Searching Efficient Models for Human Activity Recognition

Shamisa Kaspour*
Nikhil Raj*
Alankrit Mishra*
Abdulsalam Yassine*
Thiago E. Alves de Oliveira*
skaspour@lakeheadu.ca
nraj@lakeheadu.ca
amishra1@lakeheadu.ca
ayassine@lakeheadu.ca
talvesd@lakeheadu.ca
Lakehead University
Thunder Bay, Ontario, Canada

ABSTRACT

Human Activity Recognition (HAR) can be measured in various ways in a new era of growing technologies. This paper studies different classification and data processing tasks. This paper proposes using HAR to monitor the elderly while being power-efficient and respecting an individual's privacy, allowing it to be run on mobile devices like smartphones or smartwatches. Upon reviewing other methods of HAR by sensor data, we realized that they severely lacked in the areas mentioned earlier. Moreover, they either used older classification techniques or made too complex and over-the-top models for the same. We tested a total of nine methods to find the best model/method, from simple support vector machines (SVM) and convolutional neural networks (CNN) to hybrid models. The best results were produced by a simple, fully connected network (multi-layer perceptron) with the data condensed using Fisher's linear discriminant analysis (FLDA) that gave us 98.6% accuracy. Our final model satisfies both the requirements we had set; it is simplified and produces benchmark results.

CCS CONCEPTS

• **Applied computing** → Health care information systems; Health informatics; • **Human-centered computing** → Ubiquitous and mobile computing systems and tools; Mobile phones.

KEYWORDS

Human Activity Recognition, Pattern Recognition, Smartphone Sensor Data, Linear Discriminant Analysis

ACM Reference Format:
Shamisa Kaspour, Nikhil Raj, Alankrit Mishra, Abdulsalam Yassine, and Thiago E. Alves de Oliveira. 2021. Searching Efficient Models for Human Activity Recognition. In *2021 6th International Conference on Biomedical Imaging, Signal Processing (ICBSP '21), October 29–31, 2021, Xiamen, China.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3502803.3502809

1 INTRODUCTION

In the context of this work, Human Activity Recognition (HAR) is the term for classifying the movements of a person based on the data gathered from sensors. Time series data provides information on an individual's behaviours through time and may be evaluated in a variety of ways. While time-series data could be collected for a particular attribute, in some cases, it is generally a good idea to collect data for multiple attributes to be able to perform a better analysis.

In these applications, the data from these sensors often needs to be analyzed using Signal Processing methods to extract features these signals. Following this approach, streams of data are usually divided into windows that can represent parts of an activity. Then, the features extracted from these windows can be used in a machine learning classifier models. In the instance of this study, the goal of the analysis is to distinguish between various activities performed by different participants. These activities consist of walking, laying, standing, sitting, etc [3].

Current monitoring systems are expensive and need an external power system. These monitoring systems contain cameras and microphones, which require extensive upkeep whenever even a minor issue arises. Nowadays, image- and audio-based recognition systems could yield higher accuracy rates at the expense of an individual's privacy, which is increasingly a major concern when implementing and deploying these types of applications.

The data considered in this work is less invasive and was collected from inertial measurement units on smartphones. Most works on HAR from smartphone data do not consider computational power and battery constraints, many times proposing complex deep learning models. The aim of this work, on the other hand, is to investigate models and identify lightweight alternatives that could be deployed in battery-powered devices, such as wearables. Use cases for systems employing these HAR models include

[*] All authors contributed equally to this research.

ICBSP '21, October 29–31, 2021, Xiamen, China
© 2021 Association for Computing Machinery.
ACM ISBN 978-1-4503-8581-7/21/10…$15.00
https://doi.org/10.1145/3502803.3502809

the tracking elderly, Alzheimer's patients, and other populations that could benefit from having their physical activities monitored constantly.

The next section presents a brief literature review on HAR systems that also use smartphone data. Section 3 discusses the dataset, preprocessing steps, dimensionality reduction methods, machine learning models, and model design plan. Section 4 presents the details of the experiments carried. Section 5 compares the results achieved by previous works on the same dataset and the models investigated in this paper. Finally, Section 6 brings conclusion, discussion and future research directions for the present work.

2 LITERATURE REVIEW

Human Activity Recognition emerged as a new scope of smart pattern recognition over the last 8 years. The accessibility of data is not only limited to textual, number or audio/visual but also now can use sensor-based data to predict certain tasks. With the help of booming technology of smartphone, the paper [1] came up with a way to record sensor data of smartphone when it attached to a human body while performing some daily tasks.

The author of [1] observed different changes in body acceleration, gravity acceleration and few other sensor points and developed a baseline model to predict Human activity with the help of standard Support Vector Machine (SVM) that exploits fixed-point arithmetic for computational cost reduction.

A comparison with the traditional SVM shows a significant improvement in terms of computational costs while maintaining similar accuracy, which can contribute to developing more sustainable systems for Ambient Intelligence [1].

Our paper focuses on a single type of Human Activity Recognition, i.e. sensor-based Activity Recognition. It seeks the profound high-level knowledge about human activities from multitudes of low-level sensor readings. Conventional pattern recognition approaches have made tremendous progress on HAR by adopting machine learning algorithms such as decision tree, support vector machine[1], naive Bayes, and hidden Markov models[5]. It is no wonder that in some controlled environments where there are only a few labelled data or certain domain knowledge is required (e.g. some disease issues), conventional PR methods are fully capable of achieving satisfying results. However, in most daily HAR tasks, those methods may heavily rely on heuristic handcrafted feature extraction, which is usually limited by human domain knowledge [4].

For, the model to be more generalized [9] introduced the approach Deep learning to the pipeline of HAR to make the model more robust for future usage.

In the paper [7], they used a Deep Convolutional Neural Network (Convnet) for Human Activity Recognition. By time-series sensor data, they could translate 1D signals into known activities. Also, additional convolution and pooling layers in Convnet accelerate the process of feature extraction of these signals. Although, the more layers increase, the less difference would be in feature complexity level.

They examined their method on "Human Activity Recognition with Smartphones Dataset" and achieved 94.79% for accuracy of classification. One of the most important problems in previous papers was the classification of moving activities. Convnet can easily recognize these kinds of activities and properly classify them.

Human activities are hierarchical since complex activities consist of simple ones. Besides, these activities are translation invariant because different people have their own way to do one action.

In Convnet, they used softmax as their activation function, and for weight update and error cost minimization, they applied stochastic gradient descent (SGD) to mini-batches of sensor training data. Since large weight can cause the vector of weights to be stuck in a local minimum and utilized a regularization method which is called L2 regularization. In this method, they add a term to the cost function to compensate for the large weights. Moreover, dropout in this architecture caused avoiding the overfitting. Dropout temporarily removes some nodes that force the neurons not to rely on the presence of other particular neurons. They applied dropout only to the fully connected layer. Based on the greedy-wise tuning of hyperparameters approach, they started to calculate the performance by increasing the layers of Convnet. They adjusted the number of feature maps, filter size, and pooling size in that order. The maximum number of layers was 4. Increasing the filter size raised the performance while increasing the pooling size did not have any potential effect on the performance. Besides, in the validation set, increasing the performance between layers 3 and 4 is small. Convnet performs very well with the accuracy of 93.7% and outperforms other state-of-the-art approaches [7].

The authors of [8] proposed an approach to detect 6 different actions of a human being, using data collected from a smartphone. They used Support Vector Machines (SVM), to classify and identify the actions. The authors divided their tasks into different modules - Data acquisition & data processing, Feature extraction, Training and Recognition. Data acquisition is done by collecting data and processing signals from the sensors of a smartphone. With the raw data collected, the authors extracted features from it to feed it to a SVM. Data from an embedded accelerometer and gyroscope was collected at 50Hz. Each sensor recorded three values for each of the three orthogonal measurement axes x, y, and z. The raw data acquired by the sensors was organized in windows of 2.56 seconds with 128 recorded values. The authors reported an accuracy of 89.59% for the SVM, which shows a promising research direction for HAR systems using inertial data.

3 METHODOLOGY AND JUSTIFICATION

3.1 Human Activity Recognition Using Smartphones Dataset

The dataset "Human Activity Recognition Using Smartphones" [2] is the object of study in this work. The dataset is publicly available in the UCI machine learning repository. The dataset consists of the recordings of 30 participants performing activities of daily living (ADL) while carrying a waist-mounted smartphone with embedded inertial sensors. The participants volunteering in the experiment were in an age bracket of 19-48 years.

Each person performed six activities (Walking, Walking Upstairs, Walking Downstairs, Sitting, Standing, Laying) wearing a smartphone (Samsung Galaxy S II) on the waist. The data to characterize activities was captured using the smartphone's embedded 3-axis accelerometer and 3-axis gyroscope to measure linear accelerations

Table 1: Model Parameters

Model	Configuration
CNN	1D Convolution(Filters: 64, Kernel Size: 7, Activation: ReLU), 1D Convolution(Filters: 64, Kernel Size: 7, Activation: ReLU), Dropout Layer,1D Max Pooling(Pool Size: 2), Flattening Layer, Dense Neuron Layer(No. of neurons: 512, Activation: ReLU), Dropout Layer, Batch Normalization Layer, Dense Neuron Layer(No. of neurons: 256, Activation: ReLU), Dropout Layer, Dense Neuron Layer(No. of neurons: 128, Activation: ReLU), Dense Neuron Layer(No. of neurons: 6, Activation: Softmax)
SVM	Kernel: RBF Regularization: L2
MLP	Dense Neuron Layer(No. of neurons: 16, Activation: ReLU), Dense Neuron Layer(No. of neurons: 16, Activation: ReLU), Dense Neuron Layer(No. of neurons: 6, Activation: Softmax)

and angular velocities respectively at a constant rate of 50Hz. The signals collected by the sensors were further processed to time and frequency domain features using a 2.56 second window, a complete list of features can be found in the [6].

The experiments have been video-recorded to label the data manually. The obtained dataset has been randomly partitioned into two sets, where 70% of the volunteers were selected for generating the training data and 30% the test data. The dataset has 561 features, 6 activity labels and 10,299 Data sample (7352 in the train set and 2947 in the test set).

The data set was further processed to form a 2-dimensional data frame for training and another one for testing. The training set consist of subset of samples, X_train, and a subset of target activity labels, Y_Train, the X_train and Y_train pair is used as input data in Fig. 1. The test set follows an analogous structure. The training set was further divided into 5 folds for a stratified cross-validation scheme with the objective of estimating how classifiers would generalize to different datasets. The test set was reserved to evaluate the performance of trained models in unseen data.

3.2 Model Design Plan

The input data is encoded to lower dimensionality spaces to reduce the complexity of candidate models and identify lightweight alternatives that could be deployed in battery-powered devices, such as wearables. The Principal Component Analysis (PCA) and Fisher Discriminant Analysis (FLDA) were the transformations used to encode the standardized input data, as seen in Fig. 1. The models trained on the PCA-encoded data were also trained on the raw data to evaluate the dimensionality reduction impact on the models' performance.

The raw data and its PCA/FLDA encoded versions were analyzed by CNN, SVM, and Feed-forward neural network classifiers, as shown in Fig. 1. Table 1 presents the model configurations used in our experiments.

3.3 Fisher Linear Discriminant Analysis (FLDA)

PCA is among the most popular methods for dimensionality reduction. In this method, we find the direction in which data have the largest variance, and consequently, we project the data in that direction. However, PCA does not include the label of the data and it can cause problems. To avoid this issue, Fisher LDA can be used. It also reduces the dimension significantly, which leads to time complexity reduction of our program. In FLDA, we calculate the mean of each class and try to project the data in a way that these means are as far as possible. Thus, it will ease the process of classification. We will try to use FLDA in our data and observe how the data spread is behaving in feature space as compared to our previous implementation i.e. PCA. Fig. 2 presents the X_train samples encoded to the FLDA space, projected on 3 components.

3.4 Principal Component Analysis (PCA)

We can see that the magnitude dimension (i.e. features of the samples) is high. Each sample here is 561 features, from which not all features would be as useful if we consider them individually. To visualise the variance of the data we would be using Principal Component Analysis. This technique will show us how the data is spread according to their principal component. This would also help us understand the useful Principal components that can be used for training the model. By using this approach we can decrease the dimensions of sample data, which will in turn would make the model more efficient to train. It will also us to tackle any imbalance in the data. Fig. 3 presents the cumulative explained variance for per number of principal components for the X_train samples encoded.

3.5 Deep Convolutional Neural Network

According to this paper [7], a Deep convolutional Neural Network (CNN) in HAR can outperform other proposed methods. In fact, CNN, with pooling and convolution layers can easily extract the necessary features for the recognition process. By choosing the proper layers for convolution and pooling, the size of kernels and feature maps, and a fully connected layer for classification, we can achieve a high accuracy to distinguish the movements. According to the above mentioned paper, CNN works better for moving activities. Thus, we try to implement this method to represent the accuracy of these movements.

3.6 Support Vector Machine (SVM)

Support Vector Machines are supervised learning models that are used for classification as well as regression, although, classification is its wider use. It works on linearly separable data.

The goal of a Support Vector Machine is to find a decision boundary with maximum margin. This decision boundary is also a hyperplane, which can be N-dimensional, depending on the data. SVM works on certain data points known as support vectors. These support vectors are the closest data points to the decision boundary and help in generating the decision boundary itself. Things like the position and orientation of the decision boundary is decided by the support vectors.

Support Vector Machines were considered to be the best machine learning techniques before neural networks evolved to the stage

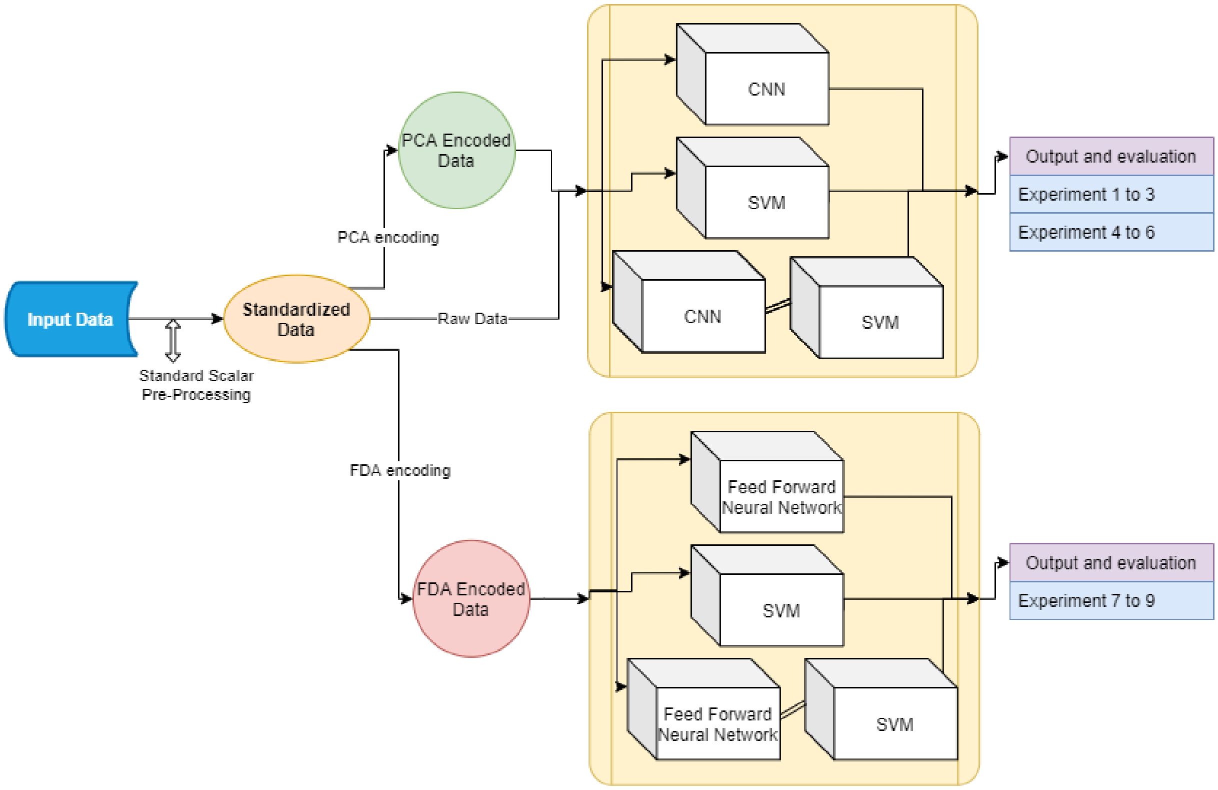
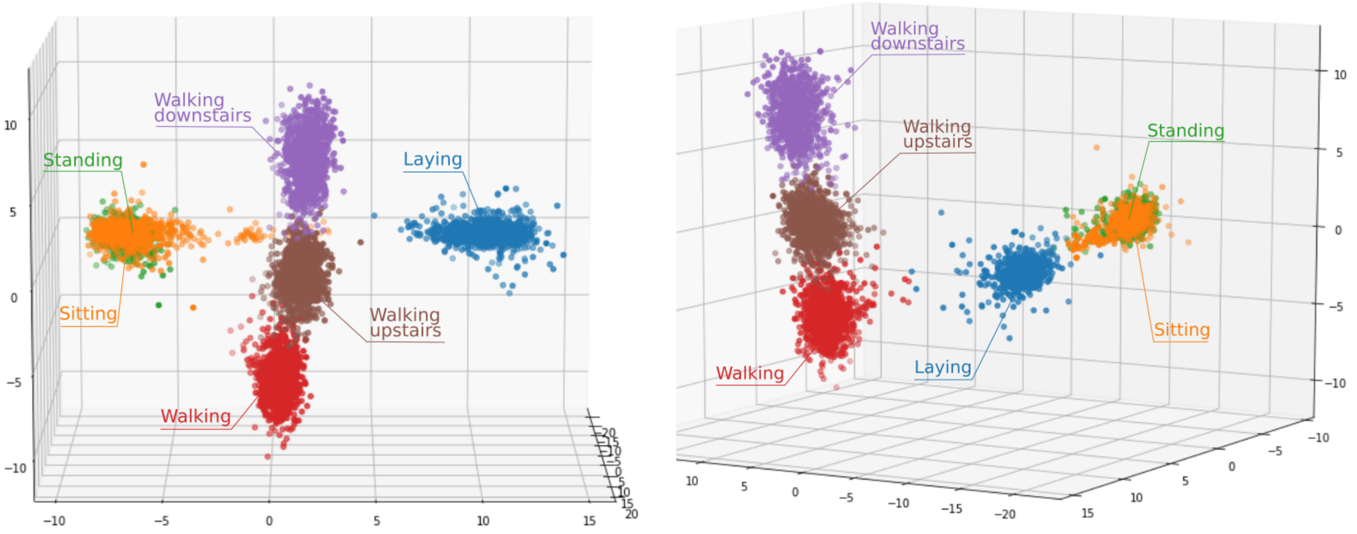


Figure 1: The whole architecture of the project (9 implementations)



(a) 0 degree

(b) 120 degree

Figure 2: FLDA Data spread (Activities according to their colors [blue: laying, orange: sitting, green: standing, red: walking, purple: walking downstairs, brown: walking upstairs])

that they are today. Even today, they provide good accuracy with classification datasets and are widely used for several applications.

We chose SVMs as one of our methods because they offer robust result without a lot of computational resources and two of the

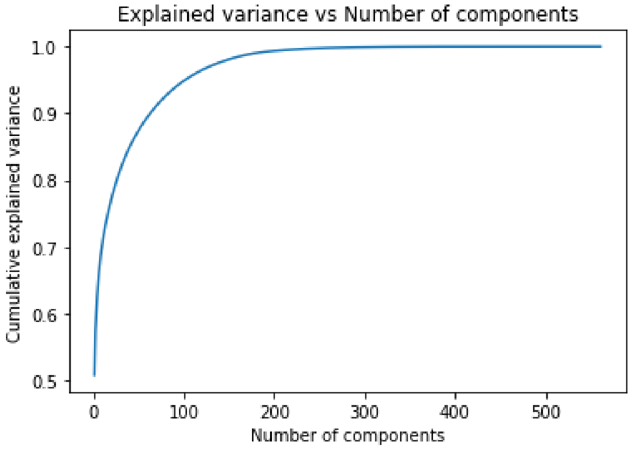


Figure 3: Explained Variance VS Number of Components

papers that we reviewed have used the same and reported great accuracy.

4 EXPERIMENTS

We determined to use PCA and FLDA for data visualization and CNN and SVM for classification in this project. There are nine methods that we executed to explore the performance in each method. The architecture of the whole implementation is shown in Figure 1.

First and foremost, standardization of the dataset results in samples with consistent and comparable features in terms of scales. PCA has been applied to the dataset to decrease the number of dimensions for the samples. Most of the variance of data is at 95% of variance spread (3); thus, we reduced the dimension into the first 102 components of the feature space. Similarly, the data plot produced by the FLDA transformation demonstrates that, despite some overlap across classes, the features are predominantly linearly separable, as shown in 2.

Convolutional Neural Networks are likely to outperform several other machine learning models in classification problems. These networks may be utilized for pictures as well as high-dimensional datasets. Besides, as demonstrated in [7], these networks have achieved the highest precision, recall and accuracy scores in the same dataset subject to this study. To confirm these findings, we apply CNNs to both Raw and PCA-transformed data, (Experiments 1 and 4 from Table 2). It was not possible to execute CNN on FLDA data because this data has five features and will not perform well with complex architectures like CNN. Thus, for FLDA data, we used a simple Feed Forward Neural Network, Experiment 7 from Table 2.

Moreover, since the classes are mostly linearly separable and SVMs are maximum margin methods for classifying linear data, we also applied SVMs to the raw, PCA, and FLDA data (Experiments 2, 5 and 8 from Table 2).

According to recent papers addressing the Human Activity Recognition problem, the highest classification performances are achieved by combinations of SVM and CNN. Following this trend, we extract the deep features of the last layer of CNNs and MLPs, and feed them to the SVMs, Experiments 3, 6 and 9 from Table 2.

All experiments were conducted with 5 fold cross-validation to increase the performance of the models.

Table 2: Results of the Experiments

Model	Precision	Recall	F1-score	Accuracy
HAR-Multi-SVM [1]	0.9	0.89	0.89	0.893
HAR-SVM(2016) [8]	-	-	-	0.895
HAR-ConvNet(2016) [7]	0.94	0.94	-	0.937
CNN Raw Data (Exp.1)	0.97	0.96	0.96	0.964
SVM Raw Data (Exp.2)	0.97	0.96	0.96	0.97
CNN + SVM Raw Data (Exp.3)	0.97	0.97	0.97	0.966
CNN PCA data (Exp.4)	0.92	0.92	0.92	0.923
SVM PCA data (Exp.5)	0.94	0.94	0.94	0.94
CNN + SVM PCA data (Exp.6)	0.93	0.93	0.93	0.926
MLP FLDA data (Exp.7)	**0.99**	**0.99**	**0.99**	**0.986**
SVM FLDA data (Exp.8)	0.94	0.94	0.94	0.94
MLP + SVM FLDA data (Exp.9)	**0.98**	**0.98**	**0.98**	**0.98**

5 RESULTS

Table 2 presents the results for the experiments carried out in this paper and the ones achieved in previous works using the same dataset. The models utilized in 2012 using only SVMs on raw data [1, 8] achieved a baseline accuracy of approximately 89 percent on the HAR smartphone sensor dataset [2]. We can also see that HAR-ConvNet(2016) [7] model obtains a significant improvement in terms of accuracy. But the model used to classify activities in [7] is computationally complex.

In the second part of Table 2 we showed the results and performance of all the nine experiments that have been done during the project. Even though the PCA transformation decreases the dimensionality necessary to describe the data, lowering model complexity, the scores of these models considerably drop when compared to alternative model combinations.

When evaluating raw standardized data, it is also worth noting that SVM performed the best in terms of accuracy.

And finally, experiments 7, 8, and 9 revealed a new benchmark result in terms of precision, recall and accuracy. Experiment 7 (i.e., MLP NN on FLDA encoded data) yields the best performance out of the above three experiments. The computation complexity is much much lower because the FLDA encoded data has only five components. Figure 4 depicts the confusion matrix for each activity classification in experiment 7. Fig. 4 shows that most of the confusion arises between the “Sitting” and “Standing” activities with a minor contribution coming from the “Walking”, “Walking upstairs and downstairs” activities. The “Sitting” and “Standing” are stationary activities while “Walking”, “Walking upstairs and downstairs” involve the same primary motions for limbs and torso.

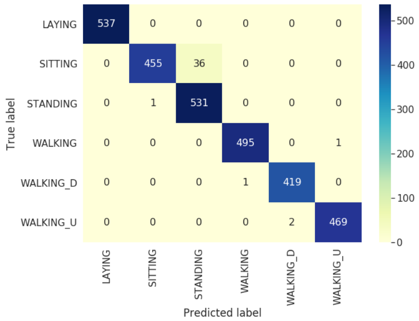


Figure 4: Confusion matrix for multi-layer perceptron on FLDA

The results demonstrate that the less complex MLP Neural network, when applied to FLDA encoded data, outperforms more complex CNN-based alternatives.

6 DISCUSSION AND CONCLUSION

In summary, the goal of this paper is to investigate lightweight models that could be deployed in battery-powered devices, such as wearables, to recognize the different kinds of movements people do in their daily lives.

The nine experiments presented in this work demonstrate that a combination of FLDA transformed data with a multi-layer perceptron classifier may provide superior accuracy (98.6%), precision (99%), and recall (99%) scores compared to other, often more sophisticated, combinations of methods. An SVM classifier applied to features derived from a multi-layer perceptron on FLDA transformed data achieves comparable accuracy, but it is slightly inferior in terms of precision and recall.

When compared to the baseline results achieved by a CNN (0.94 precision, 0.94 recall, accuracy 0.937) presented in [7] for the task of human activity recognition from smartphone data, the results presented in this paper achieve an improvement in all metrics (0.99 precision, 0.99 recall, accuracy 0.986) and at the same time presents a less complex model alternative with the combination of FLDA and MLP classifier.

The cross-validated results also lead to the conclusion that for the dataset subject to this study, FLDA leads to higher classification scores than PCA, with the advantage of transforming the data to a lower-dimensional space that is more favorable to simpler classifiers.

The future work direction will concentrate on deploying the FLDA and MLP to wearable devices in order to evaluate how the system will perform on real-time data. The models from Experiments 7, 8, and 9 will be deployed to low-power wearables to evaluate their impact on battery life. In the future, we are also planning to use our model to evaluate the level of activity on elders, persons with disabilities, and other populations of interest.

REFERENCES

[1] Davide Anguita, Alessandro Ghio, Luca Oneto, Xavier Parra, and Jorge L Reyes-Ortiz. 2012. Human activity recognition on smartphones using a multiclass hardware-friendly support vector machine. In *International workshop on ambient assisted living*. Springer, 216–223.

[2] Davide Anguita, Alessandro Ghio, Luca Oneto, Xavier Parra, and Jorge Luis Reyes-Ortiz. 2013. A public domain dataset for human activity recognition using smartphones.

[3] Oresti Banos, Juan-Manuel Galvez, Miguel Damas, Hector Pomares, and Ignacio Rojas. 2014. Window Size Impact in Human Activity Recognition. *Sensors* 14, 4 (2014), 6474–6499. https://doi.org/10.3390/s140406474

[4] Yoshua Bengio. 2013. Deep learning of representations: Looking forward. In *International Conference on Statistical Language and Speech Processing*. Springer, 1–37.

[5] O. D. Lara and M. A. Labrador. 2013. A Survey on Human Activity Recognition using Wearable Sensors. *IEEE Communications Surveys Tutorials* 15, 3 (2013), 1192–1209. https://doi.org/10.1109/SURV.2012.110112.00192

[6] UCI Machine Learning Repository. 2012. Human Activity Recognition Using Smartphones Data Set. https://archive.ics.uci.edu/ml/datasets/human+activity+recognition+using+smartphones. (Accessed on 07/06/2021).

[7] Charissa Ann Ronao and Sung-Bae Cho. 2016. Human activity recognition with smartphone sensors using deep learning neural networks. *Expert systems with applications* 59 (2016), 235–244.

[8] D. N. Tran and D. D. Phan. 2016. Human Activities Recognition in Android Smartphone Using Support Vector Machine. In *2016 7th International Conference on Intelligent Systems, Modelling and Simulation (ISMS)*. 64–68. https://doi.org/10.1109/ISMS.2016.51

[9] Jindong Wang, Yiqiang Chen, Shuji Hao, Xiaohui Peng, and Lisha Hu. 2019. Deep learning for sensor-based activity recognition: A survey. *Pattern Recognition Letters* 119 (2019), 3–11.