Noname manuscript No.
(will be inserted by the editor)

User Adaptive Models for Activity and Emotion Recognition using Deep Transfer Learning and Data Augmentation

Enrique Garcia-Ceja[1] · Michael Riegler[2,5] · Anders K. Kvernberg[3] · Jim Torresen[4]

Received: date / Accepted: date

Abstract Building predictive models for human-interactive systems is a challenging task. Every individual has unique characteristics and behaviors. A generic human-machine system will not perform equally well for each user given the between-users differences. Alternatively, a system built specifically for each particular user will perform closer to the optimum. However, such a system would require more training data for every specific user, thus, hindering its applicability for real world scenarios. Collecting training data can be time consuming and expensive. For example, in clinical applications it can take weeks or months until enough data is collected to start training machine learning models. End users expect to start receiving quality feedback from a given system as soon as possible without having to rely on time consuming calibration and training procedures. In this work we build and test user adaptive models (UAM) which are predictive models that adapt to each users' characteristics and behaviors with reduced training data. Our UAMs are trained using deep transfer learning and data augmentation and were tested on two public datasets. The first one, is an activity recognition dataset from accelerometer data. The second one, is an emotion recognition dataset from speech recordings. Our results show that the UAMs have a significant increase in recognition performance with reduced training data with respect to a general model. Furthermore, we show that individual characteristics such as gender can influence the models' performance.

Enrique Garcia-Ceja
E-mail: enrique.garcia-ceja@sintef.no

1 SINTEF Digital, Oslo, Norway.
2 Simula Research Laboratory, Oslo, Norway.
3 Department of Informatics, University of Oslo.
4 Department of Informatics and RITMO, University of Oslo.
5 Kristiania University College, Norway.

Keywords transfer learning · user adaptation · personalized models · deep learning · emotion recognition · activity recognition

1 Introduction

The automatic monitoring of human behavior using multimedia data, such as speech, videos or sensor information, has gathered a lot of interest in recent years since it is able to provide contextual information about a user. Being able to monitor human behavior in a continuous and unobtrusive manner, is of special interest for applications in sports (Mitchell et al., 2013), recommendation systems (Soleymani et al., 2018) and health care (Garcia-Ceja et al., 2018; Avci et al., 2010), to name a few. By knowing the current state of a user, personalized assistance and services can be delivered when required. For example, appropriate music play lists can be recommended based on the user's current activity (exercising, walking, working, etc.) (Wang et al., 2012). Elderly people at an early stage of dementia could also benefit from this type of system, e.g., by monitoring their hygiene related activities (wash hands, shower, brush teeth, etc.) and sending reminder messages when appropriate (Richter et al., 2016). The increasing popularity of wearable devices such as smartphones and smartwatches makes them an ideal platform for continuous support and interventions. These devices have several types of sensors like accelerometers, microphones, Wifi, temperature, light, etc. Previous work has shown that machine learning methods can be used to analyze the generated sensor data to infer users' behaviors and mood states (Zenonos et al., 2016; Grünerbl et al., 2015; Sanchez et al., 2015). One of the challenges in automatic behavior monitoring systems is that each person is different and possesses distinct characteristics, thus, a single machine learning *general model* (GM) will not perform optimally on all users. The solution to this is to have specific machine learning models for each person, also called *user-dependent models* (UDMs). A challenge with this approach is that UDMs require a lot of training data for the given user. In some settings, e.g., in the medical field, collecting training data is expensive and time consuming. Therefore, the need for models that do not require too much training data for each specific user becomes important.

An important aspect of interactive and behavior monitoring systems is model personalization. Model *personalization/adaptation* refers to training and adapting predictive models (e.g., classifiers) for a specific user according to his/her own attributes. Building a model with data from many users and using it to predict behaviors for a target user will introduce noise due to the diversity among users. For example, Lane et al. (2011) showed that there is a significant difference for the *walking* activity between two different groups of people (20-40 and $>$ 65 years old). Lockhart and Weiss (2014) showed that there are large differences in performance between GMs and UDMs (GMs perform worse compared to UDMs).

Our main goal is to have accurate models adapted to each users' characteristics that can be trained with limited amounts of labeled data. In this work we propose the use of deep transfer learning and data augmentation via random oversampling to train models that adapt to each user as more data becomes available over time. Our research question is: *Is it possible to adapt a general model to a particular user for activity and mood recognition using deep transfer learning and data augmentation?*. Thus, we aim to test the following main hypothesis for the selected datasets for activity and emotion recognition tasks:

H: For the selected activity and emotion recognition datasets, an adapted model for a particular user using deep transfer learning and data augmentation will perform better than a general model when labeled data is limited.

We start by building a GM, which is then adapted to each user by refining the last neural network layers using small amounts of training data for that particular user. Furthermore, we generated more training data by augmentation using random oversampling. For our experiments, we used two public available datasets. The first one is for activity recognition tasks and the second one is for emotion recognition from speech audio files. We chose those two domains since they exemplify common situations where there is user diversity. Furthermore, these application domains have gained a lot of attention in recent years but limited attention has been paid to user adaptation issues. In both cases, we assume that the within-user data distribution does not change over time, i.e., there is no concept drift (Gama et al., 2014). This assumption is particularly true for the two tested datasets since they were collected within small periods of time. For the activity dataset, each user's data was collected on one day, usually within 1-2 hours. For the emotion dataset, there was a single recording session with every individual for about 2 hours. We also assume that all users have labeled data for all possible classes. For the activity recognition case, we only considered users (18 of them) from the database that performed all activities since some users collected data only for a subset of activities.

The main contribution of this paper is a method based on the combination of deep transfer learning and data augmentation to build user adaptive models for activity and emotion recognition. These two domains represent common multi-user scenarios with different sensor modalities. Even though we demonstrated the applicability of the method on two specific datasets, we believe that this approach can be used for other use cases and types of sensors.

This paper is organized as follows. Section 2 presents the background about different types of models and an overview of transfer learning. In section 3, we present the related work about activity/emotion recognition and adaptive models. Next, in section 4, we describe the approach used to build the UAMs. Section 5 presents the datasets used in our experiments and the details of the preprocessing and feature extraction. In section 6, we explain our experiments and the obtained results and provide a discussion in section 7. Finally, in section 8 we draw the conclusions.

2 Background

When training predictive models that depend on user behavior, there are different strategies: general models, user-dependent models, mixed models and user adaptive models. In the following subsections we will introduce the characteristics of those types of models. Then, we present the background of what *transfer learning* is and its implementation within the context of deep learning.

2.1 Model Types

For interactive systems that involve making predictions based on user behavior, four different types of models can be identified: 1) *General*, 2) *User-Dependent*, 3) *Mixed* and 4) *User-Adaptive* models.

General Models (GM): Also known as *User-Independent Models*, *Impersonal Models*, etc. The advantage of these models is that they do not require data from the *target user* and can be used 'out of the box'. Given a set of users in a database, the GM is built by using all the aggregated data from all users to train a predictive model. When a new *target user* (not in the original database) needs a prediction, it can use the GM. The procedure to validate this type of model is as follows: for each target user u_t the data from all other users u_i, $i \neq t$ are used as the training set and the data from user u_t is used as the testing set. GMs usually perform worse than UDMs because there are some users with far from average behaviour.

User-Dependent Models (UDM): Also called *User-Specific Models* or *personal models.* These types of models are trained using just data from the target user. To estimate their performance, usually k-fold cross validation is performed on each user with her/his own data. Often, UDMs offer the best performance for a specific user, since they capture the specific characteristics of each user. However, they require a lot of training data for each user and in some domains, collecting training data is expensive and time consuming. Apart from that, in the training process overfitting can easily happen and distort the results if not accounted for.

Mixed Models (MM): This type of model does not make any distinction between users. That is, all the data is aggregated without distinguishing users and then, k-fold cross validation is performed. This means that some data points from the same user can end up in both, the training and testing sets. The performance results of MMs may not be representative of how a system will behave or generalize to new observations when different users are involved. This type of setting is common when the data does not depend on human behavior like in object recognition from images, weather prediction, and so on.

User Adaptive Models (UAM): These models are intended to be used by specific users, just as the UDM, but they require less training data. UAMs try to combine the best characteristics of GMs and UDMs. Usually, a GM is first trained and when there are more available training data, it is refined incrementally to increase its performance over time. Section 3 will present the adaptation techniques that have been previously used.

2.2 Artificial Neural Networks

Artificial neural networks (ANNs) are connectionist models mainly composed of nodes (neurons) and edges between them. Neurons receive input signals, process them and produce an output that can serve as the input for other neurons. Usually, a neuron's output is computed using a non-linear function of the sum of its inputs called an activation function. Typical activation functions are: sigmoid, hyperbolic tangent, ReLU, gaussian, etc. Edges have associated weights that are learned during the training process. Neurons are grouped into layers and signals propagate from the input layer to the last layer. Deep neural networks consist of multiple layers of neurons Haykin (1994).

2.3 Transfer learning

Transfer learning is a technique for training machine learning models. The basic idea is based on the concept that humans acquire knowledge and learn new things constantly. People can apply their previous experiences to learn new things and adapt to new situations. The goal of transfer learning therefore is to transfer the knowledge obtained by one task to another task (also through different domains) (Bengio, 2012). Transfer learning is very popular in the domain of image classification and robotics (Alnujaim et al., 2018; Sevakula et al., 2018; Peng et al., 2017; Devin et al., 2017). These methods try to transfer the knowledge from previous tasks to a new task. Often, the new task has fewer high-quality training data (Pan and Yang, 2010). For example, given a classifier trained with reviews for certain products, we may want to build a new classifier based on the previous one in order to use it in new products while there is still not enough labeled data for those new products (Blitzer et al., 2007).

Transfer learning with convolutional neural networks (CNNs) (LeCun et al., 1998; Shin et al., 2016) has become very popular for image classification. Since training CNNs can be a computationally intensive task, the idea is to train a big CNN once and then just fine tune the last layers for specific applications.

The advantage of using transfer learning is related to generalization and the size of the datasets used for training. The training technique is useful for the problem of having small datasets. When training deep learning models, the difficulty lies in achieving good results, while at the same time being able to generalize over different, but similar inputs. In this context, it means that a neural network should be able to achieve good results on similar data (e.g., different users or images) even though it has never seen this data before. For example, a neural network using a dataset of hand gestures from a limited set of users to train should after training be able to detect other, similar gestures from different users. The aim of the training is to go from a state of undergeneralization to a state of generalization. The issue of a small dataset for training is that there are too few samples to be able to reach a good rate of generalization, and the algorithm tends to overfit.

Another advantage of using transfer learning is that training from scratch is very costly in terms of data required and resources to be used. Not only does it require hardware to keep it in memory, but even with powerful hardware the training can take several weeks for big datasets. On the other hand, using transfer learning can reduce the time aspect and the needed data to a small proportion.

3 Related work

In this section we present related work about activity recognition, emotion recognition and finally, model adaptation.

3.1 Activity recognition

In recent years, Human Activity Recognition has become an important research area because of the potential range of applications in different domains such as health care, mental health care, elder care, sports monitoring systems, etc. (Martínez-Pérez et al., 2012; Garcia-Ceja et al., 2018).

Monitoring user activities and assisting them in their everyday lives has great potential in pervasive and health care applications, thus, allowing people to keep living independently and with healthy lifestyles. The objective of these types of systems is to monitor and provide opportunistic assistance automatically. This is of special importance for groups of people who require constant assistance such as elderly people, persons with chronic diseases, mental disorders, etc.

Recently, accelerometers have become very common for activity recognition because its small size and they can be found in many devices such as smartphones, watches, etc. For example, Brezmes et al. (2009) implemented a real time activity recognizer on a mobile phone. This was one of the first works to take advantage of a mobile's phone accelerometer without the need of attaching several sensors to the body. Mannini and Sabatini (2010) used five bi-axial accelerometers located at the hip, wrist, arm, ankle, and thigh and they reported accuracies between 93% and 98.5% for seven different activities (sitting, lying, standing, walking, stair climbing, running and cycling). Mitchell et al. (2013) used smartphone accelerometers to recognize sporting activities. Shoaib et al. (2014) made an extensive analysis of the impact of using just an accelerometer or gyroscope or a combination of both when placing sensors in different parts of the user's body. More recently, López-Nava and Muñoz-Meléndez (2018) used captured motion from upper and lower limbs for daily living action recognition.

3.2 Emotion recognition

Emotion recognition is the task of identifying human emotions. In the machine learning field, this process is usually conducted by analyzing facial expressions and/or speech patterns (Tarnowski et al., 2017; Ayadi et al., 2011).

One of the motivations that has driven a lot of research within this field is human-machine interaction, as pointed out by Chatterjee et al. (2018). Given the potential of these types of technologies, different approaches have been proposed. For example, Lalitha et al. (2014) they recognized seven different emotions based on speech signals by extracting mainly time domain features. The authors use a Support Vector Machine with a Radial Basis Function for the classification achieving a recognition rate of 81%. Lin and Wei (2005) used a Hidden Markov Model which is capable of modeling temporal dependencies. They trained their model to classify five emotions including anger, happiness, sadness, surprise and a neutral state. Their reported recognition rate was 99.5% for the gender independent case. An interesting approach using image representations was recently proposed by Badshah et al. (2017). They extracted image spectrograms and use them to train a Convolutional Neural Network, thus, avoiding the need to generate handcrafted features. They achieved an accuracy of 61.7% per spectrogram for seven emotions and 84.3% per speaker.

Recently, emotion recognition has also been used in the mental health care domain. Given the popularity of smartphones, several authors have proposed speech analysis from phone calls for detecting the mental state of users. Grünerbl et al. (2015) analyzed phone calls from bipolar disorder patients to detect depressive, normal and manic states. They achieved an average recognition accuracy of 70% using a Naive Bayes classifier. Other similar work is by Karam et al. (2014) in which they used a Support Vector Machine. They achieved an Area Under the Curve (AUC) of 0.81 for hypomania and 0.67 for depression. As can be seen, emotion recognition systems are important in many real world use cases.

3.3 Model adaptation

Given the importance of adapting systems to each user's needs, previous authors have proposed different methods in different domains and with different sensors. Table 1 presents a list of previous research that aim to adapt predictive models with reduced training data. From this table, it can be observed that several of them are based on clustering.

For example, Xu et al. (2015) and Garcia-Ceja et al. (2016) used a similar approach based on clustering for stress detection with sensor data. They aimed to find clusters of similar users and train a model with the data just from users within the same cluster. The rationale behind this is that users in different clusters are very different and thus, their data will have a negative impact on the final model. In the former work (Xu et al., 2015), they extracted

Table 1 Related works on user adaptation methods.

Work	Method	Domain	Sensors
Vo et al. (2013)	k-medoids clustering	activity recognition	accelerometer
Abdallah et al. (2012)	incremental and active learning	activity recognition	accelerometer
Fallahzadeh and Ghasemzadeh (2017)	Uninformed cross-subject transfer learning	sports activities recognition	3D motion tracker
Rokni et al. (2018)	CNN transfer learning	activity recognition	accelerometer
Lane et al. (2011)	Community similarity networks	activities, transportation	GPS, phone usage, audio, acceleration, Wifi, etc.
Garcia-Ceja and Brena (2015)	Class similarities clustering	activity recognition	accelerometers
Parviainen et al. (2014)	likelihood distributions	activity and environment recognition	acceleration, GPS, WLAN, Bluetooth, GSM, audio
Xu et al. (2015)	Cluster users	stress detection	EEG, ECG, EMG, GSR
Garcia-Ceja et al. (2016)	Cluster behavioral vectors	stress detection	smartphone's accelerometer
Maxhuni et al. (2016)	Decision tree transfer learning	stress detection	smartphones
Lu et al. (2012)	Maximum A Posteriori	stress detection	audio
Vildjiounaite et al. (2017)	Hidden Markov Models	stress detection	mobile phones and wrist bracelets

statistical features from physiological signals such as electrocardiography, galvanic skin response, electroencephalography, electromyography and saturation of peripheral oxygen. In this case, they got a maximum performance of 0.852 with two clusters. In the later work (Garcia-Ceja et al., 2016), they used the silhouette index to find the optimum number of clusters instead of defining them manually. Instead of physiological signals, they extracted features from accelerometer data collected with a smartphone and trained Decision Trees and a Naive Bayes classifier. The maximum accuracy was 60% with the Naive Bayes adapted classifier trained only with data from similar users. Garcia-Ceja and Brena (2015) also used a clustering approach for activity recognition but the difference is that they performed the grouping in a per class basis rather than clustering users. In their experiments, the authors used a decision tree as classifier and for the same dataset used in the present work (WISDM) they reported an accuracy of around 80% with 30% of adaptation data. Vo et al. (2013) proposed an adaptation algorithm for activity recognition that produced an 11% accuracy increase compared to a general model. Their method relies on k-medoids clustering and a Support Vector Machine that first trains a model using data from user A and then personalizes it for another person B; however they did not specify how should user A be chosen. This can be seen as a 1 → n relationship in the sense that the base model is built using data from a specific user A and the adaptation of all other users is based solely on A. The drawback of this approach is that user A may be very different from all other users which could lead to poor final models. One of the disadvantages of clustering is that as the dimension increases, the notion of distances is lost which will result in poor groupings. Furthermore, when clustering users, one needs to specify how each user will be represented in a feature vector format which requires some expertise, testing and will depend on the dataset.

Lu et al. (2012) proposed an adaptive system for stress recognition from audio data based on Maximum a posteriori. They start with a general model and as more data is available, they update it accordingly. They used two different adaptation schemes, supervised and unsupervised adaptation. In the first one, the new data is explicitly labeled whereas in the latter case, self-training (Yarowsky, 1995) is used to generate the new labels. Their adaptation method was able to achieve better performance (82.9%) compared to the general model (71.3% in indoor settings). Vildjiounaite et al. (2017) proposed a method to build stress detectors with Hidden Markov Models and Maximum a posteriori. They achieved an accuracy of 75% which is similar to the state of the art but only using unlabeled data. Parviainen et al. (2014) also proposed a method based on Maximum a posteriori but for activity recognition. Their system asks the users to only provide binary feedback to indicate if the system's prediction was correct or not, reducing annotation effort considerably. Based on this feedback, their algorithm adapts the classifier parameters to each particular user over time. The precision before adaptation was 50% and after adaptation, it increased to 68%. One of the limitations of Maximum a posteriori approaches is that they are tied to a particular classifier. Abdallah et al. (2012) proposed an incremental and active learning approach for activity recognition to adapt a classification model as new sensory data arrives. The novelty of their method is that when the system is not sure about a prediction it asks the users for the correct label. Lane et al. (2011) built adapted models for activity recognition for each user by first building Community Similarity Networks (CSN) for different data dimensions such as: anatomical similarity, lifestyle similarity and sensor-data similarity. The method consists of finding similarities between users based on different attributes and then weighting them to train a boosting classifier. Their method produced accuracy increases between 9.1% and 46.8% for different datasets. One of the limitations of this approach is that it requires someone to specify similarity functions for each type of attribute.

Transfer learning has been applied in the literature to reduce the amount of labeled data required by a machine learning model. For example, Maxhuni et al. (2016) used decision tree transfer learning for stress detection which consists of building a decision tree for all users and finding the most similar trees for the target user. Then, the users' data corresponding to those trees are transferred to build the final model. Their maximum accuracy was 71.58%. The limitation of this type of transfer learning is that it would require considerable effort to be implemented with another type of classifier. Fallahzadeh and Ghasemzadeh (2017) developed a cross-subject transfer learning algorithm for activity recognition. Their method uses the similarity between the training data and new observations from the target user to adapt and retrain the model. They achieved an accuracy of 87%. Recently, deep learning methods have been demonstrated to produce outstanding results in different tasks such as computer vision, speech recognition, text analysis, to name a few. In this work, we used deep transfer learning for model adaptation. One of the advantages of this approach is that it does not require explicit modeling of the users. Also, deep learning provides flexibility on architectural choices depending on the application while keeping the main training process unchanged. This means that different architectures can be used such as fully connected neural networks, convolutional neural networks, etc. Rokni et al. (2018) have already applied deep transfer learning on a convolutional neural network to personalize human activity recognition models getting an approximate overall accuracy of 95%. This research differs from previous work in the following ways:

1. We conducted several experiments on two different domains: activity recognition from inertial sensors and mood recognition from audio. Most of the works in Table 1 tested their methods on a single task, either activities or emotion with the exception of (Lane et al., 2011; Parviainen et al., 2014).
2. We compared five different types of models for validation and completeness (see section 6). This allowed us to measure the effect of adaptation data in different scenarios as opposed to (Rokni et al., 2018) where they trained different models but all of them already including adaptation data.
3. We tested the effect of using adapted models on other users (which resulted in poor accuracies). This type of evaluation was not conducted in any of the reviewed works (Table 1).
4. We evaluated the impact of training general models by pre-selecting users (same and different gender). This type of analysis was not conducted in any of the reviewed works (Table 1).
5. We evaluated the impact of training user adaptive models with all the data and just data from users with the same gender. This type of analysis was not conducted in any of the reviewed works (Table 1).
6. We evaluated the impact of generating synthetic data via random oversampling and combining it with transfer learning. This type of analysis was not conducted in any of the reviewed works (Table 1).

4 Deep transfer learning for user adaptive models

In this section, we describe how the UAMs are built. As mentioned before, deep transfer learning approaches are commonly used with CNNs for image recognition tasks. Commonly, transfer learning is used to train models to recognize new categories of images. In the present work, we apply transfer learning for user adaptation purposes instead of learning new categories but the idea is the same. The approach consists of having a deep neural network with two types of layers: *fixed layers* and *adaptive layers* (see Figure 1). *Fixed layers'* parameters are learned once and remain the same throughout the network's lifetime. *Adaptive layers'* parameters can change after the entire network has being trained.

The process of building a UAM for a target user u_t consists of training the entire network with all the available data from all other users different from u_t, i.e., a general model. As more data (*adaptation data*) for the target user u_t is available, it is used to fine tune the *adaptive layers* of the network. This is done by training the network through more epochs but without modifying the *fixed layers*.

Sometimes, even the adaptation data may not be enough. An approach to compensate for this is to generate new *synthetic* data. A UAM with data augmentation consists of *expanding* or generating new synthetic instances from the *adaptation data*. In this work we used random oversampling as augmentation technique which consists of copying n instances from the *adaptation data* with replacement.

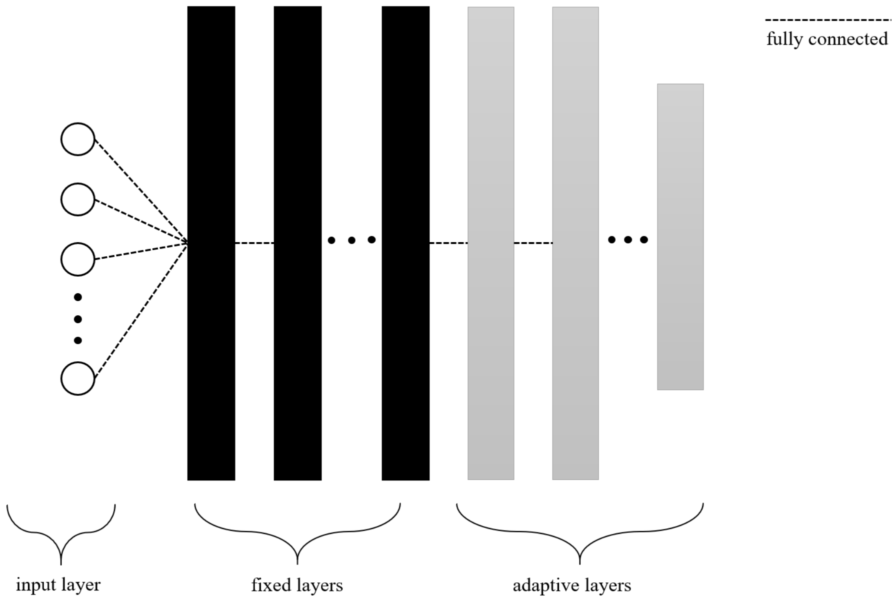


Fig. 1 A general overview of deep transfer learning. Fixed layers do not change during the transfer learning step, whereas adaptive layers do.

5 Datasets and preprocessing

For our experiments, we used two public available datasets: An activity recognition dataset from accelerometer data and, an emotion recognition dataset from speech. The criteria for selecting the datasets were:

1. The domain is representative for user adaptation, i.e, there are user variations that impact the performance.
2. It must contain data collected from several users.
3. The information of which user generated each instance must be included.
4. Each class has sufficient examples per user.

Given these constraints, the following public datasets were found.

5.1 Activity recognition dataset

This dataset was collected by 36 subjects while performing 6 different activities (Kwapisz et al., 2011; Wisdm, 2012). We included only users who performed all 6 activities (18). The data was recorded using the accelerometer in a smartphone with a sampling rate of 20 Hz. The dataset contains 43 features which were extracted from fixed length windows of 10 seconds each. Some of the features are: average, standard deviation, average absolute difference, average resultant acceleration, time between peaks, binned distribution, etc. For the complete list please see (Kwapisz et al., 2011; Wisdm, 2012). One of the features only had zeros and three features had missing values, thus, those four features were discarded. The features were normalized between 0 and 1. The activities include: 1) walking downstairs, 2) jogging, 3) sitting, 4) standing, 5) walking upstairs and 6) walking on flat ground. In this dataset not all the users performed all the activities. We considered only the users that performed all activities (18 users). The total number of instances is 3,153. Table 2 shows the class count.

Table 2 Count of activity classes.

Class	**Count**
Downstairs	326
Jogging	950
Sitting	254
Standing	183
Upstairs	358
Walking	1082

5.2 Emotion recognition dataset

This database consists of audio recorded emotional utterances spoken by 10 actors (Burkhardt et al., 2005; EmotionDB, 1999). The emotions are: happy, angry, anxious, sad, bored and disgusted as well as neutral. The participants were five males and five females. Their age range is 21-35. The database contains about 500 utterances and ten different texts. For the audio feature extraction, we used a python library called pyAudioAnalysis developed by Giannakopoulos (Giannakopoulos, 2015, 2016). The feature extraction process consists of two steps: short-term feature extraction and mid-term feature extraction. The first, splits the input signal into short-term windows and computes a set of features for each frame. In total, 34 short-term features were computed. Some of the features are zero crossing rate, energy, entropy of energy, spectral centroid, spectral spread, MFCCs, chroma deviation, etc. For the full list of features and their description please refer to the work of Giannakopoulos (2015). This process will produce a sequence of short-term feature vectors for the whole signal. Then, mid-term features are extracted from those sequences. The mid-term features were the mean and standard deviation. This resulted in 68 features (mean and standard deviation for each of the 34 short-term features). The features were normalized between 0 and 1. The total number of resulting instances after feature extraction was 3,188. Table 3 shows the class count.

Table 3 Count of speech emotion classes.

Class	**Count**
Anxiety	335
Disgust	327
Happy	387
Bored	480
Sad	524
Anger	724
Neutral	411

6 Experiments and Results

Our experiments were performed for five iterations to account for variability and the average results are reported here. For each iteration, the first 50% of each users' data is used as the test set. The remaining 50% of the data points are used to build the *adaptation set*, i.e., the data points used to adapt the model to a particular user. The reason for choosing the first 50% as the test set was to avoid overfitting since contiguous data points are expected to be very similar. We tested different percentages for the adaptation set. From 10%

to 40% relative to the total number of data points for each particular user with increments of 10 (e.g., see Figure 14). The adaptation data points were chosen at random from the remaining 50% of the total users data. The rest of the points, (from 40% to 10%, respectively) were discarded. The reported accuracies and F-scores were computed by averaging them over all users and iterations. The accuracy is the percentage of correctly classified instances. The F-score is the harmonic mean of precision and recall.

For comparison purposes and as a baseline, we trained five types of models:

1. **GM:** General Model.
2. **UAM:** User Adaptive Model.
3. **GM_all:** Similar to the General Model but updated with additional epochs using all the training data (excluding data from the user under consideration). The purpose of this is to validate whether the performance difference with respect to the UAM is just due to additional epochs.
4. **GM_rand:** Similar to the User Adaptive Model (just the adaptive layers are updated) but the adaptation data is chosen at random from all other users different than the target user. The purpose of this model is to validate whether the performance difference with respect to the UAM is just due to fine tuning the adaptive layers through additional epochs.
5. **RF:** Random Forest. This model was trained *only* with the adaptation data. The purpose of this is to show if the data from other users provides useful information or if the adaptation data is enough to build good models. The reason for choosing a Random Forest is because it is one of the more powerful models and does not require a lot of data. A deep neural network requires more data but the adaptation data is very limited and not sufficient to train such a model.

Figure 2 depicts the first four types of models tested for a particular user u_{n+1}. First, the GM is trained and three different copies of it are used to build the GM_all, GM_rand and UAM. The small circles mean that random samples of the data are used for training. No circles mean that the entire data is used for training.

In order to assess how a UAM of a particular user performs on a different user, we held out n users (5 for the activity and 3 for the emotions dataset) and tested each of the UAMs of all other users on them. One possible approach to compensate for the lack of labelled training data is to generate synthetic labelled data. Thus, we used random oversampling on the adaptation data to generate more labeled examples. Oversampling is the process of randomly choosing and copying data points from a set. In machine learning, this is typically used to generate more instances of the minority class in imbalanced datasets (Kotsiantis et al., 2006). Specifically, we over-sampled at 50% relative to the whole training data from all other users. We used the over-sampled instances to train two additional user adapted models:

User adaptive model with random oversampling (UAM+RO). This is the same as the UAM but the adaptive layers are updated using the over-sampled instances.

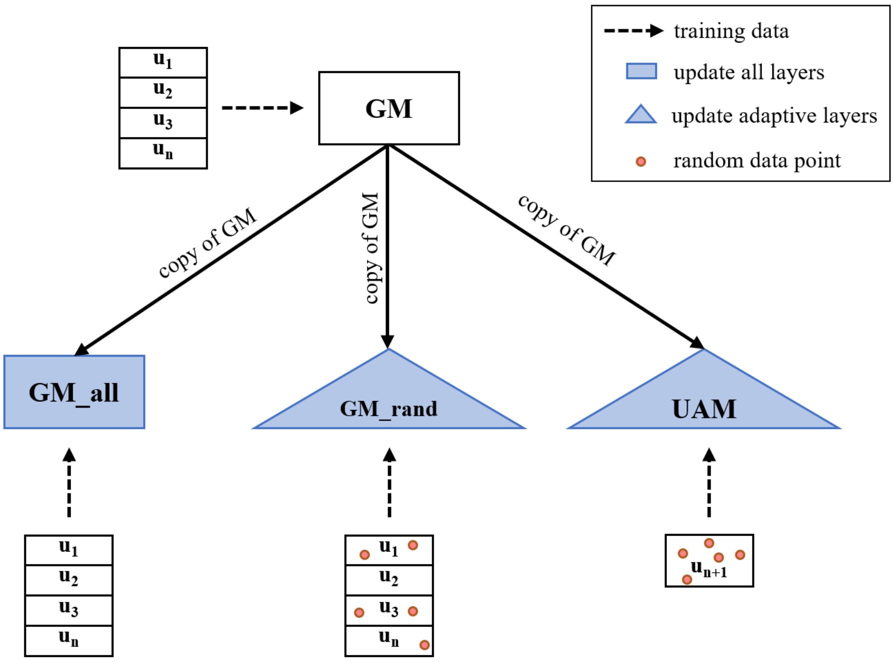


Fig. 2 The four types of models tested on a particular user u_{n+1}.

Complete user adaptive model with random oversampling (CUAM+RO). This model is a copy of the General model and then, the *complete* set of layers (fixed and adaptive layers) are updated with additional training epochs using the over-sampled instances.

In the following subsections we present the experiments and results for each of the two datasets.

6.1 Activity recognition dataset

For the activity recognition task, the fully connected deep neural network consisted of the following elements:

- An input layer consisting of 39 units corresponding to the 39 features.
- A *fixed* dense hidden layer consisting of 512 units with a RELU activation function and with a dropout rate of 0.2.
- An *adaptive* dense hidden layer consisting of 128 units with a RELU activation function and a dropout rate of 0.2.
- An *adaptive* dense output layer of 6 units (number of activity classes) with a softmax activation function.

This is a typical simple common network architecture. The parameters and architecture were empirically determined by using one of the hold-out users to test a GM. The general model was trained with an Adam optimizer (Kingma and Ba, 2014) with the default parameters provided in the original paper: $learning\_rate = 0.001$, $beta1 = 0.9$, $beta2 = 0.999$ and $\epsilon = 1e - 08$. At training time, the network was fed with batches of size 16 and for 50 epochs. 100 additional epochs were performed on copies of the GM in order to obtain the GM_all, the GM_rand and the UAM. Figure 3 shows the resulting average F-score with standard deviation bars for different percentages of adaptation data. For completeness, the respective accuracy plots are included in Appendix A.

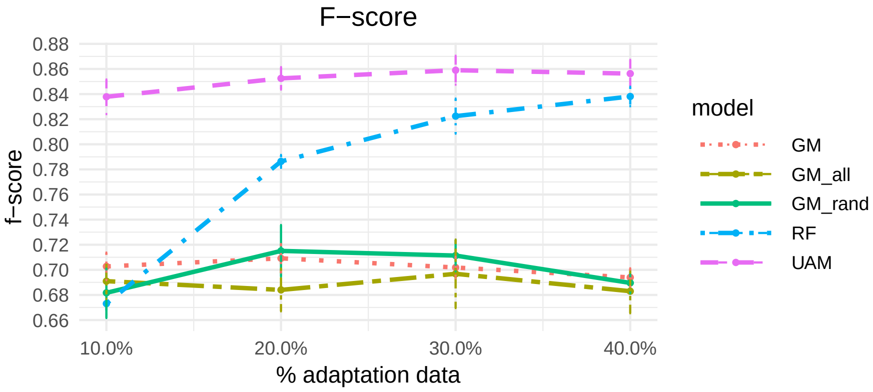
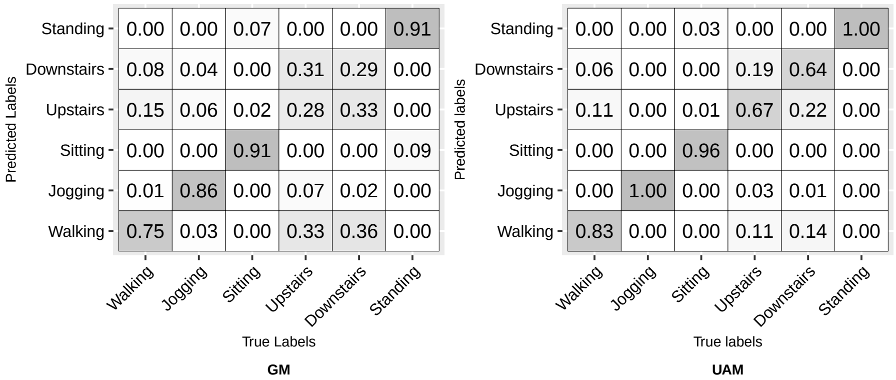


Fig. 3 Activity dataset F-score with 10%-40% adaptation data. The UAM performs better compared to the other models.

It can be clearly seen that the UAM performs better than the other models and that the performance increases as more adaptation data is added. As expected, the GM remains almost constant (with some variations due to random initialization). The GM_rand performed very similarly to the GM which supports the idea that the UAM performed much better because of the adaptation data used for each particular user. With just 10% of adaptation data, the F-score increased approximately 14% (from 70% of the GM to 84% for the UAM). The Random Forest increased its performance very quickly as more adaptation data was available. This was expected since this model is built specifically with only data from the target user, however the UAM performs better since it complements the lack of sufficient adaptation data.

Figure 4 shows the resulting confusion matrices for the GM and UAM of the aggregated predictions of all users for the 5 iterations. Here, it can be seen that the recall (diagonal) for all activities increased when using the UAM. Walking up and downstairs are the two most confused classes. They are confused against each other but in addition also with walking. This makes sense since all classes in one or another way include similar movements. Nevertheless, we can observe that the confusion is reduced significantly from GM to UAM.



	Walking	Jogging	Sitting	Upstairs	Downstairs	Standing
Standing	0.00	0.00	0.07	0.00	0.00	0.91
Downstairs	0.08	0.04	0.00	0.31	0.29	0.00
Upstairs	0.15	0.06	0.02	0.28	0.33	0.00
Sitting	0.00	0.00	0.91	0.00	0.00	0.09
Jogging	0.01	0.86	0.00	0.07	0.02	0.00
Walking	0.75	0.03	0.00	0.33	0.36	0.00

	Walking	Jogging	Sitting	Upstairs	Downstairs	Standing
Standing	0.00	0.00	0.03	0.00	0.00	1.00
Downstairs	0.06	0.00	0.00	0.19	0.64	0.00
Upstairs	0.11	0.00	0.01	0.67	0.22	0.00
Sitting	0.00	0.00	0.96	0.00	0.00	0.00
Jogging	0.00	1.00	0.00	0.03	0.01	0.00
Walking	0.83	0.00	0.00	0.11	0.14	0.00

Fig. 4 Activity dataset confusion matrices with 40% adaptation data.

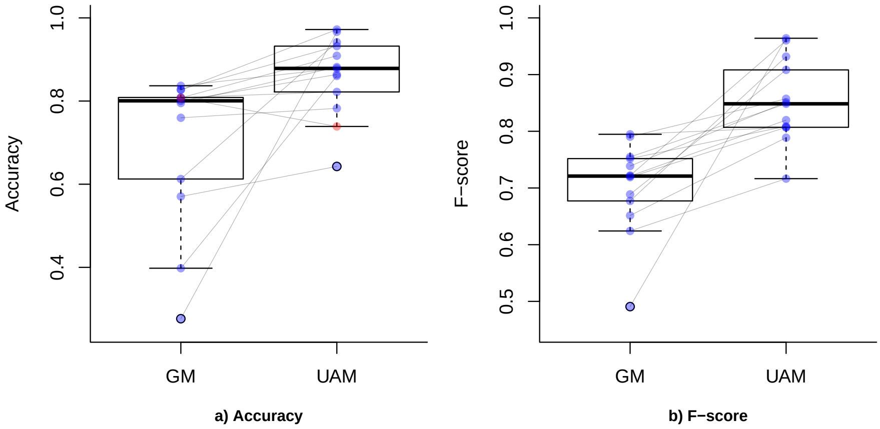


Fig. 5 Activity dataset paired-boxplots between the general model and the user adaptive model. **a) Accuracy.** GM median: 0.800, UAM median: 0.878. **b) F-score** GM median: 0.720, UAM median: 0.848.

Table 4 Performance on the hold-out users for the activity dataset.

Model	**Accuracy**	**F-score**
GM	80.0%	77.5%
GM_all	78.5%	76.8%
GM_rand	77.9%	75.7%
UAM	71.5%	71.9%
RF	54.4%	46.8%

Figure 5 shows the paired-boxplots for the accuracy and F-score between the GM and the UAM. Here, we can see that for all users the accuracy and F-score increased when adding adaptation data via deep transfer learning except for one. The median accuracy increase was 7.8% whereas the median F-score increase was 12.7%. The Cohen's d effect size for the accuracy was

0.75 (medium) and for the F-socre it was 1.2 (large). This error reduction is significant, specially if the system is intended to be used for prolonged periods of time because it could alleviate the accumulated error effect in the long term.

To investigate how the UAM performs on different users, 5 users were held out and the UAM from all other users were tested on those 5. Table 4 shows the average results. Interestingly, when a UAM is used on a user it was not intended for, the accuracy drops even below the general model. In this case the UAM accuracy was 71.5% and for the general model it was 80.0%. The F-score was also lower for the UAM. The performance of the Random Forest was the lowest one since it was trained with just adaptation from a different user. This suggests that once a UAM is built for a specific user, it should not be used on other users. This results also support the idea that data from users are very different from each other. This will be further investigated in the emotion recognition dataset since it also includes some demographic information about the participants like gender and age.

Figure 6 shows the results with random oversampling. From this figure, it can be seen that the CUAM+RO performed better than the plain UAM without data augmentation, however, the UAM combined with random oversampling (UAM+RO) achieved the best F-score results.

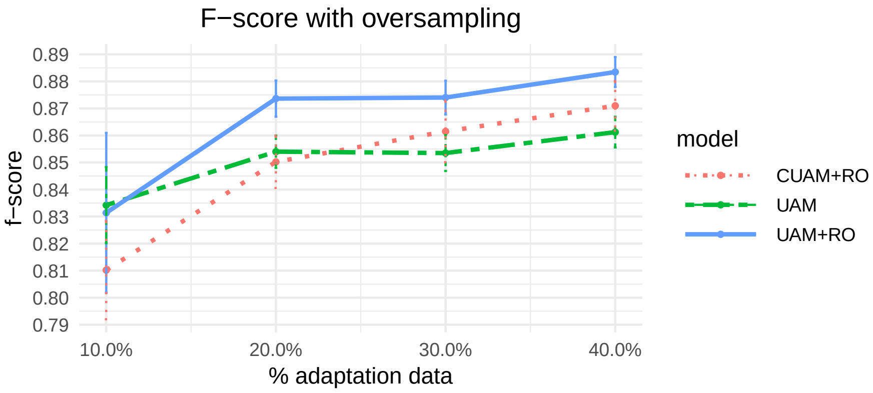


Fig. 6 Activity dataset F-score with random oversampling.

6.2 Speech emotion dataset

For the emotion recognition task, the deep neural network consisted of the following elements:

- An input layer with 68 units corresponding to the 68 audio features.
- A *fixed* dense hidden layer with 128 units with a RELU activation function and with a dropout rate of 0.2.

– An *adaptive* dense hidden layer with 64 units with a RELU activation function and a dropout rate of 0.2.
– An *adaptive* dense output layer of 7 units (number of emotions classes) with a softmax activation function.

As in the activity recognition task, the parameters and architecture were empirically determined by using one of the hold-out users. At training time, the network was fed with batches of size 16 and for 100 epochs. 100 additional epochs were performed on copies of the GM in order to obtain the GM_all, the GM_rand and the UAM.

Figure 7 shows the resulting average F-score with standard deviation bars for different percentages of adaptation data.

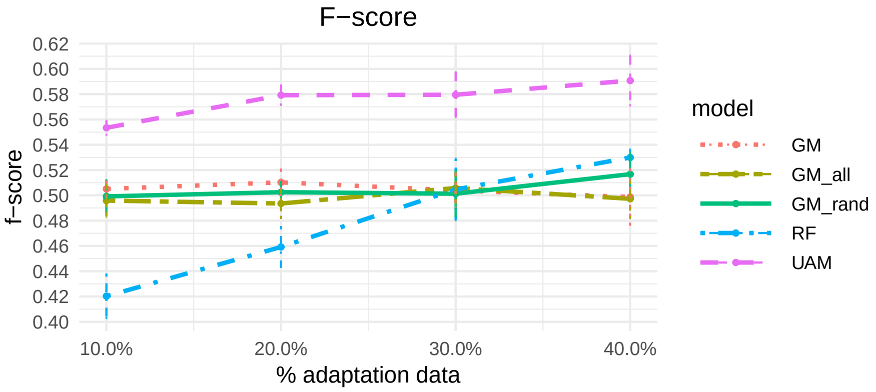
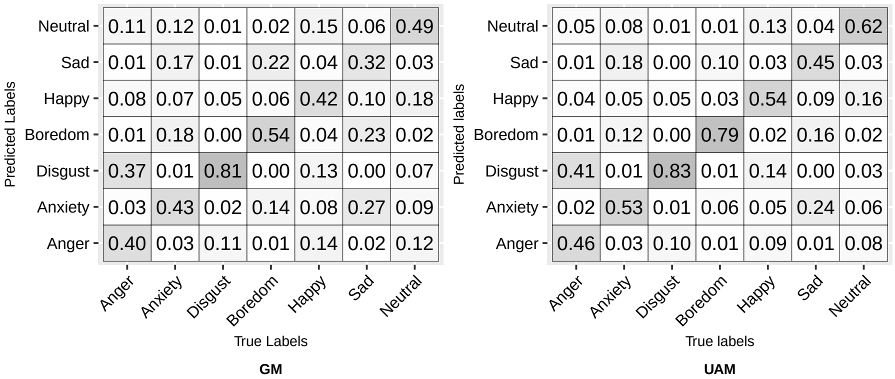


Fig. 7 Speech emotion dataset F-score with 10%-40% adaptation data. UAM achieves the highest F-score.

Again, the UAM performed better than the other approaches, and the performance increased as more adaptation data was added. All the general models performed very similar. With 10% of adaptation data, the F-score increased approximately 5% (from 50% of the GM to 55% for the UAM). Figure 8 shows the confusion matrices when the adaptation data is 40%. As with the activity task, here, all emotions experienced an increase in terms of recall. *Anger* and *disgust* were often confused which makes sense since they are even difficult to differentiate for a human Aviezer et al. (2008); Hutcherson and Gross (2011).

Figure 9 shows the paired-boxplots for the accuracy and F-score between the GM and the UAM. Here, the UAM was better for all users. In this case, the median accuracy increase was 9.5% and the median F-score increase was 7.0%. The accuracy Cohen's d effect size was 3.0 (large) and the F-score effect size was 1.9 (large). Table 5 shows the average results when testing the UAM on 3 hold-out users. The UAM had a lower accuracy and F-score compared with the general models which is consistent with the results from the activity recognition dataset. Again, Random Forest had the lowest accuracy. To



Predicted Labels	Anger	Anxiety	Disgust	Boredom	Happy	Sad	Neutral
Neutral	0.11	0.12	0.01	0.02	0.15	0.06	0.49
Sad	0.01	0.17	0.01	0.22	0.04	0.32	0.03
Happy	0.08	0.07	0.05	0.06	0.42	0.10	0.18
Boredom	0.01	0.18	0.00	0.54	0.04	0.23	0.02
Disgust	0.37	0.01	0.81	0.00	0.13	0.00	0.07
Anxiety	0.03	0.43	0.02	0.14	0.08	0.27	0.09
Anger	0.40	0.03	0.11	0.01	0.14	0.02	0.12

Predicted labels	Anger	Anxiety	Disgust	Boredom	Happy	Sad	Neutral
Neutral	0.05	0.08	0.01	0.01	0.13	0.04	0.62
Sad	0.01	0.18	0.00	0.10	0.03	0.45	0.03
Happy	0.04	0.05	0.05	0.03	0.54	0.09	0.16
Boredom	0.01	0.12	0.00	0.79	0.02	0.16	0.02
Disgust	0.41	0.01	0.83	0.01	0.14	0.00	0.03
Anxiety	0.02	0.53	0.01	0.06	0.05	0.24	0.06
Anger	0.46	0.03	0.10	0.01	0.09	0.01	0.08

Fig. 8 Speech emotion dataset confusion matrices with 40% adaptation data.

understand this behavior, we first plotted the accuracies of the UAMs tested with one of the hold-out users (see Figure 10). Here, we can see that the best performing UAMs were trained with male users which is the same gender of the hold-out user. UAMs of female users performed worse when tested on a male user.

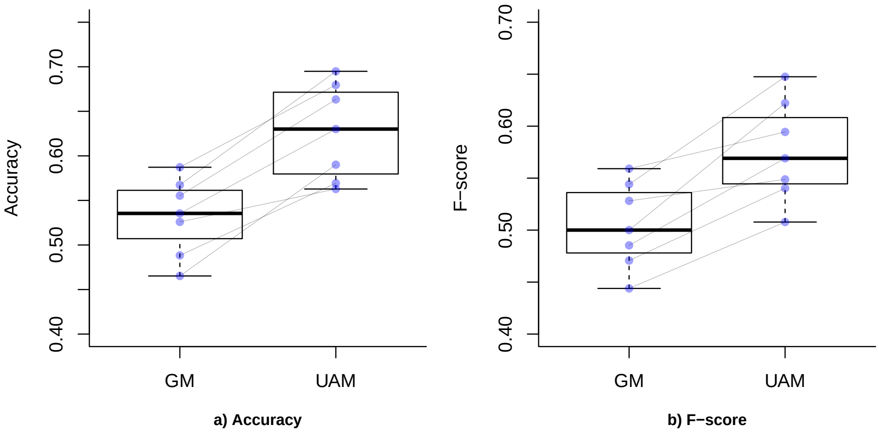


Fig. 9 Speech emotion dataset paired-boxplots between the general model and the user adaptive model. **a) Accuracy.** GM median: 0.535, UAM median: 0.630. **b) F-score** GM median: 0.500, UAM median: 0.570.

To further analyze this, we performed a leave-one-user-out evaluation. For each user we trained three models: One with data from all other users, another one with data from users of the same gender and another one with data from users of different gender.

Figure 11 shows the F-score. The median F-score for the model trained with the same gender was the highest one, however, the differences were not

Table 5 Performance on the hold-out users for the speech emotion dataset.

Model	Accuracy	F-score
GM	45.4%	45.3%
GM_all	45.1%	45.0%
GM_rand	44.6%	44.5%
UAM	43.5%	42.9%
RF	33.1%	28.8%

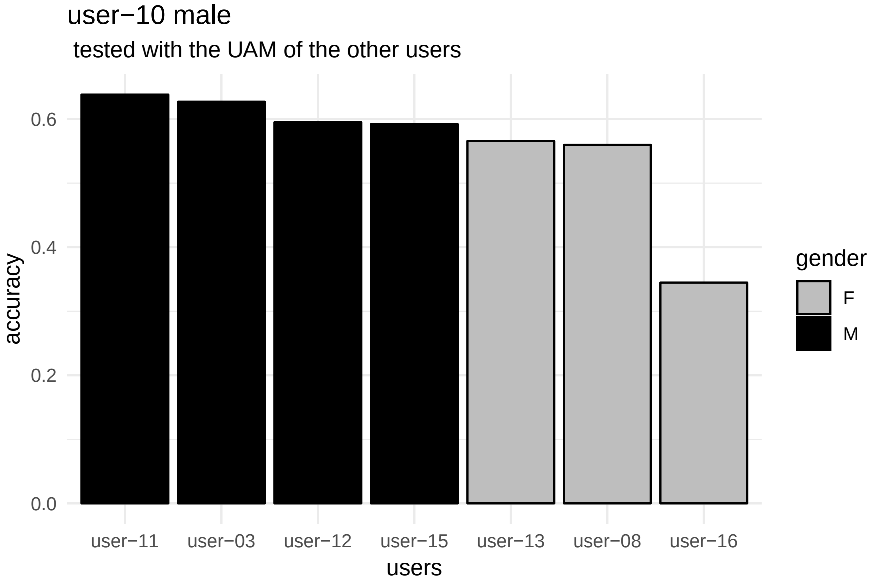


Fig. 10 Speech emotion dataset user-10 tested with all other users adapted models.

significant compared to training with all users. On the other hand, the models trained with data from users of different genders performed much worse, and the differences were significant. The statistical significance was assessed with a paired Wilcoxon signed rank test.

From these results, we can see that the particular characteristics of each user had a high impact, thus, the need to build models that adapt to the attributes of each person becomes very important.

Given the previous results, we hypothesized that deriving a UAM from a GM that only includes information from similar users would produce better results than deriving a UAM from a GM that includes information from all users. To test this, we performed a leave one user out validation and for each, we built two types of UAMs:

1. *UAM_all*: This is a UAM adapted from a GM trained with all the data.
2. *UAM_same*: This is a UAM adapted from a GM trained with data *just* from participants with the *same gender*.

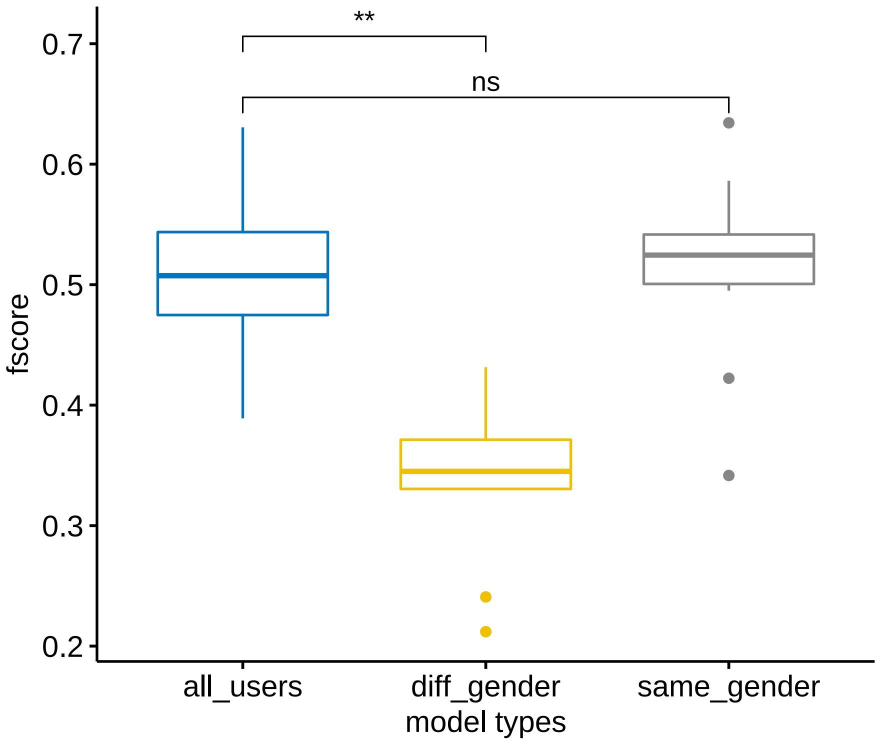


Fig. 11 Speech emotion dataset boxplots F-score.

As in previous experiments, the percent of adaptation data was from 10% to 40% with increments of 10 and five iterations were performed for each. Table 6 shows the average accuracy results across iterations. The accuracy for both models looks very similar and the differences were not statistically significant. Thus, there is not enough evidence to support our hypothesis that adapting a model from a general one that was trained with data from the same gender produces better results. The reason for this may be that the fixed layers act as feature extractors and by updating the adaptation layers differences between genders are automatically taken care of. These preliminary results suggest that building a UAM from a GM with all the data is sufficient to produce good adapted models. By adding a data pre-selection step, e.g., to just include similar users (same gender in this case) in the GM does not seem to provide additional benefits to the final UAM. The emotions dataset also contains age information about the participants. Given that the age between participants is very similar, we did not conducted experiments based on age. The age range is 21-35 years with most of the participants being between 31 and 35 years old. For the activities dataset we could not test the effect of gender or age since the database does not contain demographic information.

Table 6 Average accuracy for different percent of adaptation data.

model / % adaptation data	10%	20%	30%	40%
UAM_all	0.635	0.651	0.652	0.656
UAM_same	0.636	0.653	0.661	0.668

Figure 12 shows the F-score with random oversampling. Contrary to the activity dataset, here, the CUAM+RO model performed better than the UAM+RO. This may be due the fact that the emotion dataset has fewer users. To test this, we randomly selected only 7 users from the activity dataset. We chose 7 to make it the same as the emotion dataset (10 users minus 3 holdout users). Results are shown in Figure 13. We can see that the UAM+RO is better. It can also be seen that when the initial number of users in the training data is low, the UAM method tends to be better than CUAM+RO.

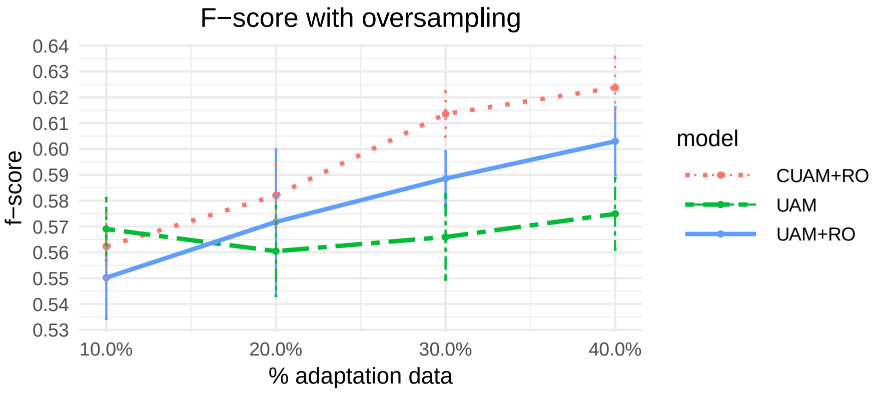


Fig. 12 Emotion dataset F-score with random oversampling.

7 Discussion

From our experiments and results, we can see that the average performance increase of the adapted model with respect to the general one was significant for both datasets. It was around 5%-14% with the additional 10% of data from the target user. For the data augmentation case (with random oversampling) the F-score of the UAM+RO performed the best with the activities dataset. With the emotion dataset, the best model was CUAM+RO. In both datasets, the use of data augmentation produced better results than the plain user adapted model (UAM). Even though the datasets used in this work are

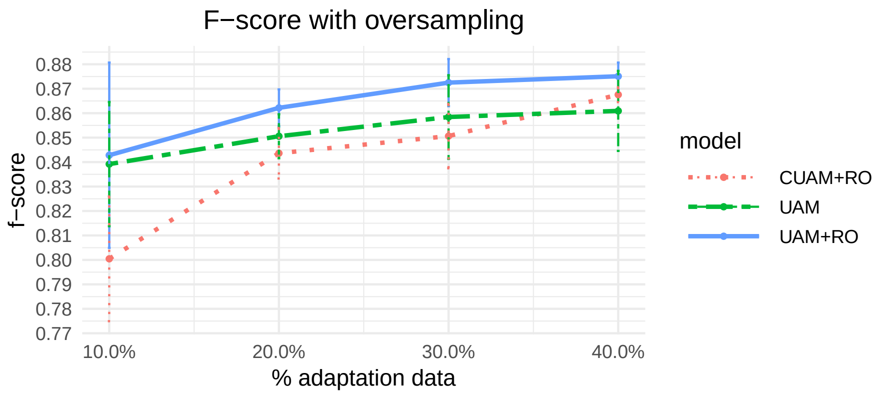


Fig. 13 Activity dataset F-score with random oversampling with only 7 users.

from very different domains, our method achieved significant performance improvements on both of them. We believe that the proposed approach could work on other datasets within the same domains but also for other tasks such as gesture recognition, stress detection, etc.

The extent to which the prediction performance of a system is useful or comparable to humans is very dependant on each application. For instance, a true positive rate of 0.2 may seem very low but if it is for predicting suicides, then it is worth it since it means saving human lives. On the other hand, the same performance is not acceptable for an activity recognition system that computes number of steps since it will provide very misleading results.

Another thing to note is that reported performance metrics across different studies vary a lot even when using the same datasets. In part, this is influenced by the type of pre-processing, feature extraction, algorithms, random initialization, and so on. Another important factor that influences the result is the validation method. In many cases, cross-validation is performed on the entire dataset without user distinction. This for example, will produce a higher accuracy compared to a leave-one-user out validation approach. In this work we tried to compensate for that by: 1) performing several iterations to account for variability. 2) Use the original features provided in the activity dataset or in the case of emotion recognition, extracting standard sound features. 3) Split train and test sets such that the train set contains the first n observations chronologically to avoid overfitting because contiguous observations may be highly correlated.

In this study, we also identified some of the limitations of our proposed method: I) Our method assumes that users' behavior do not change over time, i.e., there exist only between-user differences and not within-user differences. The datasets used in this study were collected on the same day by each participant. In real life, users also change behaviors over time. This needs to be further evaluated with longer-term datasets. II) In our current experiments, the UAMs were derived by just updating the wights of the adaptive layers or all layers (CUAM+RO). For future work, it would be interesting to explore more dynamical network architectures, i.e., instead of just updating the weights, the number of units in each layer could be adjusted or new adaptive layers could be added or removed in a per user basis. III) For data augmentation, we explored the use of random oversampling that just duplicates data points which can lead to overfitting. It would be worth exploring the use of other methods to generate synthetic data such as SMOTE (Chawla et al., 2002) and Generative Adversarial Networks (Goodfellow et al., 2014). IV) Another limitation of our approach is that it requires at least some amount of labeled data for the target user. We believe it is also worth exploring new methods of adaptation that would not require any labeled data at all for a particular user, e.g., by using semi-supervised learning methods (Chapelle et al., 2006) to generate the adaptation data set without any user intervention. A possible semi-supervised approach could be self-learning (Scudder, 1965) as used by Garcia-Ceja and Brena (2016) to build adaptive models requiring zero labeling from the user.

8 Conclusions

In this work we used deep transfer learning and data augmentation to build UAMs with small amounts of training data. We showed that with just 10% of additional labeled data, the performance of the UAM increased from 5%-14% with respect to the GM. We conducted experiments to find whether individual characteristics affect model performance. Our results on the emotion dataset showed that models built with participants of different genders had significant lower accuracies. However, when building a UAM, gender differences did not have any significant impact on the final model. Thus, our results showed that the adapted models trained with deep transfer learning methods were robust to the differences in underlying characteristics of the users (gender, in this case). Furthermore, we conducted experiments to evaluate the effect of adapting a model via data augmentation using random oversampling. The results showed that random oversampling outperformed the plain UAM. We also evaluated a combination of UAMs with random oversampling which obtained the best results on the activity recognition dataset. For the emotion recognition task, the CUAM+RO in which all layers are updated achieved better results than the UAM+RO. For future work, we would also like to analyze the impact of different features and explore whether or not some features are more important for some users. In our experiments, we did not use demographic information as features. For future directions we would also like to evaluate the impact of incorporating this type of information into the models.

In summary, we proposed a method that based on the results on two datasets, demonstrated to be effective for increasing the prediction performance when limited training data is available. The method was tested on two distinct tasks (activity and emotion recognition) and proved to be reliable on both of them. We believe this method can be used for other use cases such as gesture recognition or health monitoring systems but additional experiments are required including datasets collected during longer periods of time to test the effects of within-user variance.

Acknowledgements This publication is part of the INTROducing Mental health through Adaptive Technology (INTROMAT) project, funded by the Norwegian Research Council (259293/o70) and as a part of the RCN Centres of Excellence scheme, project number 262762.

A Appendix

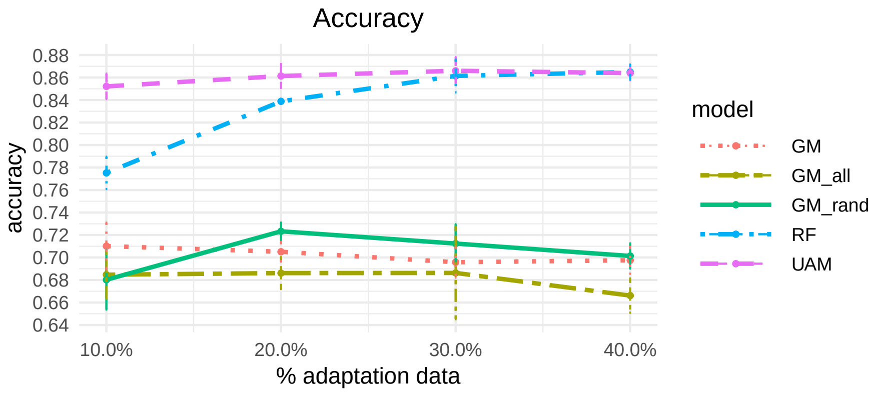


Fig. 14 Activity dataset accuracy with 10%-40% adaptation data. The UAM performs better compared to the other models.

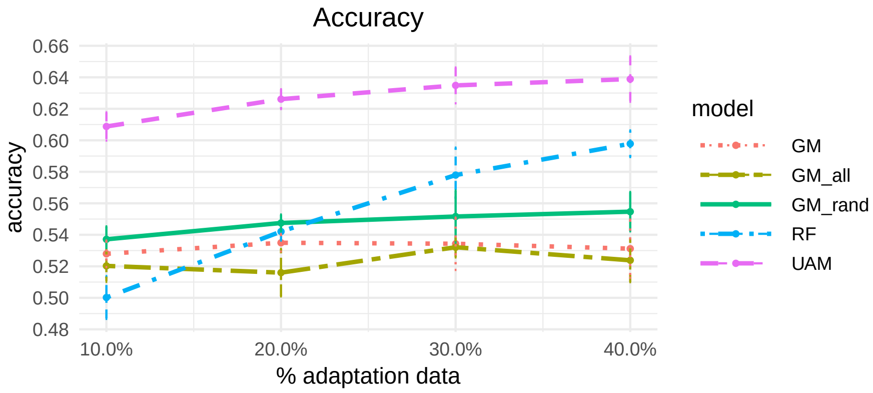


Fig. 15 Speech emotion dataset accuracy with 10%-40% adaptation data. The UAM outperforms the other approaches.

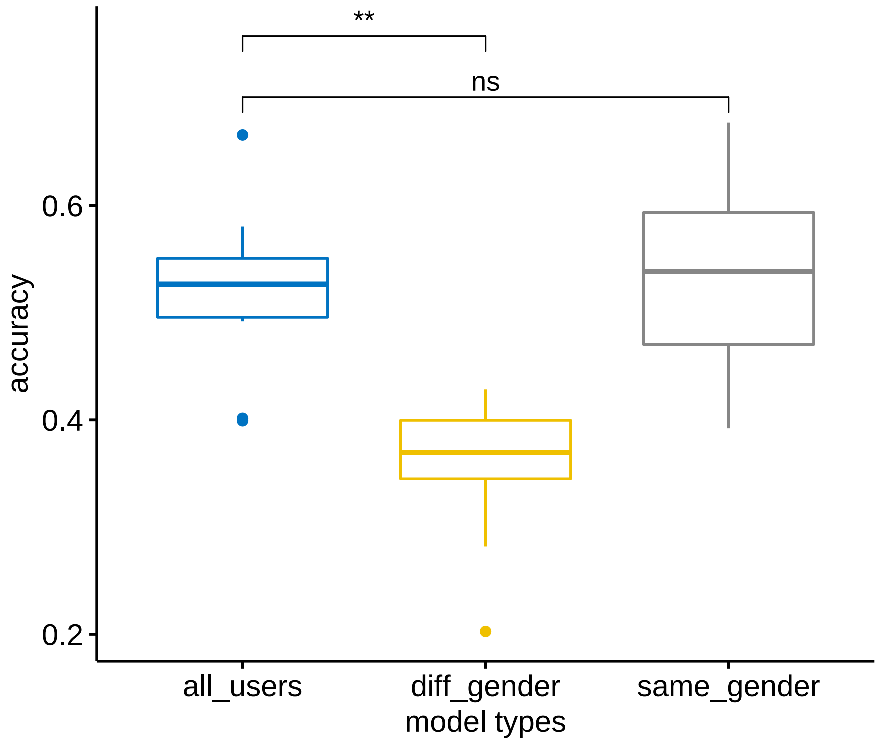


Fig. 16 Speech emotion dataset boxplots accuracy.

References

Abdallah Z, Gaber M, Srinivasan B, Krishnaswamy S (2012) StreamAR: Incremental and active learning with evolving sensory data for activity recognition. In: Tools with Artificial Intelligence (ICTAI), 2012 IEEE 24th International Conference on, vol 1, pp 1163–1170, DOI 10.1109/ICTAI.2012.169

Alnujaim I, Alali H, Khan F, Kim Y (2018) Hand gesture recognition using input impedance variation of two antennas with transfer learning. IEEE Sensors Journal 18(10):4129–4135, DOI 10.1109/JSEN.2018.2820000

Avci A, Bosch S, Marin-Perianu M, Marin-Perianu R, Havinga P (2010) Activity recognition using inertial sensing for healthcare, wellbeing and sports applications: A survey. In: Architecture of Computing Systems (ARCS), 2010 23rd International Conference on, pp 1–10

Aviezer H, Hassin RR, Ryan J, Grady C, Susskind J, Anderson A, Moscovitch M, Bentin S (2008) Angry, disgusted, or afraid? studies on the malleability of emotion perception. Psychological science 19(7):724–732

Ayadi ME, Kamel MS, Karray F (2011) Survey on speech emotion recognition: Features, classification schemes, and databases. Pattern Recognition 44(3):572 – 587, DOI https://doi.org/10.1016/j.patcog.2010.09.020, URL http://www.sciencedirect.com/science/article/pii/S0031320310004619

Badshah AM, Ahmad J, Rahim N, Baik SW (2017) Speech emotion recognition from spectrograms with deep convolutional neural network. In: 2017 International Conference on

Platform Technology and Service (PlatCon), pp 1–5, DOI 10.1109/PlatCon.2017.7883728

Bengio Y (2012) Deep learning of representations for unsupervised and transfer learning. In: Proceedings of ICML Workshop on Unsupervised and Transfer Learning, pp 17–36

Blitzer J, Dredze M, Pereira F (2007) Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In: Proceedings of the 45th annual meeting of the association of computational linguistics, pp 440–447

Brezmes T, Gorricho JL, Cotrina J (2009) Activity Recognition from Accelerometer Data on a Mobile Phone. In: Omatu S, Rocha M, Bravo J, Fernndez F, Corchado E, Bustillo A, Corchado J (eds) Distributed Computing, Artificial Intelligence, Bioinformatics, Soft Computing, and Ambient Assisted Living, Lecture Notes in Computer Science, vol 5518, Springer Berlin / Heidelberg, pp 796–799

Burkhardt F, Paeschke A, Rolfes M, Sendlmeier WF, Weiss B (2005) A database of german emotional speech. In: Ninth European Conference on Speech Communication and Technology

Chapelle O, Schölkopf B, Zien A, others (2006) Semi-supervised learning. MIT press Cambridge

Chatterjee J, Mukesh V, Hsu H, Vyas G, Liu Z (2018) Speech emotion recognition using cross-correlation and acoustic features. In: 2018 IEEE 16th Intl Conf on Dependable, Autonomic and Secure Computing, 16th Intl Conf on Pervasive Intelligence and Computing, 4th Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress(DASC/PiCom/DataCom/CyberSciTech), pp 243–249

Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP (2002) Smote: synthetic minority oversampling technique. Journal of artificial intelligence research 16:321–357

Devin C, Gupta A, Darrell T, Abbeel P, Levine S (2017) Learning modular neural network policies for multi-task and multi-robot transfer. In: Robotics and Automation (ICRA), 2017 IEEE International Conference on, IEEE, pp 2169–2176

EmotionDB (1999) Berlin database of emotional speech. http://emodb.bilderbar.info/docu/, accessed: 28 January 2018

Fallahzadeh R, Ghasemzadeh H (2017) Personalization without user interruption: boosting activity recognition in new subjects using unlabeled data. In: Proceedings of the 8th International Conference on Cyber-Physical Systems, ACM, pp 293–302

Gama J, liobait I, Bifet A, Pechenizkiy M, Bouchachia A (2014) A survey on concept drift adaptation. ACM Computing Surveys (CSUR) 46(4):44

Garcia-Ceja E, Brena R (2015) Building personalized activity recognition models with scarce labeled data based on class similarities. In: García-Chamizo JM, Fortino G, Ochoa SF (eds) Ubiquitous Computing and Ambient Intelligence. Sensing, Processing, and Using Environmental Information, Springer International Publishing, Cham, pp 265–276

Garcia-Ceja E, Brena RF (2016) Activity recognition using community data to complement small amounts of labeled instances. Sensors 16(6), DOI 10.3390/s16060877, URL http://www.mdpi.com/1424-8220/16/6/877

Garcia-Ceja E, Osmani V, Mayora O (2016) Automatic stress detection in working environments from smartphones' accelerometer data: A first step. IEEE Journal of Biomedical and Health Informatics 20(4):1053–1060, DOI 10.1109/JBHI.2015.2446195

Garcia-Ceja E, Riegler M, Nordgreen T, Jakobsen P, Oedegaard KJ, Trresen J (2018) Mental health monitoring with multimodal sensing and machine learning: A survey. Pervasive and Mobile Computing 51:1 – 26, DOI https://doi.org/10.1016/j.pmcj.2018.09.003, URL http://www.sciencedirect.com/science/article/pii/S1574119217305692

Giannakopoulos T (2015) pyaudioanalysis: An open-source python library for audio signal analysis. PLOS ONE 10(12):1–17, DOI 10.1371/journal.pone.0144610, URL https://doi.org/10.1371/journal.pone.0144610

Giannakopoulos T (2016) Python audio analysis library. https://github.com/tyiannak/pyAudioAnalysis, accessed: 28 January 2018

Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y (2014) Generative adversarial nets. In: Advances in neural information processing systems, pp 2672–2680

Grünerbl A, Muaremi A, Osmani V, Bahle G, Öhler S, Trster G, Mayora O, Haring C, Lukowicz P (2015) Smartphone-based recognition of states and state changes in bipolar disorder patients. IEEE Journal of Biomedical and Health Informatics 19(1):140–148, DOI 10.1109/JBHI.2014.2343154

Haykin S (1994) Neural networks: a comprehensive foundation. Prentice Hall PTR

Hutcherson CA, Gross JJ (2011) The moral emotions: A social–functionalist account of anger, disgust, and contempt. Journal of personality and social psychology 100(4):719

Karam ZN, Provost EM, Singh S, Montgomery J, Archer C, Harrington G, Mcinnis MG (2014) Ecologically valid long-term mood monitoring of individuals with bipolar disorder using speech. In: Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on, IEEE, pp 4858–4862

Kingma DP, Ba J (2014) Adam: A method for stochastic optimization. CoRR abs/1412.6980, URL http://arxiv.org/abs/1412.6980, 1412.6980

Kotsiantis S, Kanellopoulos D, Pintelas P, others (2006) Handling imbalanced datasets: A review. GESTS International Transactions on Computer Science and Engineering 30(1):25–36

Kwapisz JR, Weiss GM, Moore SA (2011) Activity recognition using cell phone accelerometers. SIGKDD Explor Newsl 12(2):74–82, DOI 10.1145/1964897.1964918, URL http://doi.acm.org/10.1145/1964897.1964918

Lalitha S, Madhavan A, Bhushan B, Saketh S (2014) Speech emotion recognition. In: Advances in Electronics, Computers and Communications (ICAECC), 2014 International Conference on, IEEE, pp 1–4

Lane ND, Xu Y, Lu H, Hu S, Choudhury T, Campbell AT, Zhao F (2011) Enabling Large-scale Human Activity Inference on Smartphones Using Community Similarity Networks (Csn). In: Proceedings of the 13th International Conference on Ubiquitous Computing, ACM, New York, NY, USA, UbiComp '11, pp 355–364, DOI 10.1145/2030112.2030160, URL http://doi.acm.org/10.1145/2030112.2030160

LeCun Y, Bottou L, Bengio Y, Haffner P (1998) Gradient-based learning applied to document recognition. Proceedings of the IEEE 86(11):2278–2324

Lin YL, Wei G (2005) Speech emotion recognition based on hmm and svm. In: 2005 International Conference on Machine Learning and Cybernetics, vol 8, pp 4898–4901 Vol. 8, DOI 10.1109/ICMLC.2005.1527805

Lockhart JW, Weiss GM (2014) The benefits of personalized smartphone-based activity recognition models. In: Proceedings of the 2014 SIAM International Conference on Data Mining, pp 614–622, URL http://epubs.siam.org/doi/abs/10.1137/1.9781611973440.71

López-Nava I, Muñoz-Meléndez A (2018) High-level features for recognizing human actions in daily living environments using wearable sensors. In: Multidisciplinary Digital Publishing Institute Proceedings, vol 2, p 1238

Lu H, Frauendorfer D, Rabbi M, Mast MS, Chittaranjan GT, Campbell AT, Gatica-Perez D, Choudhury T (2012) StressSense: Detecting stress in unconstrained acoustic environments using smartphones. In: Proceedings of the 2012 ACM Conference on Ubiquitous Computing, ACM, UbiComp '12, pp 351–360, DOI 10.1145/2370216.2370270, URL http://doi.acm.org/10.1145/2370216.2370270

Mannini A, Sabatini AM (2010) Machine Learning Methods for Classifying Human Physical Activity from On-Body Accelerometers. Sensors 10(2):1154–1175, DOI 10.3390/s100201154, URL http://www.mdpi.com/1424-8220/10/2/1154

Martínez-Pérez FE, González-Fraga JA, Cuevas-Tello JC, Rodríguez MD (2012) Activity inference for ambient intelligence through handling artifacts in a healthcare environment. Sensors 12(1):1072–1099, DOI 10.3390/s120101072, URL http://www.mdpi.com/1424-8220/12/1/1072

Maxhuni A, Hernandez-Leal P, Sucar LE, Osmani V, Morales EF, Mayora O (2016) Stress modelling and prediction in presence of scarce data. Journal of Biomedical Informatics 63:344 – 356, DOI https://doi.org/10.1016/j.jbi.2016.08.023, URL http://www.sciencedirect.com/science/article/pii/S1532046416301095

Mitchell E, Monaghan D, O'Connor NE (2013) Classification of sporting activities using smartphone accelerometers. Sensors 13(4):5317–5337

Pan SJ, Yang Q (2010) A survey on transfer learning. IEEE Transactions on Knowledge and Data Engineering 22(10):1345–1359

Parviainen J, Bojja J, Collin J, Leppnen J, Eronen A (2014) Adaptive activity and environment recognition for mobile phones. Sensors 14(11):20753–20778, DOI 10.3390/s141120753, URL http://www.mdpi.com/1424-8220/14/11/20753

Peng P, Tian Y, Xiang T, Wang Y, Pontil M, Huang T (2017) Joint semantic and latent attribute modelling for cross-class transfer learning. IEEE transactions on pattern analysis and machine intelligence

Richter J, Wiede C, Dayangac E, Shahenshah A, Hirtz G (2016) Activity recognition for elderly care by evaluating proximity to objects and human skeleton data. In: International Conference on Pattern Recognition Applications and Methods, Springer, pp 139–155

Rokni SA, Nourollahi M, Ghasemzadeh H (2018) Personalized human activity recognition using convolutional neural networks. CoRR abs/1801.08252, URL http://arxiv.org/abs/1801.08252, 1801.08252

Sanchez W, Martinez A, Campos W, Estrada H, Pelechano V (2015) Inferring loneliness levels in older adults from smartphones. Journal of Ambient Intelligence and Smart Environments 7(1):85–98

Scudder I H (1965) Probability of error of some adaptive pattern-recognition machines. Information Theory, IEEE Transactions on 11(3):363–371, DOI 10.1109/TIT.1965.1053799

Sevakula RK, Singh V, Verma NK, Kumar C, Cui Y (2018) Transfer learning for molecular cancer classification using deep neural networks. IEEE/ACM Transactions on Computational Biology and Bioinformatics pp 1–1, DOI 10.1109/TCBB.2018.2822803

Shin HC, Roth HR, Gao M, Lu L, Xu Z, Nogues I, Yao J, Mollura D, Summers RM (2016) Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning. IEEE transactions on medical imaging 35(5):1285–1298

Shoaib M, Bosch S, Incel OD, Scholten H, Havinga PJM (2014) Fusion of Smartphone Motion Sensors for Physical Activity Recognition. Sensors 14(6):10146–10176, DOI 10.3390/s140610146, URL http://www.mdpi.com/1424-8220/14/6/10146

Soleymani M, Riegler M, Halvorsen P (2018) Multimodal analysis of user behavior and browsed content under different image search intents. International Journal of Multimedia Information Retrieval 7(1):29–41

Tarnowski P, Koodziej M, Majkowski A, Rak RJ (2017) Emotion recognition using facial expressions. Procedia Computer Science 108:1175 – 1184, DOI https://doi.org/10.1016/j.procs.2017.05.025, URL http://www.sciencedirect.com/science/article/pii/S1877050917305264, international Conference on Computational Science, ICCS 2017, 12-14 June 2017, Zurich, Switzerland

Vildjiounaite E, Kallio J, Mntyjrvi J, Kyllnen V, Lindholm M, Gimel'farb G (2017) Unsupervised stress detection algorithm and experiments with real life data. In: Oliveira E, Gama J, Vale Z, Lopes Cardoso H (eds) Progress in Artificial Intelligence, Springer International Publishing, pp 95–107

Vo QV, Hoang MT, Choi D (2013) Personalization in mobile activity recognition system using k-medoids clustering algorithm. International Journal of Distributed Sensor Networks 9(7):315841

Wang X, Rosenblum D, Wang Y (2012) Context-aware mobile music recommendation for daily activities. In: Proceedings of the 20th ACM international conference on Multimedia, ACM, pp 99–108

Wisdm (2012) Activity prediction dataset. http://www.cis.fordham.edu/wisdm/dataset.php, accessed: 28 January 2018

Xu Q, Nwe TL, Guan C (2015) Cluster-based analysis for personalized stress evaluation using physiological signals. Biomedical and Health Informatics, IEEE Journal of 19(1):275–281, DOI 10.1109/JBHI.2014.2311044

Yarowsky D (1995) Unsupervised word sense disambiguation rivaling supervised methods. In: Proceedings of the 33rd annual meeting on Association for Computational Linguistics, Association for Computational Linguistics, pp 189–196

Zenonos A, Khan A, Kalogridis G, Vatsikas S, Lewis T, Sooriyabandara M (2016) HealthyOffice: Mood recognition at work using smartphones and wearable sensors. In: 2016 IEEE International Conference on Pervasive Computing and Communication Workshops (PerCom Workshops), pp 1–6, DOI 10.1109/PERCOMW.2016.7457166

Enrique Garcia-Ceja is a researcher at SINTEF Digital. Previously, he was a Postdoc fellow at the Robotics and Intelligent Systems Group, University of Oslo. He received his

Ph.D. degree in intelligent systems from Tecnolgico de Monterrey University, Mexico in 2016. His main research topic is the analysis of wearable sensors data using machine learning to understand user contextual information such as physical activity, location and mood.

Michael Riegler is a senior researcher at Simula Research Laboratory. He received his Masters degree from Klagenfurt University with distinction and finished his PhD at the University of Oslo in two and a half years. His PhD thesis topic was efficient processing of medical multimedia workloads. His research interests are medical multimedia data analysis and understanding, image processing, image retrieval, parallel processing, crowdsourcing, social computing and user intent. He is involved in several initiatives like the MediaEval Benchmarking initiative for Multimedia Evaluation. Furthermore he is part of an expert group for the Norwegian Council of Technology on Machine Learning for Healthcare.

Anders K. Kvernberg studied informatics at the University of Oslo, and wrote his master's thesis on mental health prediction using machine learning. After finishing his master's degree he started working as a software engineer in Sopra Steria where he still works today. His interests include, but are not limited to, software development, distributed systems and machine learning.

Jim Torresen is a Professor in computer science at the University of Oslo, Norway and head of the Robotics and Intelligent Systems group. He is also a principle investigator in the RITMO Centre for Interdisciplinary Studies in Rhythm, Time and Motion. His research interests include artificial intelligence, ethical aspects of AI and robotics, machine learning, robotics, and applying this to complex real-world applications which have resulted in approximately 180 scientific peer-reviewed papers in international journals, books and conference proceedings. He is a member of the Norwegian Academy of Technological Sciences (NTVA) and the National Committee for Research Ethics in Science and Technology (NENT).