2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)
Bari, Italy. October 6-9, 2019

Human Activity Recognition Based on Evolution of Features Selection and Random Forest*

Christine Dewi and Rung-Ching Chen

Abstract— **Human Activity Recognition is a promising area having potential to benefit the human society by developing assistive technologies in order to aid elderly, chronically ill and for people with special needs. Accurate activity recognition is challenging because human activity is complex and highly diverse. A comparative study on Human Activity Recognition (HAR) dataset based on four methods, Random Forest (RF), Support Vector Machines (SVM), K-Nearest Neighbors (KNN), and Linear Discriminant Analysis (LDA) with different features for selecting the best classifier among the models to test the dataset, has been carry out in this paper. The best classifier is choosing by accuracy of the model. We compare the result of dataset with and without important features selection by RF methods *varImp()*, Boruta, and Recursive Feature Elimination (RFE) to get the best accuracy. From the four methods, we found that the method RF have high accuracy from every group (98.16%, 98.09%, 93.6%), which is considered as a best classifier.**

Keywords— **HAR, SVM, KNN, Caret, LDA**

I. INTRODUCTION

Human activity recognition or HAR, is a challenging time series classification task. It involves predicting the movement of a person based on sensor data and traditionally involves deep domain expertise and methods from signal processing to correct engineer features from the raw data in order to fit a machine-learning model. Recent developments of sensor technologies and widespread use of small devices like smartphones, tablets etc., facilitate gathering of huge data from small portable devices of everyday use [1]. Recognition of human basic physical activities has recently been an interesting topic. With the advances of sensing techniques and data processing, current research can achieve accurate recognition of human's specific actions instead of simply tracking them [2]. The generate data from various high quality in-built sensors of small mobile devices can be collected over a length of time and transmitted by wireless technology to processing computers. Efficient and fast analysis of those time series data provides us with the opportunity of recognition, monitoring and control of various activities, from human to chemical and industrial process [1].

Activity recognition can use to perform continuous analysis of the daily activities performed by a user. Such an analysis is useful in understanding the behavior and thereby makes it possible to provide automated suggestions for reducing the risk factor for various non-communicable diseases [3]. In addition to healthcare applications, activity recognition is also useful in applications such as smart homes, security and transportation mode detection [4]. Differential recognition approaches have been proposing, including classifiers from machine learning, such as support vector machine [5], decision tree [6], and neural network [7]. Moreover, most of available methods recognize the activity set at one time with one feature set. As the growth of the activity number, challenges will be pose to both classifiers and feature selection.

The main contributions of this work can be summarized as follows: First, in this work, we will compare four classifiers method: Random Forest (RF), Support Vector Machines (SVM), K-Nearest Neighbors (KNN), and Linear Discriminant Analysis (LDA) with different features to select the best classifiers method based on accuracy of each classifier. Second, an analysis of variable importance to find out which variables are more relevant especially for classification data. The important features will select by RF methods *varImp(),* Boruta, and Recursive Feature Elimination (RFE). The classifiers for the dataset are built from the methods rf, lda, svmRadial and knn by using Caret package in R. HAR dataset stands for human activity recognition dataset, which is a collection 6 different activities, laying, walking, sitting, standing, walking upstairs, walking downstairs of a human being monitored by some social workers. This dataset come across different stages in preparation [8].

II. RELATED WORKS

In [8], the human activity recognition uses a multi-class Support Vector Machine (SVM) framework utilized to classify six full body activities and the reported accuracy is around 89%. In that work, 17 variables based feature vectors acquired by smartphones, which further used in the following classification process. Dataset consist of 561 features and 10299 instances. The accuracy calculation for this large dataset of 561 attributes is quite typical and the accuracy we get is not the correct one. In order to get proper accuracy pre-processing of the data is required. In [9][10], first propose to classify 6 actions with 3 different classifiers. The recognition accuracies in these papers range from 82% for both the custom decision tree classifier and artificial neural network classifier, to 86% for the automatically generated decision tree. Other research A. Sharma et al [11] and [7] used neural network for classification. This works achieved quite high recognition accuracy for basic activities while it was difficult to distinguish between similar activities like walking upstairs or walking downstairs.

*Research supported by Ministry of Science and Technology, Taiwan

Christine Dewi Author is PhD student at Department of Information Management, Chaoyang University of Technology, Taichung, Taiwan, e-mail: s10714904@cyut.edu.tw).

Rung-Ching Chen Author is with the Department of Information Management, Chaoyang University of Technology, Taichung, Taiwan, e-mail: crching@cyut.edu.tw).

A. Caret

The caret package (short for Classification and Regression Training) contains functions to streamline the model training process for complex regression and classification problems. The package utilizes a number of R packages but tries not to load them all at package start-up. The package suggests field includes 30 packages. Caret loads packages as needed and assumes that they installed. If a modelling package is missing, there is a prompt to install it. The package contains tools for data splitting, pre-processing, feature selection, model tuning using resampling, variable importance estimation, as well as other functionality [12].

B. Classifiers Method

Random Forests (RF) consists of a combination of decision-trees. It improves the classification performance of a singletree classifier by combining the bootstrap aggregating (bagging) method and randomization in the selection of partitioning data nodes in the construction of decision tree [13]. A RF classifier integrates a set of independent decision tree classifiers [14]. A decision tree with M leaves splits the feature space into M regions R_m *and* $1 \leq m \leq M$. For each tree, the prediction function $f(x)$ is defined as:

$$f(x) = \sum_{m=1}^{M} c_m \Pi(x, R_m) \quad (1)$$

Where M is the number of regions in the feature space, R_m is a region corresponding to m, c_m is a constants corresponding to m, and 1 is the indicator function:

$$\Pi(x, R_m) = \begin{cases} 1, & if\ x \in R_m \\ 0, & otherwise \end{cases} \quad (2)$$

The final classification decision is made from the majority vote of all trees.

K-Nearest Neighbors (KNN) is a supervised classification technique that can see as a direct classification method because it does not require a learning process. It just requires the storage of the whole data. To classify a new observation, the K-NN algorithm uses the principle of similarity between the training set and new observation to classify. Linear Discriminant Analysis (LDA) is the most commonly used as dimensionality reduction technique in the pre-processing step for pattern-classification and machine learning applications. The goal is to project a dataset onto a lower-dimensional space with good class-separability in order avoid overfitting (curse of dimensionality) and reduce computational costs. A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane. In other words, given labelled training data (*supervised learning*); the algorithm outputs an optimal hyperplane which categorizes new examples. In two-dimensional space this hyperplane is a line dividing a plane in two parts where in each class lay in either side.

III. Methodology

A. Evolution of Features Selection Method

Recursive feature elimination (RFE) offers a rigorous way to determine the important variables before you even feed them into a Machine Learning algorithm. Guyon et al. [15] proposed RFE which is applied in cancer classification by using SVM. RFE employs all features to build a SVM model, and then ranks the contribution of each feature in the SVM model which generates a ranked feature list, and finally eliminates irrelevant features meaning less contribution to the SVM model. Moreover, Recursive Feature Elimination (RFE) is a powerful algorithm for feature selection, which depends on the specific learning model. A brief explanation of the RFE algorithm is based on this following step:

1. Fit the model using all independent variables. Calculate variable importance of all the variables.
2. Each independent variable is ranked using its importance to the model.
3. Drop the weakest variable (worst ranked) and builds a model using the remaining variables and calculate model accuracy.
4. Repeat step 4 until all variables are used.
5. Variables are then ranked according to when they were dropped. For classification use accuracy and kappa as a metrics.

Boruta is a feature ranking and selection algorithm based on RF algorithm. The advantage with Boruta is that it clearly decides if a variable is important or not and helps to select variables that are statistically significant. Besides, you can adjust the strictness of the algorithm by adjusting the p values that defaults to 0.01 and the *maxRun*(). *maxRun()* is the number of times the algorithm is run. The higher the *maxRun(),* the more selective you get in picking the variables. The default value is 100. Another way to look at feature selection is to consider variables most used by various machine learning algorithms the most to be important. Depending on how the machine learning algorithm learns the relationship between *X*'s and *Y*, different machine learning algorithms may possibly end up using different variables to various degrees. We use *train()* the desired model using the caret package. Then, use *varImp()* to determine the feature importance by RF.

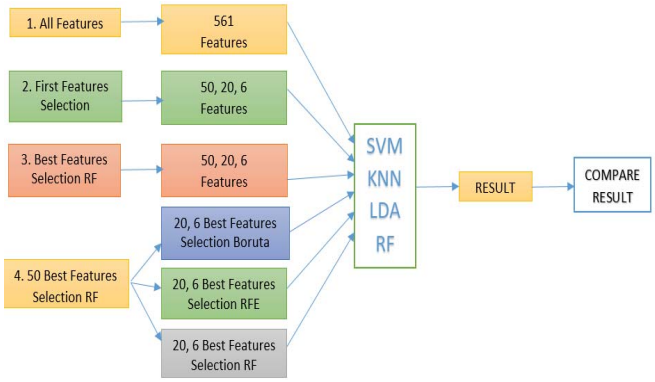

Fig. 1. The workflow of this research.

Fig. 1 describes the workflow of this research. The experiment divided into some groups. Random Forest (RF), Support Vector Machines (SVM), K-Nearest Neighbors (KNN), and Linear Discriminant Analysis (LDA) used for activity classification from 561, 50, 20, and 6 features. RF feature selection method *varImp()* from Caret package will be used to choose the important features data set. After that we use Boruta and RFE to select important features and compare with the previous result. The next step is compared the accuracy result of data set using the same classifiers with features selection method (RF, RFE, Boruta) and without features selection method.

B. Data Set

The simulation use HAR data set publicly available from UCI (University of California Irvine) machine learning repository [16]. The data set get from a group of 30 volunteers within an age bracket of 19-48 years. Each person performed six activities (WALKING, WALKING_UPSTAIRS, WALKING_DOWNSTAIRS, SITTING, STANDING, LAYING) wearing a smartphone (Samsung Galaxy S II) on the waist. Using its embedded accelerometer and gyroscope, we captured 3-axial linear acceleration and 3-axial angular velocity at a constant rate of 50Hz. The experiments have been video-recorded to label the data manually. The obtained dataset has randomly partitioned into two sets, where 70% of the volunteers selected for generating the training data and 30% the test data.

The sensor signals (accelerometer and gyroscope) were pre-processed by applying noise filters and then sampled in fixed-width sliding windows of 2.56 sec and 50% overlap (128 readings/windows). The sensor acceleration signal, which has gravitational and body motion components separated using a Butterworth low-pass filter into body acceleration and gravity. The gravitational force is assuming to have only low frequency component, therefore a filter with 0.3 Hz cut off frequency used. From each window, a vector of 561 features obtained by calculating variables from the time and frequency domain. The details can be find in [8].

The features selected for this database come from the accelerometer and gyroscope 3-axial raw signals *tAcc-XYZ* and *tGyro-XYZ*. These time domain signals (prefix 't' to denote time) were captured at a constant rate of 50 Hz. Then they filtered using a median filter and a third order low pass Butterworth filter with a corner frequency of 20 Hz to remove noise. Similarly, the acceleration signal was then separated into body and gravity acceleration signals (*tBodyAcc-XYZ* and *tGravityAcc-XYZ*) using another low pass Butterworth filter with a corner frequency of 0.3 Hz. These signals used to estimate variables of the feature vector for each pattern: *'-XYZ'* is used to denote 3-axial signals in the *X, Y* and *Z* directions that can be seen on Table I.

TABLE I. 3-AXIAL SIGNAL IN THE X, Y AND Z

No	Features
1	*tBodyAcc-XYZ*
2	*tGravityAcc-XYZ*
3	*tBodyAccJerk-XYZ*
4	*tBodyGyro-XYZ*
5	*tBodyGyroJerk-XYZ*
6	*tBodyAccMag*
7	*tGravityAccMag*
8	*tBodyAcc-XYZ*
9	*tGravityAcc-XYZ*
10	*tBodyAccJerk-XYZ*
11	*tBodyGyro-XYZ*
12	*tBodyGyroJerk-XYZ*
13	*tBodyAccMag*
14	*tGravityAccMag*
15	*tBodyAccJerkMag*
16	*tBodyGyroMag*
17	*tBodyGyroJerkMag*

The set of variables that estimated from these signals can be seen on Table II.

TABLE II. VARIABLES FROM THE SIGNAL

No	Variables
1	Mean value
2	Standard deviation
3	Median absolute deviation
4	Largest value in array
5	Smallest value in array
6	Signal magnitude area
7	Energy measure
8	Interquartile range
9	Signal entropy
10	Auto regression coefficients
11	Correlation coefficient
12	Index of the frequency component with largest magnitude
13	Weighted average
14	Skewness of the frequency domain signal
15	Kurtosis of the frequency domain signal
16	Energy of a frequency interval within the 64 bins of the FFT of each window.
17	Angle between vector

Additional vectors obtained by averaging the signals in a signal window sample. These are used on the *angle()* variable can be seen on Table III.

TABLE III. *ANGLE()* VARIABLES

No	*Angle()* variable
1	*gravityMean*
2	*tBodyAccMean*
3	*tBodyAccJerkMean*
4	*tBodyGyroMean*
5	*tBodyGyroJerkMean*

IV. EXPERIMENT RESULTS AND DISCUSSION

A. Environment and Dataset

The data consist of 561 features and 10299 instances from UCI HAR dataset. Furthermore, we use caret package from R studio language to conduct the experiment. Evaluation for this research is based on calculation of accuracy. Accuracy is how often the model trained is correct, which depicted by using Confusion Matrix. A confusion matrix is the summary of prediction results on a classification problem [8]. The number of correct and incorrect predictions summarized with count values and separated by each class, which is the key to the confusion matrix. The higher the accuracy value, the better the method. Accuracy calculation in this paper based on formula below:

$$Accuracy = (TP + TN)/(TP + TN + FP + FN) \quad (3)$$

Where: TP = True positive; FP = False positive; TN = True negative; FN = False negative.

B. Experiment Result

Table IV describe the result of classification accuracy of different classifiers with all 561 features. As we can see method RF have high accuracy 98.57% compare to others.

TABLE IV. CLASSIFICATION ACCURACY OF DIFFERENT CLASSIFIERS WITH ALL FEATURES

Total Features	RF	KNN	SVM	LDA
561	**0.9857**	0.9748	0.9796	0.9823

Table V and Fig 2 show the result of classification accuracy of different classifiers without best features selection. In this works, we choose the first 50, 20, and 6 features from the total features in the dataset. Moreover, RF method have the best accuracy 96.53% for 50 features, 79.97% for 20 features, and 72.28% for 6 features. The trend of accuracy for all classifiers method decrease when we limit the features that is why we need to select the important features to get the best accuracy. Furthermore, to process all data it will takes a long time; the best features selection is needed. The processing time will faster than before, if we limit the features and data.

TABLE V. CLASSIFICATION ACCURACY OF DIFFERENT CLASSIFIERS WITHOUT FEATURES SELECTION

Total Features	RF	KNN	SVM	LDA
50	**0.9653**	0.9332	0.9394	0.8931
20	**0.7997**	0.7527	0.5334	0.6819
6	**0.7228**	0.6996	0.6294	0.5456

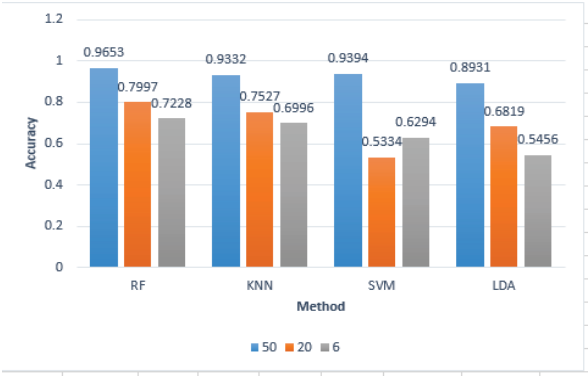

Fig. 2. Classification accuracy of different classifiers without best features selection

In this experiment, we use method *varImp()* from RF to select the important features from 561 features become 50, 20, and 6 features. Best features selection by method *varImp()* RF, RFE, and Boruta can be seen on Fig 3, 4 and 5. In addition, Table VI explain about 6 best features from RF, RFE, and Boruta method. Moreover, from 561 features we can choose 20 and 6 important features.

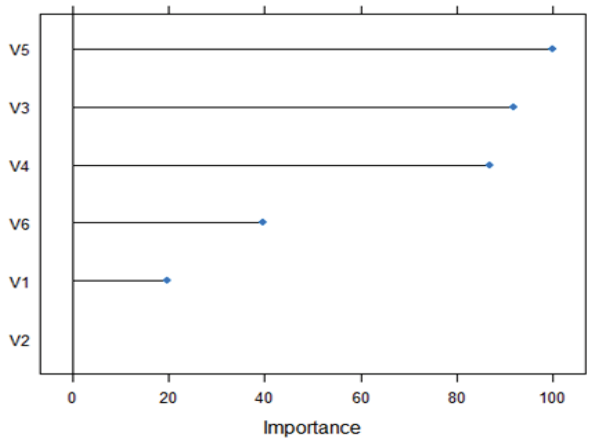

Fig. 3. Features Selection by Random Forest

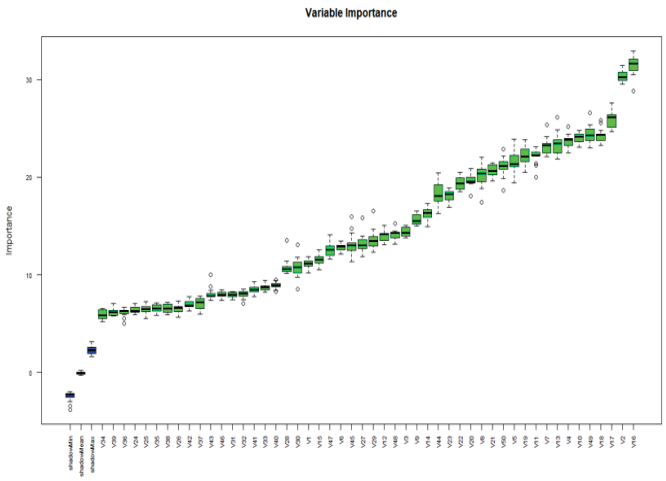

Fig. 4. Features Selection by Boruta

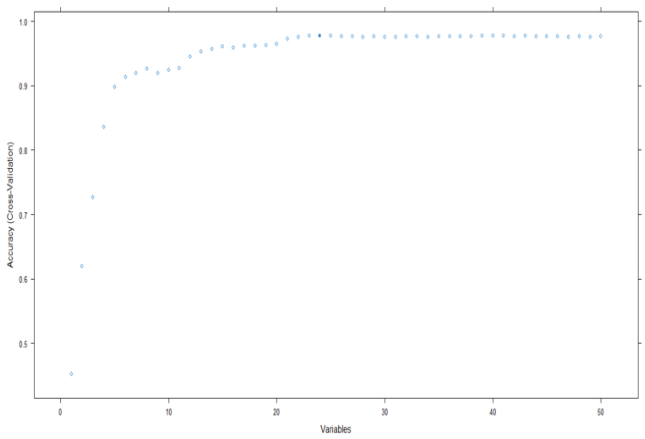

Fig. 5. Features Selection by Recursive Features Elimination.

TABLE VI. FEATURES SELECTION

ID	6 Features (RF)	6 Features (Boruta)	6 Features (RFE)
V1	*tGravityAcc-mean()-X*	*tGravityAcc-arCoeff()-X*	*tBodyAcc-correlation()-X*
V2	*tGravityAcc-mean()-Y*	*tBodyAcc-correlation()-X*	*tGravityAcc-min()-X*
V3	*tGravityAcc-min()-X*	*tGravityAcc-arCoeff()-Y*	*tGravityAcc-arCoeff()-X*
V4	*tBodyAccJerk-energy()-X*	*tGravityAcc-min()-Y*	*tGravityAcc-mean()-X*
V5	*fBodyAcc-mad()-X*	*tGravityAcc-arCoeff()-Y*	*angle(X gravityMean)*
V6	*fBodyAccJerk-bandsEnergy()-1*	*angle(Y gravityMean)*	*angle(Y gravityMean)*

Table VII and Fig 6 show the result of classification accuracy of different classifiers with best features selection. The graph shows RF method have the best accuracy 98.16% for 50 features, 98.09% for 20 features, and 93.6% for 6 features. Although the trend of accuracy decrease, but RF method still receive the high accuracy, compare with other classifiers. It is proof that even we limit the features we can get the best accuracy if we choose the important features. As we can see on Table 4 RF method receive 93.6% for 6 features and it is higher than on Table V RF method receive 72.28% for 6 features.

TABLE VII. CLASSIFICATION ACCURACY OF DIFFERENT CLASSIFIERS WITH FEATURES SELECTION BY RANDOM FOREST

Total Features	RF	KNN	SVM	LDA
50	**0.9816**	0.9728	0.9503	0.9169
20	**0.9809**	0.9659	0.9564	0.8971
6	**0.936**	0.9217	0.8808	0.846

TABLE VIII. CLASSIFICATION ACCURACY OF DIFFERENT CLASSIFIERS WITH FEATURES SELECTION (20 FEATURES)

Best Features Method	Total Features	RF	KNN	SVM	LDA
Boruta	20	0.9741	0.9428	0.9475	0.8965
RFE	20	0.9653	0.9094	0.9094	0.8365
RF	20	**0.9809**	0.9659	0.9564	0.8971

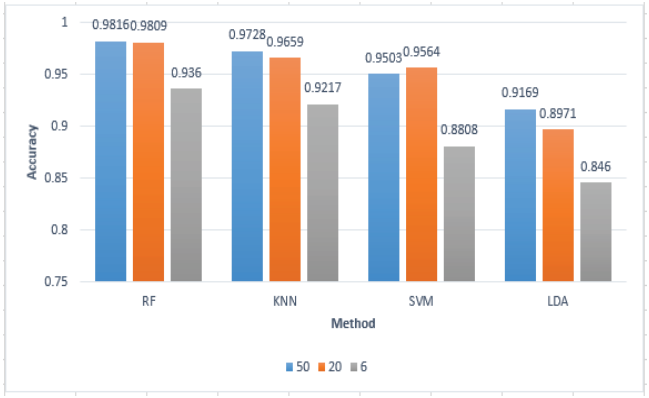

Fig. 6. Classification accuracy of different classifiers with best features selection by Random Forest.

Fig 7 shows the classification accuracy of different classifiers with 6 best features selection. In addition, RF method has high accuracy above 93.6% and LDA become the worst method only gets accuracy between 80% and 85%.

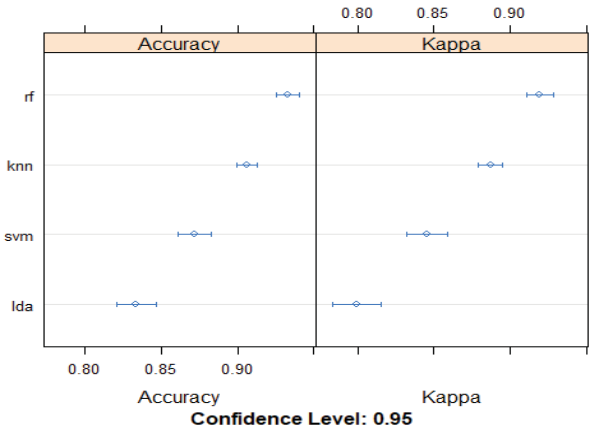

Fig. 7. Classification accuracy of different classifiers with best features selection RF method (6 features).

TABLE IX. CLASSIFICATION ACCURACY OF DIFFERENT CLASSIFIERS WITH FEATURES SELECTION (6 FEATURES)

Best Features Method	Total Features	RF	KNN	SVM	LDA
Boruta	6	0.8924	0.864	0.8692	0.7786
RFE	6	**0.9394**	0.8385	0.8331	0.705
RF	6	0.936	0.9271	0.8808	0.846

Table VIII and IX describe the result of classification accuracy of different classifiers with features selection method Boruta, RFE, and RF (20 and 6 features). The result from Table IV, V, VII, VIII and IX shows that RF method has high accuracy in all experiment group. In Table VII and VIII we can see RF method has high accuracy about 98.09% with 20 features. Furthermore, in Table IX RFE become the best features selection method with 6 features and get 93.94% for accuracy using RF method.

V. Conclusion and Future Work

In this work, we compare four classifiers method, there are Random Forest (RF), Support Vector Machines (SVM), K-Nearest Neighbors (KNN), and Linear Discriminant Analysis (LDA) with different features (50, 20, 6) to select the best classifiers method based on accuracy of each classifier. RF methods *varImp()* very helpful and efficient to select the important features. The accuracy of the dataset is depending on how much quality is the data and by the method which is trained. HAR dataset collect 6 different activities, laying, walking, sitting, standing, walking upstairs, walking downstairs of a human being monitored by some social workers. Table IV, V, VII, VIII and IX shows that RF method has high accuracy in all experiment group. We can conclude from this experiment that RF is the best classifiers method. In addition, best features selection by RF method can help to choose the important features, so we should not use all the features in the dataset. Furthermore, it will affect the processing time and it could give the best accuracy and more feature is the higher dimension of data. For example, on Table VII RF method receive 93.6% for 6 features with features selection and it is higher than on Table V RF method receive 72.28% for 6 features without features selection and compare to Table IV RF method receive 98.57% with 561 features. We can combine features selection method to get the best features from dataset. RFE best features method can improve the accuracy RF method from 93.6% to 93.94% as we can see in the Table IX. In the future, we would like to set up our own dataset with more different activities besides the six activities tested in this work and use different method.

Acknowledgment

This paper is supported by Ministry of Science and Technology, Taiwan. The Nos are MOST-107-2221-E-324 -018 -MY2 and MOST-106-2218-E-324 -002, Taiwan. This research is also partially sponsored by Chaoyang University of technology (CYUT) and Higher Education Sprout Project, Ministry of Education, Taiwan, under the project name: "The R&D and the cultivation of talent for Health-Enhancement Products".

References

[1] K. Nakano and B. Chakraborty, "Effect of dynamic feature for human activity recognition using smartphone sensors," in *Proc. of 2017 IEEE 8th International Conference on Awareness Science and Technology (iCAST 2017)*, IEEE, Taichung, Taiwan, pp. 539-543, 2017.

[2] Y. Guo, D. Tao, W. Liu, and J. Cheng, "Multiview cauchy estimator feature embedding for depth and inertial sensor-based human action recognition," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 47, no. 4, pp. 617–627, 2017.

[3] A. Jain and V. Kanhangad, "Human activity classification in smartphones using accelerometer and gyroscope sensors," *IEEE Sensors Journal*, vol. 18, no. 3, pp. 1169-1177, Feb. 2018.

[4] J. Wannenburg and R. Malekian, "Physical activity recognition from smartphone accelerometer data for user context awareness sensing," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 47, no. 12, pp. 3142–3149, Dec. 2017.

[5] D. Anguita, A. Ghio, L. Oneto, X. Parra, and J. Reyes-Ortiz, "Human activity recognition on smartphones using a multiclass hardware friendly support vector machine," in *Proc. of International Conference on Ambient Assisted Living and Home Care*, pp.216-223, 2012.

[6] T. Phan, "Improving activity recognition via automatic decision tree pruning," in *Proc. of ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pp. 827-832, 2014.

[7] J. Yang, J. Wang, and Y. Chen, "Using acceleration measurements for activity recognition: An effective learning algorithm for constructing neural classifiers," *Pattern Recognition Letters*, vol. 29, no. 16, pp. 2213-2220, 2008.

[8] D. Anguita, A. Ghio, L. Oneto, X. Parra, and J.L. Reyes-Ortiz, "A public domain dataset for human activity recognition using smartphones," *in Proc. of the European symposium on artificial neural networks (ESANN)*, Bruges, pp. 437–442, 2013.

[9] M. Ermes, J. arkka, and L. Cluitmans, "Advancing from online to online activity recognition with wearable sensors," *in Proc. of 30th Annual International IEEE EMBS Conference*, pp. 4451–4454, 2008.

[10] J. Arkka, M. Ermes, P. Korpip, J. Mantyjarvi, J. Peltola, and I. Korhonen, "Activity classification using realistic data from wearable sensors," *IEEE Trans. Inf. Technol. Biomed*, pp. 119–128, 2006.

[11] A. Sharma, Y. Lee, and W. Chung, "High accuracy human activity monitoring using neural network," *in Proc. of the third international conference on convergence and hybrid information technology*, pp. 430–435. 2008.

[12] K. BhanuJyothi, K.H. Bindu, and D. Suryanarayana, "A comparative study of random forest & k – nearest neighbors on har dataset using caret", *International Journal of Innovative Research in Technology*, vol. 3, issue 9, pp 6-9, 2017.

[13] P. Mandha, G. Lavanya Devy, and S. Vizianand Row, "A random forest based classification model for human activity recognition," in *Proc. of International Conference on Innovative Applications in Engineering and Information Technology (ICIAEIT-2017)*, pp. 294-300, 2017.

[14] L. Breiman, "Random forests," *Journal of Machine Learning,*, vol. 45, pp. 5–32, 2001.

[15] L. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene Selection for Cancer Classification using Support Vector Machines," Machine Learning, vol. 46, pp. 389-422, 2002.

[16] M. Lichman. (2013) UCI machine learning repository. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/, Irvine, CA: University of California, School of Information and Computer Science.