgBoost	0.98	0.98	0.98	0.98
RF	0.970	0.971	0.970	0.970

Table 8. Results of identification using the classifiers.

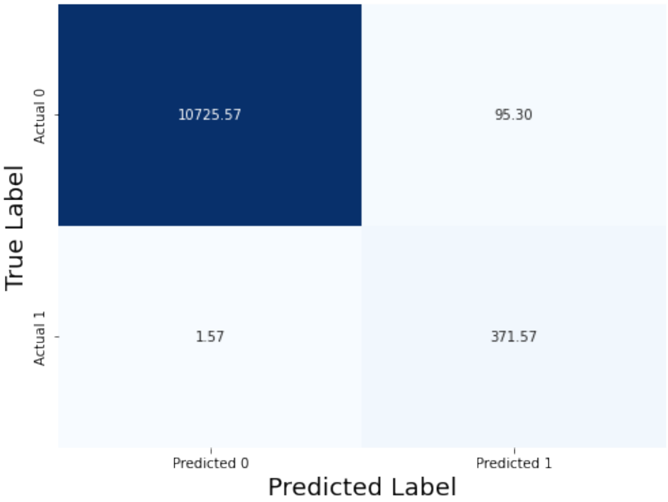

Fig. 5. Average binary classification results using Random Forest.

available. It could be interesting to see if it is possible for a person who was not in the dataset to try and get authenticated by the system. Perhaps by imitating characteristic movements, it may be possible to pose a threat to the classification system.

It is also not clear if this data is sufficient to be used on a longer term or regular basis. It could for example be possible for a person to perform the gestures in another way than as they did in this particular dataset. The situation would then change and the models may have to be retrained. Automating this process could be interesting to work on in the future. This could then be a part of a system or application that combines all aspects and can then be used for authentication testing to perform more tests and research.

Future research can be done to see where the limit lies in terms of performance metrics. Collecting more datasets will improve the diversity of the data and can make the classifiers perform better.

6 CONCLUSION

We started with using the dataset collected with earable sensors making use of accelerometer and gyroscope data. The data from all persons was processed by extracting features and combining all resulting smaller dataset into a larger one. Then, analysis was done to find the optimal setup for the final file and which classifiers should be used for authorization and identification. During the identification stage, it became clear that not all gestures were equally informative. In the authentication phase, Random Forest performed the best with lower error rates on average.

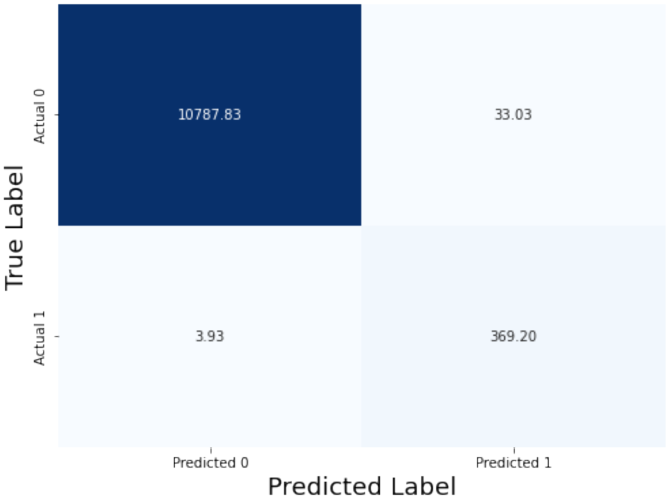

Fig. 6. Average binary classification results using xgBoost.

To answer which gesture is most useful in determining a person's identity, it can be said that the best one seems to be the chin raise, followed closely by chewing with both being over 99 percent accurate. The worst performing gesture is speaking with over 92 percent and 91 percent when using xgBoost and Random Forest respectively.

The question which machine learning algorithm performs best in identifying users gives a convincing answer. It is clear that xgBoost performed the best and can be considered the best algorithm out of these. But using the Random Forest classifier gives nearly similar results. All in all, the Random Forest machine learning algorithm is most suitable for authentication and also scores high on identification.

REFERENCES

[1] Andrea Ferlini, Alessandro Montanari, Ananta Narayanan Balaji, Cecilia Mascolo, and Fahim Kawsar. 2023. EarSet: a Multi-Modal In-Ear dataset. https://doi.org/10.5281/zenodo.8142332

[2] Dennis Guse. 2011. *Gesture-based User Authentication on Mobile Devices using Accelerometer and Gyroscope.* Technical Report.

[3] Shayla Islam, Tahera Hossain, Md. Atiqur Rahman Ahad, and Sozo Inoue. 2020. *Exploring human activities using eSense earable device.* 169–185 pages. https://doi.org/10.1007/978-981-15-8944-711

[4] Bhavesh Kumar Jaisawal, Yusuf Perwej, Sanjay Kumar Singh, Susheel Kumar, Jai Pratap Dixit, and Niraj Kumar Singh. 2023. An Empirical Investigation of Human Identity Verification Methods. *International journal of scientific research in science, engineering and technology* (1 2023), 16–38. https://doi.org/10.32628/ijsrset2310012

[5] Jianwei Liu, Wenfan Song, Leming Shen, Jinsong Han, and Kui Ren. 2022. Secure User Verification and Continuous Authentication Via Earphone IMU. *IEEE transactions on mobile computing* (1 2022), 1–15. https://doi.org/10.1109/tmc.2022.3193847

[6] Pratap Chandra Sen, Mahimarnab Hajra, and Mitadru Ghosh. 2019. *Supervised Classification Algorithms in Machine Learning: A Survey and review.* 99–111 pages. 11 https://doi.org/10.1007/978-981-13-7403-6\{_

[7] Yong Wang, Tianyu Yang, Chunxiao Wang, Feng Li, Pengfei Hu, and Yiran Shen. 2024. BudsAuth: towards Gesture-Wise continuous user authentication through earbuds vibration sensing. *IEEE internet of things journal* (1 2024), 1. https://doi.org/10.1109/jiot.2024.3380811

[8] Zi Wang, Yili Ren, Yingying Chen, and Jie Yang. 2021. Earable authentication via acoustic toothprint. https://doi.org/10.1145/3460120.3485340