Analyzing Data from Smartphone Sensors to Recognize Human Activities using Machine Learning

Research Project
Masters in Data Analytics

Abhishek Angne
Student ID: 18136923

School of Computing
National College of Ireland

Supervisor: Mr. Christian Horn

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Abhishek Angne
Student ID:	x18136923
Programme:	MSc. in Data Analytics
Year:	2019
Module:	MSc Research Project
Supervisor:	Mr. Christian Horn
Submission Due Date:	12/12/2019
Project Title:	Analyzing Data from Smartphone Sensors toRecognize Human Activities using Machine Learning
Word Count:	6832
Page Count:	27

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	12th December 2019

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Analyzing Data from Smartphone Sensors to Recognize Human Activities using Machine Learning

Abhishek Angne
x18136923

Abstract

There's a massive demand in the field of Human Activity Recognition(HAR) for surveillance and healthcare domains by leveraging machine learning algorithms, artificial intelligence, on data from smartphone-based sensors. This project delves deep into recognizing human activities by examining data extracted from multimodal sensors present in ubiquitous devices like the smartphone and the smartwatch with the help of machine learning algorithms. HAR has its roots deeply embedded in the fields of medical science and surveillance systems with applications like fitness recommendation and activity monitoring respectively. The CRISP-DM methodology has been used to analyze the Extrasensory dataset by Vaizman and Ellis (2017) and team to classify various activities. Feature extraction, ensemble methods, and deep learning algorithms have been leveraged to analyze the dataset.

The research dives in to the use of different machine learning classifiers, with different feature selection and hyper-parameter optimization approaches. In order to perform HAR based classification, models based on Random Forest, Extreme Gradient Boosting classifier with 10-fold cross validation with stratified sampling, and Multilayer Perceptron were created for pre-labelled data to predict user activities from the Extrasensory dataset. The Multilayer Perceptron classifier performed best out of all the classifiers with an accuracy of 91.5%.

Keywords: Human Activity Recognition, Pervasive Computing, Machine Learning, Data Mining, Classification.

1 Introduction

Everyone wants to have a life with longevity and good health. Fitness trackers made by technology giants like Google, Apple and Fitbit with the help of dedicated operations systems like the Wear OS by Google and watchOS by apple are being used extensively to understand and analyze our daily activities. The backbone of these wearable devices is extensive scientific research in the field of human activity recognition (HAR). HAR has been hitting the headlines in recent years due to extensive demands in the fields of surveillance and healthcare. The progress of ubiquitous devices like smartphones, smartwatches, and wearables have greatly helped in the development of this domain. In simpler words, activity recognition is the process of predicting a person's activity using sensor-based data from ubiquitous devices like the smartphone.

However, there are multiple aspects that still need to be looked into for further development. Ensemble methods, deep learning frameworks, and many other machine learning techniques have been constantly used to test and classify activities. Single sensor classification and multimodal sensor classification techniques have been utilized to understand the activity patterns of individuals and further used to share their activity reports for health improvement purposes.

1.0.1 Research Question

"Can data extracted from sensors present in smartphones and wearable devices be used to recognize human activities to recommend better health care solutions?"

Objectives:
1) Identifying human activity at a given timestamp.
2) Analyze multi-modal sensor-readings(features) and their vitality in predicting a human activity.
3) Creating a significant model for human activity recognition using machine learning algorithm.

This research aims at performing activity recognition with the help of classification using machine learning algorithms. It will be of immense use in the fields of fitness and medical science. There are multiple aspects wherein the activities analyzed from human activity data to work on fitness-related problems like injuries, obesity, surveillance trackers and many more. The ACL (anterior cruciate ligament) injury is one such injury wherein gyroscope sensor values are being used to analyze walking speeds and vertical motion to provide quicker recovery recommendations (Malik; 2017). Patients diagnosed with Parkinson's disease can be given better exercise recommendations with the help of multi-modal sensors in specific targeted variable ubiquitous devices (Cheng et al.; 2017). In terms of surveillance systems, government organizations can leverage the use of electronic sensors to track the movements and activities of individuals under surveillance. The analyzed data can be further used to predict behavioral patterns, personality traits, context recognition, situation response of the individual and much more. The most challenging aspects of this field are data collection and data privacy. Leveraging one such novel research by (Vaizman and Ellis; 2017), wherein both these aspects were taken care of, this research project aims to classify unsupervised data on human activities whilst maintaining their privacy.

2 Related Work

Human Activity Recognition goes hand-in-hand with the Internet of Things(IoT) as the IoT sensors are utilized and analyzed to perform HAR (Sakr et al.; 2018). The HAR domain has primarily been focusing on the domain of supervised learning for over a decade. However, recent developments have been moving towards the onset of unsupervised methods for human activity recognition based research. Ambient Assisted Living (AAL) is another domain that's improvising with incoming research and nowadays even social media influencers have been very committed towards the idea of living life in a better, healthier manner. Wearable technology with the help of improvements in the field of pervasive computing has soared greater heights in this decade and it's only going to get bigger from here. Gabrielle's team of researchers (Civitarese et al.; 2019) performed significant research in HAR, particularly AAL in the year 2019, wherein, they introduced a new framework that comprised of an approach that supported active(dynamic) learning in a collective manner. The analysis was performed on eight complex activities and it was observed the classification worked best with the random forest classifier in comparison to Support Vector Machines, Naive Bayes, and Multi-Layer Perceptron(MLP). The algorithm was called newNectar and it gave great values of 0.87 each for precision and recall.

The research performed in this project uses Random Forest Classifier. A few other researchers also preferred the use of Random Forests for the use for HAR in comparison to other classifiers.

2.1 An Exploratory Survey on Machine Learning Techniques used in Human Activity Recognition

The research conducted by (Gao et al.; 2019) and team have highlighted the urgency and progress of HAR in terms of its applications. The research proposes the use of an algorithm that is based on the Stacking Denoising Autoencoder(SDAE) and LightGBM. The proposed approach worked the best with an average accuracy of 95.99 % against Convolutional Neural Networks, XGBoost, Statistical feature selection and SDAE alone. (Vaizman et al.; 2018a) and team who are creators of the 'Extrasensory Dataset' (Vaizman and Ellis; 2017) used in this research are the inspiration for the research that is performed. Vaizman and team have highlighted the importance of HAR in the field of analytics, surveillance, medical science and have utilized various machine learning algorithms for classification with single senor fusion, early sensor fusion and late fusion averaging for comparing the results of the classification. They concluded with multimodal sensor classification to be better over single sensor classification for HAR, using the logistic regression classifier. [1] (Li et al.; 2019) and team researched on heart disease and its diagnosis. Clustering techniques are one of the most commonly used methods to analyze biomedical data, especially, heart disease diagnosis. After comparing their proposed method of anamoly threshold, with many algorithms like SVM, Naive Bayes, XGBoost, and RandomForest, they concluded that a model classifier that was tree-based like XGBoost performed the best in arrhytmia predicion with an added advantages of parallel processing and regularization. (Gani et al.; 2019) researched about HAR that was based on a ystem that light weight and smartphone based, wherein the sensor was robust with respect to computational consumption. The system data was examined with the help of chaos theory and dynamic systems. They used Gaussian Mixture Models(GMM) for tracking location-centred activities of the 'Hajj' for geographical tracking of pilgrim activities. (Xu et al.; 2019) proposed a machine learning model on HAR analysis,wherein, they used ensemble learning models to classify human activities. The algorithms they used were XGBoost (Extreme Gradient Boosting), Random Forests, Softmax Regression and Extra trees. With the help of these, they made a model by the name CElearning. XGBoost performed significantly with mean values of 94.33 % and 90.87 %, with a standard of deviation 0.00 % each for automatic and handcrafted feature extraction respectively. The ensemble worked best against SVM, CNN, and SAE on multiple datasets.(Ronao and Cho; 2016) leveraged the potential of deep neural networks for HAR on sensor-based data, wherein, they used 'convnet', by making the most of a plethora of inherent aspects of human activities and time series impulses (1D) and convnets gave an accuracy of 94.79 % on test data and 95.75 % when coupled with data from temporal FFT(fast fourier transform). (Chetty et al.; 2015) and their team of researchers worked on HAR by

[1] http://extrasensory.ucsd.edu/

using five fold cross-validation for data partitioning. They used random forest classifier and analyzed various aspects like FPR(False Positive Rate), TPR(True Positive Rate), f-measure, ROC, RC(Recall) and Precision. They chose Random Forest(RF) as the algorithm worked best given an accuracy of 96.30 %. The team suggested that they would work on an unsupervised method in the future, like (Civitarese et al.; 2019) with active collaborative learning. Other researchers like (Wen and Zhong; 2015) and (Civitarese et al.; 2019) bring forth the importance of performing unsupervised human acitivity recognition by highlighting the weak points in supervised machine learning like the aspect of it being a subject-centric method. But, however, even Unsupervised methods result in poor answers with scant quantities of data and bigger clustering overhead. (Wen and Zhong; 2015) thereby proposed using a combination of unsupervised(unlabelled) and supervised(labelled) methodologies. In comparison to traditional HAR systems, their proposed system worked better with smaller of amounts of labelled input data.

(Lee et al.; 2019), L. Saidy and Fitri analyzed data that was extracted from a 'smart chair', a device that is capable of detecting and classifying five common day-to-day activities. The chair had sensors working hand-in-hand with Raspberry Pi for collection of raw data. Classification algorithms like random forest, extremely randomized trees(ERT) and SVM were used and ERT performed the best with 98% accuracy. (Hu et al.; 2019) and his fellow researchers worked on novel technique, which was, feature-incremental method of learning for activity recognition on sensor-based data called the 'Feature Incremental Random Forest' or FIRF. They used two specific components called the MIDGS(mutual information based diversity generation strategy) and FITGM(feature incremental tree growing mechanism) and FIRF performed significantly with better efficiency than the state-of-the-art methods. Another domain wherein, HAR is widely used is the domain of Sports. (Balli et al.; 2019) analyzed data taken from a smart watch sensors with the help of a hybrid model of random forest and PCA wherein random forest performed the best classification. (Mehrang et al.; 2018) and team experimented with data from wrist-based ubiquitous device and used the RF classifier which gave an accuracy of 89.6 with a fluctuation of 3.9 % wherein the size of the forest was 64 trees with signal segments of 13-s and a 90 % overlap. (Nweke et al.; 2018) and his team worked on data fusion and multi-class classification systems for HAR and Health Monitoring wherein they used Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA) and a few other methods and neural networks for deep learning features, followed by support vector machines(SVM) for classification. They concluded that deep learning models needed to be used further for better classification of activities.

(Subasi et al.; 2018) and his team of researchers recommended ensemble classifiers that stem from the Adaboost algorithm brought about a sigifincant growth in the overall performance of a HAR mode. (Galvn-Tejada et al.; 2018) used feature extraction and random forest for the classification of a human's indoor location. For validating results from the RF model, the out of bag(OOB) error was used as its unbiased estimator of the true prediction error. The study conducted by (Gaikwad et al.; 2019) and team worked on the efficacious FPGA execution of Multilayer Perceptron(MLP) for real-time classification of Human Activities in the year 2019. The reserachers utilized a hardware-based system for performing HAR classification instead of relying on the traditional smartphone-embedded sensors, wherein, they classified a data comprising of more than 7700 features for 20 subjects. The reserach concluded with an observation that the accuracy of the MLP model started decaying below a 16-bits models wherein decrements of between 0.9% and 3.33% in terms of accuracy were observed, specifically due to the switching between data formats. (Narang et al.; 2018) and (Sarcevic et al.; 2017) are two more studies that quite recently utilized MLP for classification. (Narang et al.; 2018) and team concluded that ANN(Artificial Neural Networks performed better in comparison to a SVM model for EEG classification using MLP in combination with the Marquardy algorithm. (Sarcevic et al.; 2017) concluded that time-domain features (TDFs) worked better in terms of classification compared to frequency domain features(FDFs) utilizing Multilayer Perceptron(MLP). The highest accuracy in terms of classification they achieved was 91.75 % and 88.51 % with TDFs and FDFs respectively.

2.2 Human Activity Recognition from the perspective of Medical Data Science and Human Heathcare

The radical improvement of technology in the direction of utilizing multimodal sensors with the help of smartwatches, smartphones, wearable devices, Zenith-cameras and other ubiquitous devices has made Human Activity Recognition an area that is ought to be researched. Human Activity Recognition is spearheading the involvement of technology, artificial intelligence, analytics in the field of Medical Data Science. The research performed by (Alvarez et al.; 2018) on recognition of behavioural activities by utilizing multimodal sensor-based data for monitoring the health of patients with conditions like Alzheimers and Parkinsons. The approach that was followed by (Alvarez et al.; 2018) and team was crucial for patients having cognitive disability. This was coupled along with patient suggestions collaborated together into an end-to-end system for healthcare monitoring. Sparse Encoder (SAE) was utilized for feature representation and cluster validation was done with the help of Logistic Regression(LR) and SVM classifiers. As a future suggestion, FoG detection with the help of proprietary data was recommended for better classification. (Witt et al.; n.d.) worked on the subject of healthcare with the help of data extracted from sensor-based wearables. The PubMed dataset they used consists of multiple records of algorithms of famous wearable tech-giants like Fitbit(now owned by Google), and Apple, and data from electrocardio(ECG) and accelerometry algorithm. The team performed regression based analysis, time series based analysis for examining parameters like sleep stage, fall counts, steo counts, and many more.

2.2.1 Identify stressful states using HAR with the help of machine learning:

High-levels of stress has taken a toll on the lives of numerous human beings in the past decade[2]. According to the Centers for Disease Control (CDC), approximately 110 million lives are lost in the battle with stress on a yearly basis. A study by (Can S; 2019) and Cem Ersoy worked on the aspect of stress detection with the help of wearable sensors wherein they used heart rate activity, accelerometer, signals and skin conductance to classify stressful, higher cognitive load, relaxed activities. They used PCA with Linear Discriminant Analysis(LDA) and SVM, kNN, Logistic Regression, MLP and Random Forest wherein Random Forest gave the highest accuracy of 88.26 % followed by kNN and MLP.(Ghaderi and Frounchi; 2015) performed feature extraction and utilized PCA and Bayesian Learning wherein the model gave an accuracy of 99 % in a classification involving 3 classes. (Said et al.; 2019) also analyzed routine stress using data extracted from wearable ubiquitous devices. Bayesian Networks performed the best 82% of times. Looking at the psychological aspect of health care, (Garcia-ceja et al.; 2018) used,

[2] https://www.slma.cc/the-science-of-stress/

Random Forest, XGBoost and Multilayer Perceptrons. The three algorithms utilized for analyzing sensor-based data of multiple individuals to perform classification of various human activities.

3 Methodology

Data Mining is the process of exploring big datasets and looking for unknown patterns or different relationships/behaviors by utilizing the power of machine learning algorithms, statistical tests, and database systems. In simpler words finding meaningful information out of raw data by analyzing various patterns of information is data mining[3] [4].

This research lays its foundations in getting the best classification method for categorizing human activities based on sensor-based data. The fundamental goal of this research is to build a model that can successfully classify activities in-the-wild i.e. without any supervision on the basis of data collected from ubiquitous devices that is a Pebble watch and smartphones, both from the Android and Apple domains.

The methodology chosen for this research is CRISP-DM. According to (Sika; 2016), CRISP-DM is the most broadly used while implementing data mining algorithms for research in a variety of fields.

The following figure will give you a visual representation of the CRISP-DM methodology[5]:

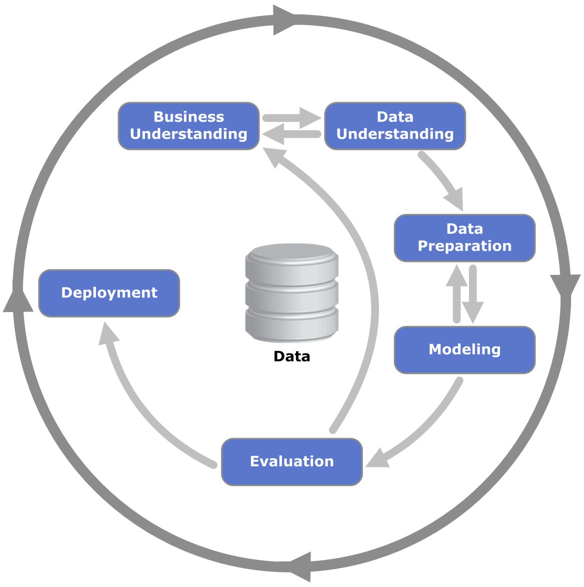


Figure 1: CRISP-DM

3.1 Business Understanding

[6] With the onset of the era of wearable devices and pervasive computing, the population became more conscious about their physical health, mental stability given the amount of

[3] https://upload.wikimedia.org/wikipedia/commons/b/b9/CRISP-DM$_{P}rocess_{D}iagram.png$

[4] https://economictimes.indiatimes.com/definition/data-mining

[5] https://upload.wikimedia.org/wikipedia/commons/b/b9/CRISP-DM$_{P}rocess_{D}iagram.png$

[6] https://www.slma.cc/the-science-of-stress/

stress levels (Said et al.; 2019).

The field of Data Mining is the process of exploring big datasets and looking for unknown patterns or different relationships/behaviors by utilizing the power of machine learning algorithms, statistical tests, and database systems.

There are a multitude of approaches that can be leveraged to perform mining on a given set of data. Identifying the one that goes with your set of requirements is a product of comprehensive, thourough surveying and research about the (Fayyad et al.; 1996) and his fellow researchers looked at data mining as an intergral aspect of the KDD which stands for Knowledge Discovery in Data Mining. The other approaches are called as SEMMA and CRISP-DM, which stand fror Sample, Explore, Modify, Model and Asses, and Cross-Industry Standard Process for Data Mining, respectively. CRISP-DM was developed by a team of experts from remarkable firms in the world of analytics like Ohra, NCR, Dailmer, Teradata and SPSS.

The methodology chosen for this research was **CRISP-DM**. According to (Sika, 2016), CRISP-DM is the most broadly used while implementing data mining algorithms for research in a variety of fields.

The initial step of CRISP-DM is understanding and analyzing the business requirements. Human activity recognition forms the core part of the application of wearable technologies, pervasive computing and ubiquitous devices like the smartphones, smartwatches, ambient living powered smart houses and many more. The research primarily focuses on context recognition for activity classification.

There are two major business applications where human activity recognition can become a headstrong pillar of, namely, the field of medical science and wearable technology.

3.1.1 Human Activity Recognition - A boon to the healthcare industry

The section on related work clearly explains how HAR has been benefiting the world of healthcare and medical science and why researchers have been analyzing multimodal sensor data to understand human activity. (Said et al.; 2019) have addressed the issue of mental health and looked at the issue from a data science perspective and analyzed how stress can negatively affect one's entire life. Patients who suffer from [7] Parkinson's disease (Alvarez et al.; 2018) can be helped by identifying their actions wherein both motor-related and non-motor issues of the patients can be identified and thus dealt with, by using sensor based wearable devices. Also, individuals with the Alzheimer's disease could be helped with the help of a sensor-based device in order to identify the onset of a problem beforehand by reading sensor ratings. (Ong et al.; 2017) worked on a 'fall detection mechanism' which was designed for the elderly. The list of applications goes on, for example, (Witt et al.; n.d.) and his team spoke about heart rate sensors in their research, whereas a lot of researchers have looked at [8] sleep patterns [9]of individuals using sensor based data. (de Chazal et al.; 2008) assessed the sleep and wake patterns with the help of a non-contact sensor called the biomotion sensor. (Vaizman and Ellis; 2017) and team analyzed daily activities using smartphone sensors, whereas, (Garcia-ceja et al.; 2018) probed in the areas of lifestyle and other day-to-day practices.

[7] https://parkinsons.ie/Professionals$_{W}hat_{I}s_{P}arkinsons$

[8] http://www.scitecheuropa.eu/sensors-monitor-sleeping-patterns/85872/

[9] https://www.tuck.com/how-sleep-trackers-work/

3.1.2 Human Activity Recognition and the World of Wearables

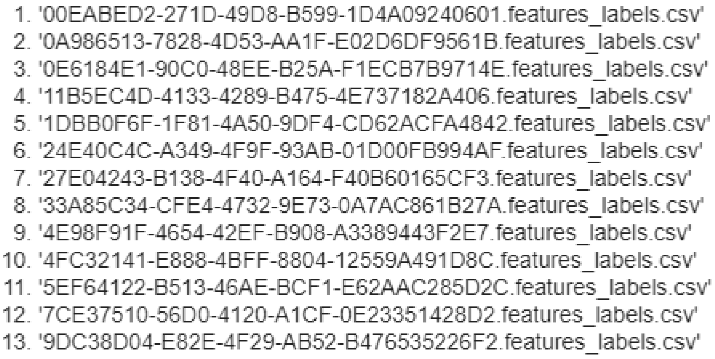
In the world of pervasive computing, crucial insights gathered by understanding a variety of patterns from the available data and finding new relationships between the available attributes is something that is being leveraged by the data science community. Behavioral patterns, activity patterns and sleep patterns can give crucial insights about the overall health of an individual with the help of deep neural networks or machine learning models. Multiple research projects like (Sutton et al.; 2003), (Kune et al.; 2016), (Alvarez et al.; 2018), (Civitarese et al.; 2019), (Vaizman et al.; 2018a), (Ignatov and Strijov; 2015), have been performed by numerous researchers on Human Activity Recognition(HAR) for bringing improvements in wearable devices, sensor identification devices, fitness smart bands, heart-rate monitoring(hrm) devices. (Sutton et al.; 2003), (Kune et al.; 2016), (Alvarez et al.; 2018), (Civitarese et al.; 2019), (Vaizman et al.; 2018a), (Ignatov and Strijov; 2015).

The issues from the previous research performed has been utilized and different models and analyses have been performed to build a thorough classification model.

1. Classifying human activities in-the-wild from multimodal sensor-based data.
2. Predicting activities based on previous training.

Exploratory data analysis was performed and a cluster of similar activities and the most relevant ones were chosen based number of minutes available and finally 4 classes of activity were chosen.

3.2 Data Understanding



```
 1. '00EABED2-271D-49D8-B599-1D4A09240601.features_labels.csv'
 2. '0A986513-7828-4D53-AA1F-E02D6DF9561B.features_labels.csv'
 3. '0E6184E1-90C0-48EE-B25A-F1ECB7B9714E.features_labels.csv'
 4. '11B5EC4D-4133-4289-B475-4E737182A406.features_labels.csv'
 5. '1DBB0F6F-1F81-4A50-9DF4-CD62ACFA4842.features_labels.csv'
 6. '24E40C4C-A349-4F9F-93AB-01D00FB994AF.features_labels.csv'
 7. '27E04243-B138-4F40-A164-F40B60165CF3.features_labels.csv'
 8. '33A85C34-CFE4-4732-9E73-0A7AC861B27A.features_labels.csv'
 9. '4E98F91F-4654-42EF-B908-A3389443F2E7.features_labels.csv'
10. '4FC32141-E888-4BFF-8804-12559A491D8C.features_labels.csv'
11. '5EF64122-B513-46AE-BCF1-E62AAC285D2C.features_labels.csv'
12. '7CE37510-56D0-4120-A1CF-0E23351428D2.features_labels.csv'
13. '9DC38D04-E82E-4F29-AB52-B476535226F2.features_labels.csv'
```

Figure 2: Number of users chosen for the analysis

13 out of 60 users were chosen as a sample of the complete dataset as. Every user has over 2600 rows of data distributed over 278 features spread between various sensors. The sensors utilized are accelerometer, gyroscope, magnetometer, location, watch accelerometer, latitude, longitude data, and audio sensors.

The data collection was performed on real-life scenarios wherein the individual is in his/her natural element, for example, their place of work, or their abode, etc. There were no instructions given to any of the participants and their complete consent and privacy were maintained throughout the entire process. This is one of the primary reasons why there wasn't any use of camera sensors like the use of Sense Cam's or the GoPro cameras. The individuals were in their element without any intervention in their freedom to perform their daily chores.

Researchers like (Civitarese et al.; 2019) suggest the use of user-reported labeling to classify various activities. (Vaizman and Ellis; 2017) and team, similarly, had a massive set of pre-labeled actions or activities with the option for the users to self-report an activity that wasn't detected or incorrectly detected by the pervasive devices.

3.3 Data Preparation

Data collection is certainly one of the most crucial aspects about this dataset (Vaizman and Ellis; 2017). Situations in-the-wild, or in simple words, in the individuals daily life were analyzed to get information about their activities. This helped in the model be unbiased towards to a supervised method of data collection.

About the Extrasensory dataset :
The dataset was created with the help of data collected 'in-the-wild' i.e. from an uncontrolled and unsupervised environment with the help of an app created for both the iPhone and Android environments. The application consists of pre-defined labels and an option for the users to self-report activities.

There was a separate environment dedicated for the Pebble watch that blended with the aforementioned environments. Katherine Ellis along with Yonatan Vaizmann gathered this data in the year 2015-2016. The dataset consists of 3,00,000 minutes (examples) gathered from sensors in ubiquitous devices of 60 different individuals with data that was collected in-the-wild i.e. the natural environment(without any supervision). Multiple sensor ratings were gathered to create the entirety of the dataset. A dynamic mode of data collection was also utilized and was called 'Active Feedback View' wherein the users reported any incorrect labels (Vaizman et al.; 2018b).

Feature Extraction: There are 4 different feature extraction methods used for treating the features in order to extract the most relevant features for the analysis.

1. Simple feature elimination by correlation
2. Boruta Algorithm in R from the package DwMR
3. Looking at skewness and kurtosis
4. Variable importance plot using RF
5. Using Agglomerative hierarchial Clustering

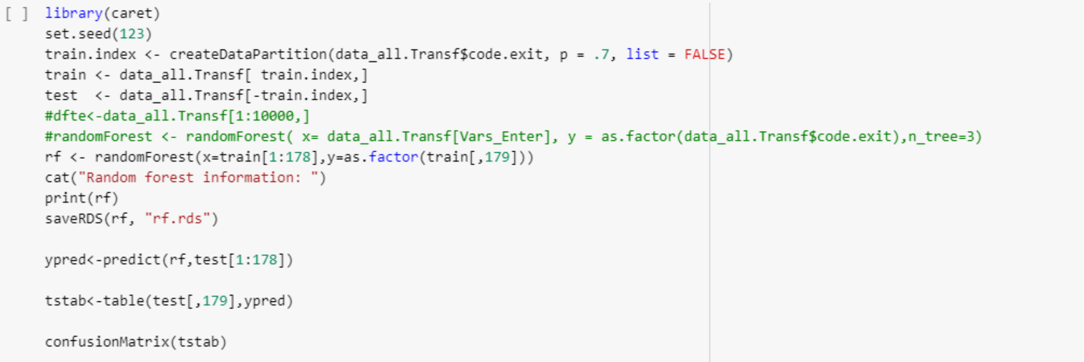
3.4 Modelling

In this research on Human Activity Recognition, post the application of multiple feature selection algorithms and multiple research surveys, three machine learning algorithms were applied to the dataset, namely, Random Forest, Extreme Gradient Boosting, and Mutilayer Perceptron (MLP) to identify between different activities.

3.4.1 Random Forest

Decision trees are used for solving problems of decision making and splitting from root to leaf level, wherein, data is split on features as per higher information gain, until the leaf nodes are pure. However, for multi-nominal class classification, decision trees ensemble into multiple trees into a forest that is random, hence the name Random Forest (Mehrang et al.; 2018) (Aliwy and Ameer; 2017). Four labelled activities were chosen for the activity classification (Mao et al.; 2017). [10]

Following figure Figure 3. shows the code for Random Forest:



```
library(caret)
set.seed(123)
train.index <- createDataPartition(data_all.Transf$code.exit, p = .7, list = FALSE)
train <- data_all.Transf[ train.index,]
test  <- data_all.Transf[-train.index,]
#dfte<-data_all.Transf[1:10000,]
#randomForest <- randomForest( x= data_all.Transf[Vars_Enter], y = as.factor(data_all.Transf$code.exit),n_tree=3)
rf <- randomForest(x=train[1:178],y=as.factor(train[,179]))
cat("Random forest information: ")
print(rf)
saveRDS(rf, "rf.rds")

ypred<-predict(rf,test[1:178])

tstab<-table(test[,179],ypred)

confusionMatrix(tstab)
```

Figure 3: Random Forest

Random Forests is an ensemble method in machine learning algorithms that follows the principle of supervised machine learning. It also follows the principle of bagging, wherein multiple decision trees run in parallel, wherein, no single tree has any interaction with another tree. [11]

3.4.2 eXtreme Gradient Boosting (XGBoost)

The name Extreme Gradient Boosting or XGBoost comes from the aspect that it performs e**X**treme **G**radient **B**oosting. It also comes from the aim to surpass the limitations of computational assets for boosted-tree algorithms. (Gao et al.; 2019) (Li et al.; 2019) and (Garcia-ceja et al.; 2018) worked on HAR multi-class classification models wherein XGBoost performed the best by giving the highest accuracy. [12]

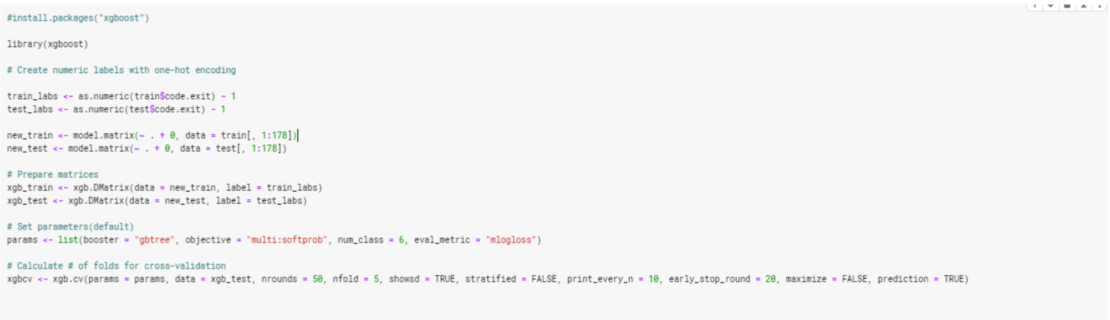
The principle of boosting is exactly different from the principle of bagging. Boosting is meant to boost the performance of weak learners and therefore while boosting, the output of 1 tree is used as the input for the next one. There is a high amount of interaction between trees.

Following figure Figure 4. shows the code for XGBoost:

[10] https://machinelearningmastery.com/bagging-and-random-forest-ensemble-algorithms-for-machine-learning/

[11] https://towardsdatascience.com/basic-ensemble-learning-random-forest-adaboost-gradient-boosting-step-by-step-explained-95d49d1e2725

[12] https://machinelearningmastery.com/gentle-introduction-xgboost-applied-machine-learning/

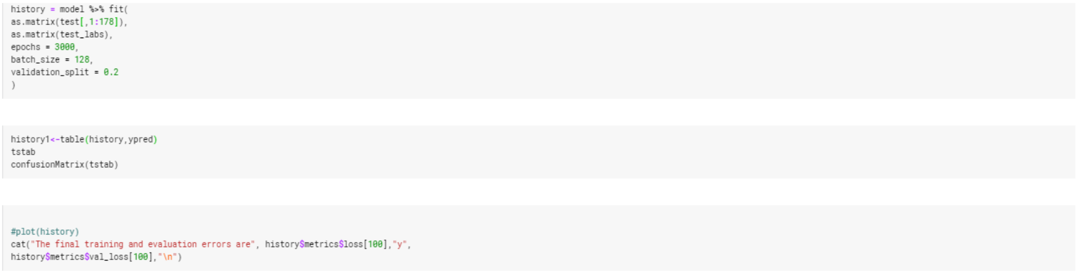


```
#install.packages("xgboost")

library(xgboost)

# Create numeric labels with one-hot encoding

train_labs <- as.numeric(train$code.exit) - 1
test_labs <- as.numeric(test$code.exit) - 1

new_train <- model.matrix(~ . + 0, data = train[, 1:178])
new_test <- model.matrix(~ . + 0, data = test[, 1:178])

# Prepare matrices
xgb_train <- xgb.DMatrix(data = new_train, label = train_labs)
xgb_test <- xgb.DMatrix(data = new_test, label = test_labs)

# Set parameters(default)
params <- list(booster = "gbtree", objective = "multi:softprob", num_class = 6, eval_metric = "mlogloss")

# Calculate # of folds for cross-validation
xgbcv <- xgb.cv(params = params, data = xgb_test, nrounds = 50, nfold = 5, showsd = TRUE, stratified = FALSE, print_every_n = 10, early_stop_round = 20, maximize = FALSE, prediction = TRUE)
```

Figure 4: XGBoost

3.4.3 Multilayer Perceptron

Artificial Neural Networks have class called as a Multilayer Perceptron (MLP). This is why MLP's are sometimes refered to as a feedforward ANN and sometimes they're informally referred as 'vanilla' neural network, in cases wherein only a single hidden layer exists. MLP's comprise of 3 layers of nodes, namely, an input later, followed by hidden and output layers, respectively. These are fully connected artificial neural networks, wherein, one node of the input layer is connected with a particular weight to each node in the following layer. [13] MLP's are great algorithms by mathematical and regression analysis. Since, classification is a special case of regression, that is, when the output variable is categorical. (Sarcevic et al.; 2017) and (Narang et al.; 2018) have both utilized the Multilayer Perceptron for Multi-class classification. 'Softmax' activation layer was used as the classification is a multi-nominal classification on HAR. The following figure Figure 5. shows the code for MLP: [14]



```
history = model %>% fit(
as.matrix(test[,1:178]),
as.matrix(test_labs),
epochs = 3000,
batch_size = 128,
validation_split = 0.2
)
```

```
history1<-table(history,ypred)
tstab
confusionMatrix(tstab)
```

```
#plot(history)
cat("The final training and evaluation errors are", history$metrics$loss[100],"y",
history$metrics$val_loss[100],"\n")
```

Figure 5: Multilayer Perceptron

Further, data normalization was the step taken for proceeding with the application stage.

3.4.4 Normalization and Model Deployment

Normalizing the Data

[13] https://en.wikipedia.org/wiki/Multilayer$_{p}erceptron$

[14] https://towardsdatascience.com/17-rules-of-thumb-for-building-a-neural-network-93356f9930af

One of the most principal aspects before applying the chosen machine learning algorithms was normalizing the data.

Applying Machine Learning Models

Three machine learning algorithms were applied to the normalized data, namely, Random Forest, XGBoost, and an Artificial Neural Network called the Multilayer Perceptron.

1. **Train-Test Split**: Before applying any machine learning models, the data was split into training and testing sets for applying the algorithms and checking the predictions. The data was split as 70% for training and 30 % for testing.

2. **Random Forest**:

   Random Forest was used to classify the chosen classes of activities, namely, label.FIX_walking, label.LYING_DOWN,label.SITTING, and Other activity.

   Here, the random forest algorithm was chosen for classification post feature normalization as it only works with normalized data. Random Forests are ensemble methods wherein a multitude of decision trees work together without any interaction amidst each other towards the goal of classification. Random Forests work on the concept of Bootstrap Aggregation (Bagging). Bagging was used in order to avoid issues with over-fitting and variance related issues.

3. **XGBoost**:

   Extreme Gradient Boosting is an ensemble method that has been used for performing multi-class classification. The parameters chosen for performing the analysis were booster as 'gbtree', objective as 'multi:softprob', num_class as 4, and the eval_metric as 'mlogloss'.

   The meaning of every parameter is as follows: The parameters objective and eval_metric are used to tell the algorithm that probabilistic classification is being performed for a multi-class output variable. The use of 'multi:softprob' and 'num_class' go hand-in-hand for letting the model know about the number of classes. The remaining parameters utilized are 'nrounds' and 'prediction, wherein, nrounds notifies the algorithm about the number of times it would iterate. We tuned the hyperparameter for experimental purposes to see the effect on the overall accuracy. Stratified sampling and '5' fold cross-validation were utilized, with an early stopping parameter set at 20.

4. **Multilayer Perceptron:**

   The **keras** library was loaded in the Kaggle kernel and keras_model_sequential was loaded. The dense layer was loaded with relu as the activation layer with an input shape of 178 that is the total number of features and the softmax activation layer. A **batch size** of 128 was used at 1000 and 3000 epochs respectively. A batch size is nothing but, a smaller cut or subset of the data, that is used in a single iteration. A smaller batch size was utilized in order to maintain the model quality. In cases of higher batch sizes, there is a possibility that the model quality may degrade. The layer dense unit value '4' represents the number of classes chosen for the classification.

Sparse categorical entropy was used as the loss function as the classification was had multiple classes and **'adam optimizer'** was used as the optimizer. Accuracy was the selected metric for the analysis.

3.5 Evaluation

Applying machine learning models on a dataset is one of the crucial phases of data science. However, evaluating the model based on various aspects is the key to getting the best understanding of various relationships between the model parameters, like, input variables and output variables. Accuracy, Precision, Recall, Kappa value, and f1-score were the multiple parameters taken into consideration for performing a 4-class classification on 179 features. After a thorough literature survey on machine learning approaches followed in the field of human activity recognition, various algorithms were taken into consideration. Random Forests, XGBoost, and the Artificial Neural Network, Multilayer Perceptron were selected. These algorithms were run at different values of epochs, number of trees and cross-validation for understanding how iterations worked on the data.

4 Ethical Considerations

The research performed by me with inspiration from Vaizman and Ellis (2017) and the team does not hamper any ethical aspects of any individual. No private information was collected and used while the entirety of this research. Individuals were classified and annotated with the help of unique IDs. [15]

5 Design Specification

Figure 6. shows the representation of design specification.

The three layers of the entire research project are database layer, application layer, and presentation layer. The database layer talks about the Data Extraction, Data Transformation(Preparation) and Feature Selection. The Application layer dives deep into the aspect of machine learning models and the model evaluation. The presentation layer speaks about the visual representations in the research.

[15] http://extrasensory.ucsd.edu/

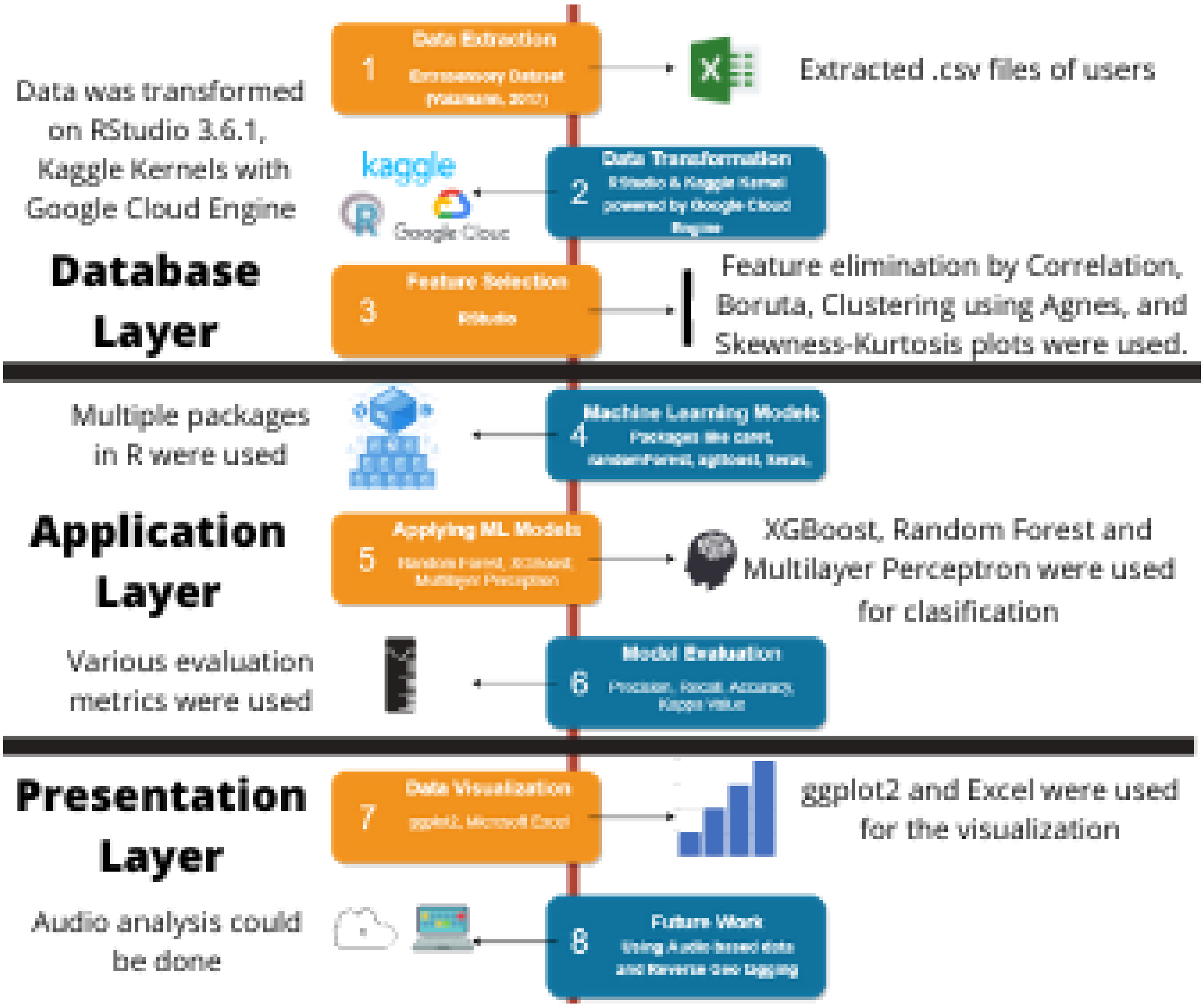

Data was transformed on RStudio 3.6.1, Kaggle Kernels with Google Cloud Engine
Database Layer
1
Data Extraction
Extrasensory Dataset (Vaizman, 2017)
Extracted .csv files of users
kaggle
Google Cloud
2
Data Transformation
RStudio & Kaggle Kernel powered by Google Cloud Engine
3
Feature Selection
RStudio
Feature elimination by Correlation, Boruta, Clustering using Agnes, and Skewness-Kurtosis plots were used.
Multiple packages in R were used
4
Machine Learning Models
Packages like caret, randomForest, xgboost, keras.
Application Layer
5
Applying ML Models
Random Forest, XGBoost, Multilayer Perceptron
XGBoost, Random Forest and Multilayer Perceptron were used for clasification
Various evaluation metrics were used
6
Model Evaluation
Precision, Recall, Accuracy, Kappa Value
Presentation Layer
7
Data Visualization
ggplot2, Microsoft Excel
ggplot2 and Excel were used for the visualization
Audio analysis could be done
8
Future Work
Using Audio based data and Reverse Geo tagging

Figure 6: Design Specification

6 Implementation

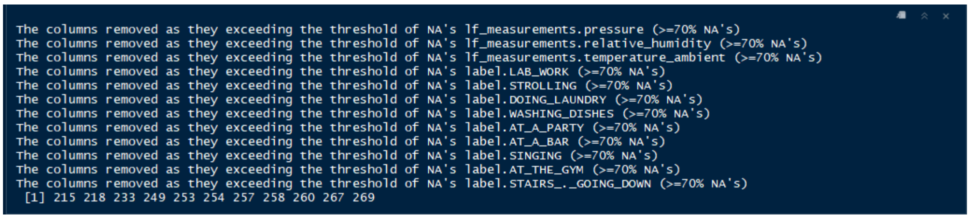
Exploring the Extrasensory Dataset:

The analysis was performed using R in collaboration with Kaggle Kernel Notebook and RStudio which is an Infrastructure as a Service (IaaS). IaaS are online services that are equipped with multiple high-level software systems and APIs. The Kaggle Kernels were chosen as they come equipped Quad Core CPUs clocked at 3.0 GHz, 16 GB of RAM and Google Cloud Engine GPU enabled with Google Drive memory. The choice of using Kaggle Kernels greatly benefits the research as it empowered better functionality of neural networks, especially, keras, with the GPU powered on.

Multiple R libraries were used and CSV files of 13 random users were selected and loaded on to the Kaggle Kernel using string extract wherein a regular expression(regex) was used to pull in all the files in the working directory. Before initialing the analysis, the dataset was treated with multiple transformations.

Handling missing values Firstly the features that had complete null values were removed. Further feature with more than 70 % values as null values were eliminated. And the remaining values were replaced with mean values. 11 features were removed after

handling NA values.
Figure 7 will give you a quick insight of the features that were eliminated.



```
The columns removed as they exceeding the threshold of NA's lf_measurements.pressure (>=70% NA's)
The columns removed as they exceeding the threshold of NA's lf_measurements.relative_humidity (>=70% NA's)
The columns removed as they exceeding the threshold of NA's lf_measurements.temperature_ambient (>=70% NA's)
The columns removed as they exceeding the threshold of NA's label.LAB_WORK (>=70% NA's)
The columns removed as they exceeding the threshold of NA's label.STROLLING (>=70% NA's)
The columns removed as they exceeding the threshold of NA's label.DOING_LAUNDRY (>=70% NA's)
The columns removed as they exceeding the threshold of NA's label.WASHING_DISHES (>=70% NA's)
The columns removed as they exceeding the threshold of NA's label.AT_A_PARTY (>=70% NA's)
The columns removed as they exceeding the threshold of NA's label.AT_A_BAR (>=70% NA's)
The columns removed as they exceeding the threshold of NA's label.SINGING (>=70% NA's)
The columns removed as they exceeding the threshold of NA's label.AT_THE_GYM (>=70% NA's)
The columns removed as they exceeding the threshold of NA's label.STAIRS_._GOING_DOWN (>=70% NA's)
 [1] 215 218 233 249 253 254 257 258 260 267 269
```

Figure 7: Handling Missing Values

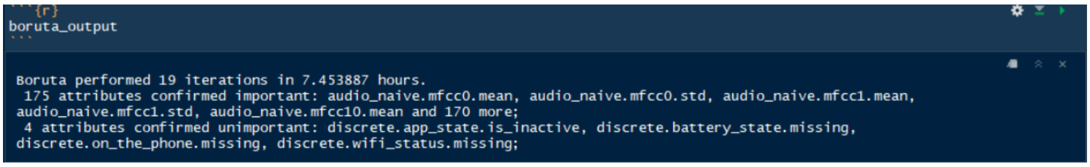
After handling missing values, the features were treated with the help of various techniques.

Feature Engineering

Feature Engineering is one of the most essential steps in the process of performing data mining.

Techniques Used:

1. A technique called as the **Boruta Algorithm in R** was used as an experimental feature selection method. It eliminated 4 features out of the remaining features. (Kursa; 2018) and have mentioned that Boruta is one of the fastest feature elimination methods and is an algorithm for those who are in a hurry. Following figure Figure 8 is the output of the Boruta Algorithm:



~~~
```{r}
boruta_output
```

Boruta performed 19 iterations in 7.453887 hours.
 175 attributes confirmed important: audio_naive.mfcc0.mean, audio_naive.mfcc0.std, audio_naive.mfcc1.mean,
audio_naive.mfcc1.std, audio_naive.mfcc10.mean and 170 more;
 4 attributes confirmed unimportant: discrete.app_state.is_inactive, discrete.battery_state.missing,
discrete.on_the_phone.missing, discrete.wifi_status.missing;
~~~

Figure 8: Boruta

2. Variable Importance Plot was calculated using Random Forest for feature selection with ntree = 100. The plot in figure Figure 9 was referred to look at the top 5 features of the dataset.

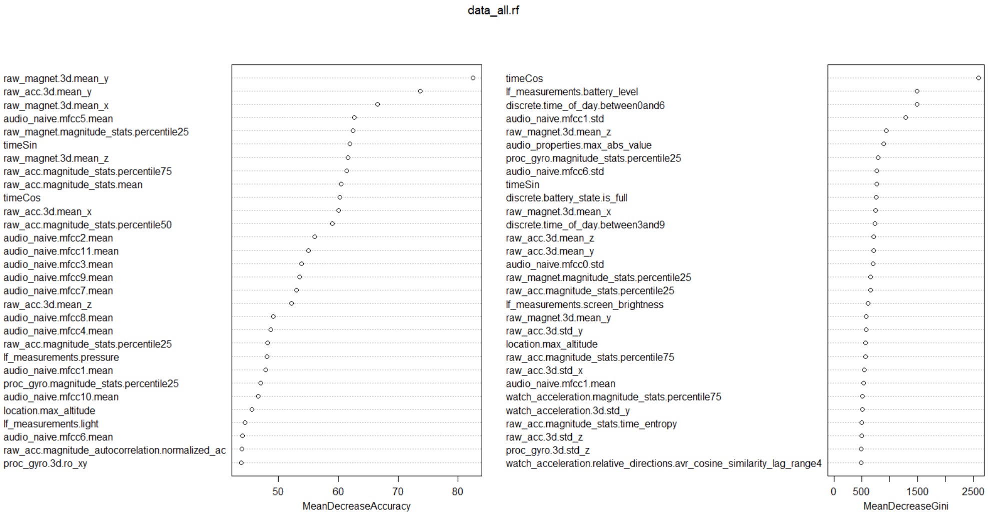


Figure 9: Variable Importance Plot using Random Forest

Both these analyses were performed to understand the various relationships amidst the independent input variables within themselves.

3. Understanding the data distribution of different sensors was performed to see how the data distribution. Statistical functions like skewness and kurtosis were utilized from the library called 'moments'. The top rows of the dataset were selected for looking at the data. A higher value of the kurtosis showed that the data does not follow a normal distribution visualized in figure Figure 10.

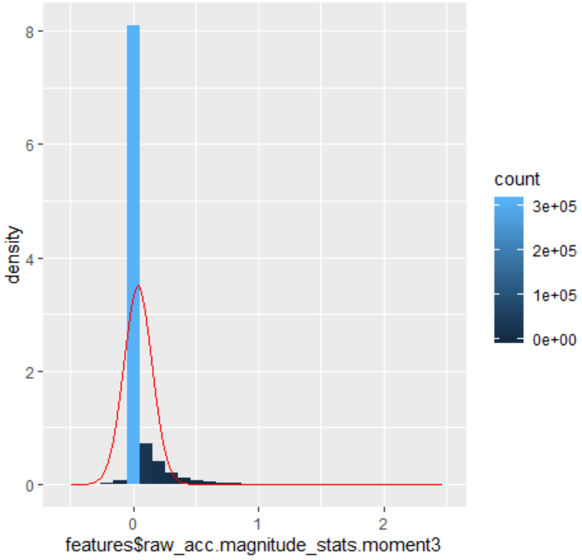


Figure 10: Distribution of sensor data

Another Q-Q plot was used to visualize the distribution of the data is depicted in Figure 11.

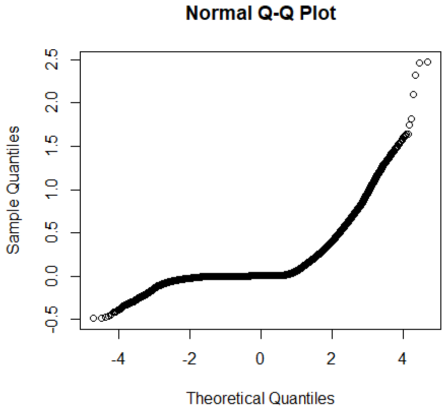


Figure 11: Distribution of data

4. Simple Elimination by Correlation

   At the initial stage, simple elimination using correlations was utilized. Total number of features present for the analysis were 280. However, post the comparison of correlation coefficients, correlation of 0.9 and above were extracted and removed for eradicating issues related to multi-collinearity by referring to Cohen's criterion. The algorithm of correlation

   Figure 12 below will help you get a better view of the entire correlation between different activities.

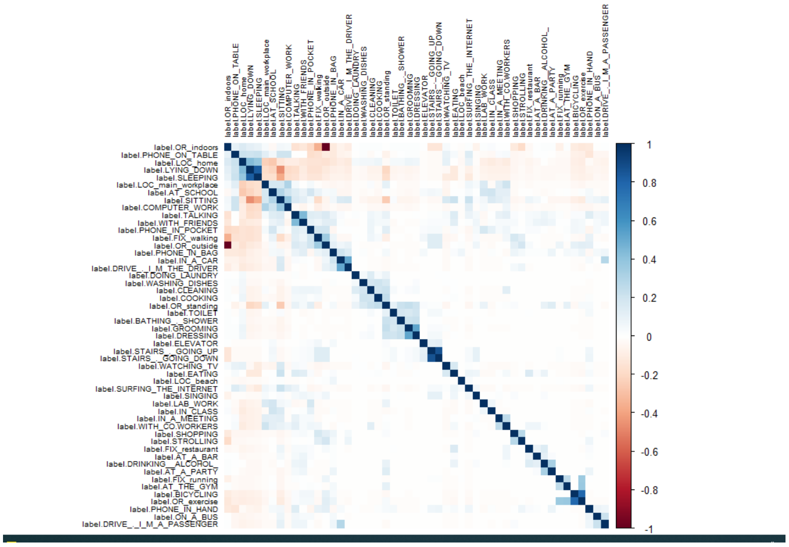


Figure 12: Correlation Plot

5. Using the Agglomerative method for performing hierarchical form of clustering. There were certain clusters that were formed here.

The cluster representation is done with the help of a dendogram visualization in figure Figure 13 :

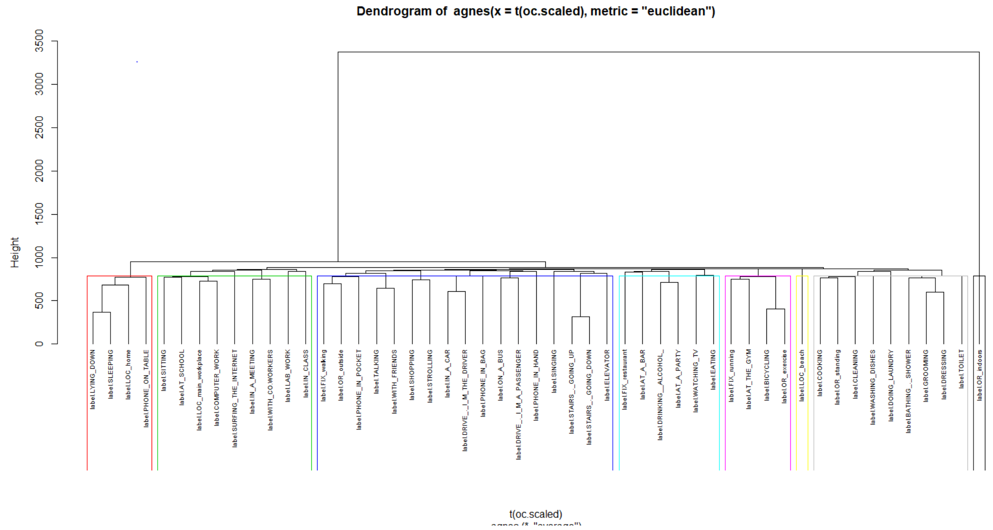


Figure 13: Agglomerative Hierarchial Clustering

Transforming the Timestamp feature: The foremost column timestamp , wherein, epoch UNIX timestamp tools for conversion were used to understand this column. Before beginning the transformation, two variables were created to represent cyclic time, one being a sine and another a cosine factor of time. This way we would not need the timestamp column for the evaluation. The goal was to convert the feature to a cyclic one, which indeed, is the nature of time to come back to the same time after 24 hours.

Notice that the separation between a point as 5 minutes prior and 5 minutes post the split is massive This is bothersome, as we need our model to learn that the timings 23:55 and 00:05 have a difference of 10 minutes, however, the way things are, those occasions will seem, by all accounts, to be 23 hours and 50 minutes separated! Sine only components and cosine only components still had issues, which is why a sine and a cosine aspect of timestamp were chosen to be fed to the machine learning algorithm in order to maintain the 24-hour cyclical of time.[16]

7 Evaluation

Model evaluation is the aspect wherein the end-user can generate recommendations out of the values provided by various metrics, parameters and plots after training and testing the model on the train and test sets. The models were evaluated on the basis of accuracy as the classifying metric.

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

$$F1-\text{score} = \frac{2*\text{Precision}*\text{Recall}}{\text{Precision}+\text{Recall}} \tag{2}$$

[16] https://ianlondon.github.io/blog/encoding-cyclical-features-24hour-time

Table 1: Summary

Algorithm	**Accuracy(%)**
Random Forest	89.17
XGBoost	90.65
Multilayer Perceptron	91.50

There were multiple other metrics to be considered. But, in terms of multi-class classification, accuracy is the metric chosen for model evaluation.

Why not Kappa values? [17]
Kappa values were not chosen as the evaluation metrics. This is because there have been instances wherein the incorrect prediction gets a high value of kappa. Therefore, kappa values were not considered as the metric.

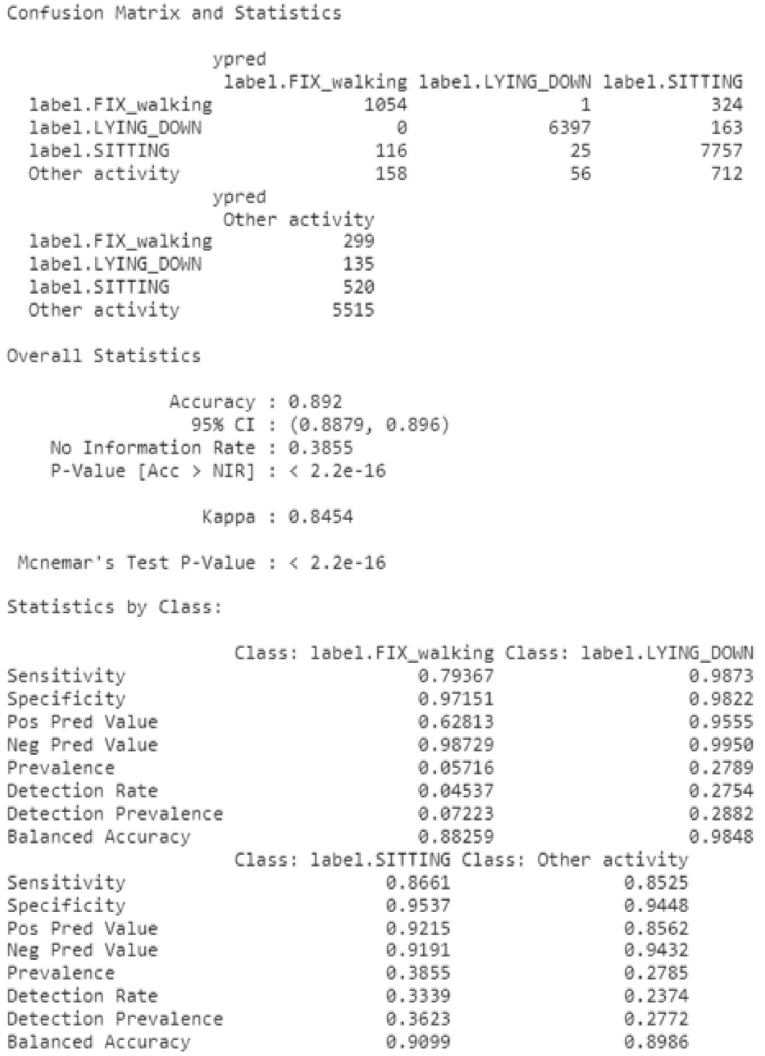
7.1 Experiment 1: Applying Random Forest to the Extrasensory dataset

The dataset was normalized before applying random forest to the set of features and labels for multi-class classification.

The normalized data was then split into two sets, training and testing, into 70 % and 30 %, respectively.

After applying random forest, we got: Accuracy: 89.17%, Kappa Value: 0.8454, Number of trees: 500, No. of variables tried at each split: 13, Out of Bag estimate of error rate: 10.6%,

[17] https://www.ncbi.nlm.nih.gov/pubmed/31557204



```
Confusion Matrix and Statistics

                    ypred
                     label.FIX_walking label.LYING_DOWN label.SITTING
  label.FIX_walking               1054                1           324
  label.LYING_DOWN                   0             6397           163
  label.SITTING                    116               25          7757
  Other activity                   158               56           712
                    ypred
                     Other activity
  label.FIX_walking             299
  label.LYING_DOWN              135
  label.SITTING                 520
  Other activity               5515

Overall Statistics

               Accuracy : 0.892
                 95% CI : (0.8879, 0.896)
    No Information Rate : 0.3855
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.8454

 Mcnemar's Test P-Value : < 2.2e-16

Statistics by Class:

                     Class: label.FIX_walking Class: label.LYING_DOWN
Sensitivity                           0.79367                  0.9873
Specificity                           0.97151                  0.9822
Pos Pred Value                        0.62813                  0.9555
Neg Pred Value                        0.98729                  0.9950
Prevalence                            0.05716                  0.2789
Detection Rate                        0.04537                  0.2754
Detection Prevalence                  0.07223                  0.2882
Balanced Accuracy                     0.88259                  0.9848
                     Class: label.SITTING Class: Other activity
Sensitivity                        0.8661                0.8525
Specificity                        0.9537                0.9448
Pos Pred Value                     0.9215                0.8562
Neg Pred Value                     0.9191                0.9432
Prevalence                         0.3855                0.2785
Detection Rate                     0.3339                0.2374
Detection Prevalence               0.3623                0.2772
Balanced Accuracy                  0.9099                0.8986
```

Figure 14: Confusion Matrix for Random Forest

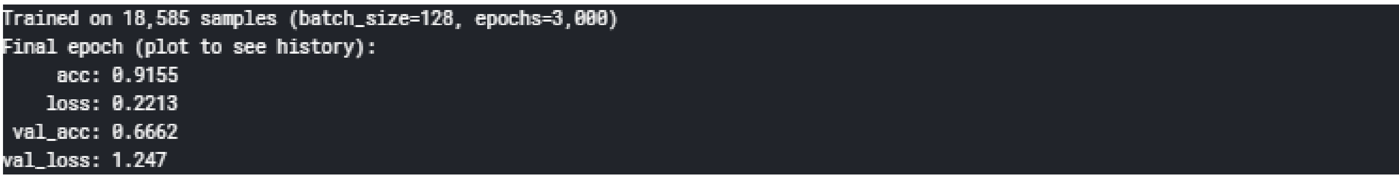
This means that our classification model was able to predict the trained class label correctly 89 % of the times. This was after the model was run through the ntree value at 50. Random Forest gave a weaker accuracy before feature selection using the multiple methods mentioned was performed. Therefore, the feature selection methods performed in the research are novel and are successful in improving the performance.

7.2 Experiment 2: Applying Multilayer Perceptron

Evaluating the Multilayer Perceptron

A perceptron is an algorithm used in ML that assists in providing information of classified answers for computing problems. Artificial Neural Networks(ANN) are non-linear models that are very easy to interpret in comparison to various statistical methods. This is because ANN models are non-parametric in nature, whereas, statistical methods in generally parametric models that require more statistical information. ANNs are supremely powerful in capturing unknown patterns present in the dataset.

Multilayer perceptrons are generally used to solve problems related to supervised



```
Trained on 18,585 samples (batch_size=128, epochs=3,000)
Final epoch (plot to see history):
     acc: 0.9155
    loss: 0.2213
 val_acc: 0.6662
val_loss: 1.247
```

Figure 15: Confusion Matrix for MLP

learning. MLs are part of a feedforward genre of ANNs, wherein MLP utilizes supervised methods of learning called backpropogation for the purpose of training. The 'n' number of layers present along with the non-linear function for activation set aside the MLP as compared to the linear perceptron.

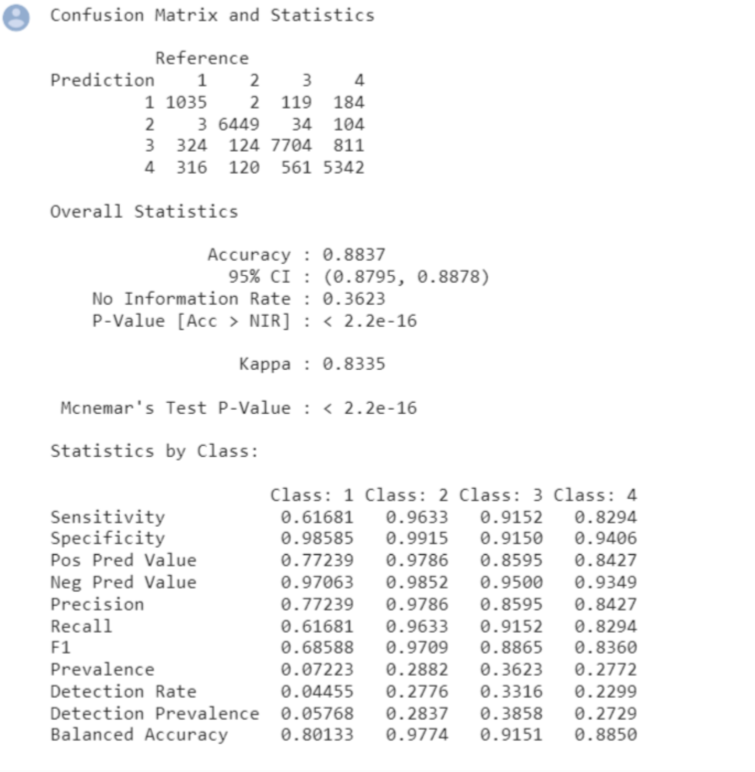
Multilayer Perceptron is an artificial neural network. The loss function was sparse categorical crossentropy. The model performed the best with 91.50 % accuracy for 3000 epochs. MLP could predict the correct label 91.50 % of the times. The batch size chosen in this case was 128.

7.3 Experiment 3: Applying Extreme Gradient Boosting

Algorithm	**nrounds**	**Accuracy %**	**no. of folds**
XGBoost	50	86.49	5
XGBoost	100	87.83	5
XGBoost	100	88.37	10

Table 2: XGBoost Accuracy after changing nrounds, number of folds for cross validation

Evaluating the XGBoost algorithm



```
Confusion Matrix and Statistics

          Reference
Prediction    1    2    3    4
         1 1035    2  119  184
         2    3 6449   34  104
         3  324  124 7704  811
         4  316  120  561 5342

Overall Statistics

               Accuracy : 0.8837
                 95% CI : (0.8795, 0.8878)
    No Information Rate : 0.3623
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.8335

 Mcnemar's Test P-Value : < 2.2e-16

Statistics by Class:

                     Class: 1 Class: 2 Class: 3 Class: 4
Sensitivity           0.61681   0.9633   0.9152   0.8294
Specificity           0.98585   0.9915   0.9150   0.9406
Pos Pred Value        0.77239   0.9786   0.8595   0.8427
Neg Pred Value        0.97063   0.9852   0.9500   0.9349
Precision             0.77239   0.9786   0.8595   0.8427
Recall                0.61681   0.9633   0.9152   0.8294
F1                    0.68588   0.9709   0.8865   0.8360
Prevalence            0.07223   0.2882   0.3623   0.2772
Detection Rate        0.04455   0.2776   0.3316   0.2299
Detection Prevalence  0.05768   0.2837   0.3858   0.2729
Balanced Accuracy     0.80133   0.9774   0.9151   0.8850
```

Figure 16: Accuracy of MLP

Extreme Gradient Boosting is an ensemble technique that follows boosting. XGBoost has a property of strengthening the weak learners by taking data from the preceding node. The model performed with an accuracy of 88.37 % at 10 folds cross-validation and nrounds set as 100. At lower value of nrounds, i.e 5, the model performed with 87.83 % for 100 folds and 86.49 % with 5 fold cross-validation at nrounds 50. The reason for the increased accuracy is due to stratified sampling, nrounds set as 10 and 10-fold cross validation and early stopping set at 20.

7.4 Experiment 4: Applying Extreme Gradient Boosting by Modifying Hyper-parameter Gamma

To check whether tuning for regularization will affect the accuracy. Generally, gamma is applied to models that are over-fitting. However, in our case the model works perfectly fine. The only reason to check the value of accuracy after the tuning of hyper-parameter 'gamma' is to check whether the accuracy would reduce by how much in case the gamma value is increased from 0 to 5. As per the expectation, gamma brings about a regularization in the process of boosting and at a gamma value of 5, the model performed at 85.32 % accuracy.

This clearly proves that hyper-parameter modification worked well with the XGBoost model and there was a significant change with the final model accuracy.

For cases wherein, this model might over-fit, the research would recommend increasing the value of the regularization parameter in smaller steps of 3 or 5.

7.5 Discussion

Human activity recognition has a strong prominence in the fields of medical sciences, health monitoring, physiotherapy and surveillance systems as well. The success of classification models like Random Forest, Extreme Gradient Boosting and Multilayer Perceptron certainly help in the overall research in this field from the perspective of machine learning.

This research proposes the identification of human activity with the help of a machine learning model based on trained label data from the extrasensory dataset. There were multiple research papers wherein, random forest was the recommended algorithm for multi-class classification of human activities. (Balli et al.; 2019), (Nweke et al.; 2018), (Mehrang et al.; 2018) to name a few have recommended the use of Random Forests as the best algorithm for Human Activity Recognition. Random forest gave extremely significant accuracy of about 89 % after 5-fold cross validation on the selected subset of data. However, in comparison to random forest, Multilayer Perceptron worked better.

(Gao et al.; 2019), (Xu et al.; 2019), and (Li et al.; 2019) found good results in terms of classification accuracy with the use of XGBoost. This was because XGBoost helps in getting good inputs from weak learners which we visualized in the variable importance plot we used as a feature selection using RF classifier.

The use of the three algorithms on the Extrasensory dataset was something that was a experiment in understanding the amazing dataset created by (Vaizman and Ellis; 2017) and team. However, in addition to these, the use of deep neural networks is an aspect yet to explored by a lot of researchers. (Cheng et al.; 2017) and team's research on HAR for HAR to aid the patients suffering from Parkinson's recommends the use of Deep Neural Networks and Gait recognition with 96.9 % accuracy with the help of gait analysis which is focused on identification of issues in the limb and other areas of the feet, muscle activity, and mechanics of the body.

8 Conclusion and Future Work

The objectives of the research question were substantially met with the proposed approach. The model was able to identify the activities at a given timestamp, by first learning and then predicting the activities on the test set, thus meeting the first two objectives and thereby a significant model for human activity recognition using machine learning algorithms. From the perspective of a recommended machine learning algorithm, XGBoost and Random Forest worked best with 88.21 % and 89.20 % accuracies respectively and the multilayer perceptron performed best with an accuracy of 91.50% at 3000 epochs.

To improve this research further, the use more deep neural networks with better optimization and a separate analysis on audio signals in the dataset are my recommendations. Use of other available sensors like ambient light sensors, temperature, watch compass sensors, humidity sensors, etc. from the Extrasensory dataset would also fetch some new insights. Also, larger amounts of data in terms of features, users and activities, could be analyzed along with time series analysis on the data.

9 Acknowledgement

I would to share my greatest gratitude to my guide Mr. Christian Horn for his continuous support and belief in me and my work right from the onset of the initial phase of the project. I greatly benefited from his feedback in all the phases of the research project and without his mentoring, this would be a difficult milestone to achieve. I would also like to thank all the professors who aided me in comprehending the concepts of data mining during the implementation of this project. Lastly, I would like to thank my parents, my family and my friends who've been a constant source of motivation and assistance, which has been of great help in my journey of the completion of this project.

References

Aliwy, A. H. and Ameer, E. H. A. (2017). Comparative Study of Five Text Classification Algorithms with their Improvements, **12**(14): 4309–4319.

Alvarez, F., Hernández, G. and Vretos, N. (2018). Behavior Analysis through Multimodal Sensing for Care of Parkinson ' s and Alzheimer ' s Patients, (March): 14–25.

Balli, S., Sağbaş, E. A. and Peker, M. (2019). Human activity recognition from smart watch sensor data using a hybrid of principal component analysis and random forest algorithm, *Measurement and Control (United Kingdom)* **52**(1-2): 37–45.

Can S, E. (2019). Continuous Stress Detection Using Wearable Sensors in Real Life: Algorithmic Programming Contest Case Study.

Cheng, W. Y., Scotland, A., Lipsmeier, F., Kilchenmann, T., Jin, L., Schjodt-Eriksen, J., Wolf, D., Zhang-Schaerer, Y. P., Garcia, I. F., Siebourg-Polster, J., Soto, J., Verselis, L., Martin-Facklam, M., Boess, F., Koller, M., Grundman, M., Monsch, A., Postuma, R., Ghosh, A., Kremer, T., Taylor, K., Czech, C., Gossens, C. and Lindemann, M. (2017). Human Activity Recognition from Sensor-Based Large-Scale Continuous Monitoring of Parkinson's Disease Patients, *Proceedings - 2017 IEEE 2nd International Conference on Connected Health: Applications, Systems and Engineering Technologies, CHASE 2017* pp. 249–250.

Chetty, G., White, M. and Akther, F. (2015). Smart Phone Based Data Mining For Human Activity Recognition, *Procedia - Procedia Computer Science* **46**(Icict 2014): 1181–1187.
URL: *http://dx.doi.org/10.1016/j.procs.2015.01.031*

Civitarese, G., Bettini, C., Sztyler, T. and Riboni, D. (2019). newNECTAR : Collaborative active learning for knowledge-based probabilistic activity recognition , *Pervasive and Mobile Computing* **56**: 88–105.
URL: *https://doi.org/10.1016/j.pmcj.2019.04.006*

de Chazal, O'Hare, E., Fox, N. and Heneghan, C. (2008). Assessment of sleep/wake patterns using a non-contact biomotion sensor, *2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 514–517.

Fayyad, U., Piatetsky-shapiro, G. and Smyth, P. (1996). From Data Mining to Knowledge Discovery in, **17**(3): 37–54.

Gaikwad, N. B., Tiwari, V., Keskar, A. and Member, S. (2019). Efficient FPGA Implementation of Multilayer Perceptron for Real-Time Human Activity Classification, *IEEE Access* **7**: 26696–26706.

Galvn-Tejada, C. E., Lpez-Monteagudo, F. E., Alonso-Gonzlez, O., Galvn-Tejada, J. I., Celaya-Padilla, J. M., Gamboa-Rosales, H., Magallanes-Quintanar, R. and Zanella-Calzada, L. A. (2018). A generalized model for indoor location estimation using environmental sound from human activity recognition, *ISPRS International Journal of Geo-Information* **7**(3).

Gani, O., Fayezeen, T., Povinelli, R. J., Smith, R. O., Arif, M., Kattan, A. J. and Iqbal, S. (2019). Journal of Network and Computer Applications A light weight smartphone based human activity recognition system with high accuracy, *Journal of Network and Computer Applications* **141**(May): 59–72.
URL: *https://doi.org/10.1016/j.jnca.2019.05.001*

Gao, X., Luo, H., Wang, Q., Zhao, F., Ye, L. and Zhang, Y. (2019). A human activity recognition algorithm based on stacking denoising autoencoder and lightGBM, *Sensors (Switzerland)* **19**(4).

Garcia-ceja, E., Riegler, M., Nordgreen, T., Jakobsen, P., Oedegaard, K. J. and Tørresen, J. (2018). Mental health monitoring with multimodal sensing and machine learning : A survey, *Pervasive and Mobile Computing* **51**: 1–26.
URL: *https://doi.org/10.1016/j.pmcj.2018.09.003*

Ghaderi, A. and Frounchi, J. (2015). Machine Learning-based Signal Processing Using Physiological Signals for Stress Detection, (November): 25–27.

Hu, C., Chen, Y., Peng, X., Yu, H., Gao, C. and Hu, L. (2019). A novel feature incremental learning method for sensor-based activity recognition, *IEEE Transactions on Knowledge and Data Engineering* **31**(6): 1038–1050.

Ignatov, A. D. and Strijov, V. V. (2015). Human activity recognition using quasiperiodic time series collected from a single tri-axial accelerometer.

Kune, R., Konugurthi, P., Agarwal, A., Rao, C. R. and Buyya, R. (2016). The anatomy of big data computing, *Softw., Pract. Exper.* **46**(1): 79–105.

Kursa, M. B. (2018). Boruta for those in a hurry, pp. 1–6.

Lee, C. C., Saidy, L. and Fitri (2019). Human activity recognition based on smart chair, *Sensors and Materials* **31**(5): 1589–1598.

Li, H., Pu, B., Kang, Y. and Lu, C. Y. (2019). Research on massive ECG data in XGBoost, *Journal of Intelligent and Fuzzy Systems* **36**(2): 1161–1169.

Malik, O. A. (2017). IMECE2012-87809, pp. 1–10.

Mao, J., Liu, Q., Song, X., Wang, H., Feng, H. and Xu, H. (2017). Combinatorial analysis of enzymatic bottlenecks of L -tyrosine pathway by p -coumaric acid production in Saccharomyces cerevisiae, *Biotechnology Letters* (94).

Mehrang, S., Pietilä, J. and Korhonen, I. (2018). An activity recognition framework deploying the random forest classifier and a single optical heart rate monitoring and triaxial accelerometer wrist-band, *Sensors (Switzerland)* **18**(2): 1–14.

Narang, A., Batra, B., Ahuja, A., Yadav, J. and Pachauri, N. (2018). Classification of EEG signals for epileptic seizures using Levenberg- Marquardt algorithm based Multilayer Perceptron Neural Network, **34**: 1669–1677.

Nweke, H. F., The, Y. W., Mujtaba, G. and Al-garadi, M. A. (2018). US CR, *Information Fusion* .
URL: *https://doi.org/10.1016/j.inffus.2018.06.002*

Ong, W., Jiunn, V., Sabri, N. and Ibrahim, Z. (2017). Image-based Human Fall Recognition Using Gaussian Mixture Model and Support Vector Machine, **10**(30): 339–344.

Ronao, C. A. and Cho, S.-b. (2016). Human activity recognition with smartphone sensors using deep learning neural networks, *Expert Systems With Applications* **59**: 235–244.
URL: *http://dx.doi.org/10.1016/j.eswa.2016.04.032*

Said, Y., Arnrich, B. and Ersoy, C. (2019). Stress detection in daily life scenarios using smart phones and wearable sensors : A survey, *Journal of Biomedical Informatics* **92**(February): 103139.
URL: *https://doi.org/10.1016/j.jbi.2019.103139*

Sakr, N. A., Abu-Elkheir, M., Atwan, A. and Soliman, H. H. (2018). Current trends in complex human activity recognition, *Journal of Theoretical and Applied Information Technology* **96**(14): 4564–4583.

Sarcevic, P., Pletl, S. and Kincses, Z. (2017). Comparison of Time- and Frequency-domain Features for Movement Classification using Data from Wrist-worn Sensors, pp. 261–266.

Sika, R. (2016). Methodologies of knowledge discovery from data and data mining methods in mechanical engineering, **7**(4): 97–108.

Subasi, A., Dammas, D. H., Alghamdi, R. D., Makawi, R. A., Albiety, E. A., Brahimi, T. and Sarirete, A. (2018). Sensor based human activity recognition using adaboost ensemble classifier, *Procedia Computer Science* **140**: 104–111.
URL: *https://doi.org/10.1016/j.procs.2018.10.298*

Sutton, R. S., Precup, D., Singh, S., Harb, J., Bacon, P.-L., Klissarov, M., Precup, D., Kulkarni, T. D., Bakker, B., Schmidhuber, J., Barto, A. G., Mahadevan, S., Salman, H., Singhal, P., Shankar, T., Yin, P., Salman, A., Paivine, W., Sartoretti, G., Travers, M. and Choset, H. (2003). Deep Reinforcement Learning with Temporal Abstraction and Intrinsic Motivation, *Conference on Intelligent Autonomous Systems* **48**(1-2): 181–211.
URL: *http://arxiv.org/abs/1803.01446%0Ahttp://pdf.aminer.org/000/337/561/automatic_discovery*

Vaizman, Y. and Ellis, K. (2017). Recognizing Detailed Human Context in the Wild from Smartphones and Smartwatches.

Vaizman, Y., Ellis, K., Lanckriet, G. and Weibel, N. (2018a). ExtraSensory App : Data Collection In-the-Wild with Rich User Interface to Self-Report Behavior.

Vaizman, Y., Ellis, K., Lanckriet, G. and Weibel, N. (2018b). Extrasensory app: Data collection in-the-wild with rich user interface to self-report behavior, *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, ACM, New York, NY, USA, pp. 554:1–554:12.
URL: *http://doi.acm.org/10.1145/3173574.3174128*

Wen, J. and Zhong, M. (2015). Expert Systems with Applications Activity discovering and modelling with labelled and unlabelled data in smart environments, *Expert Systems With Applications* **42**(14): 5800–5810.
URL: *http://dx.doi.org/10.1016/j.eswa.2015.04.005*

Witt, D. R., Kellogg, R. A., Snyder, M. P. and Dunn, J. (n.d.). ScienceDirect Windows into human health through wearables data analytics, *Current Opinion in Biomedical Engineering* **9**: 28–46.
URL: *https://doi.org/10.1016/j.cobme.2019.01.001*

Xu, S., Tang, Q., Jin, L. and Pan, Z. (2019). A Cascade Ensemble Learning Model for Human Activity Recognition with Smartphones, pp. 1–18.