Top-K Human Activity Recognition Dataset

https://doi.org/10.3991/ijim.v14i18.16965

Moses Lesiba Gadebe (✉)
Tshwane University of Technology, Gauteng, South Africa
gadebeml@tut.ac.za

Okuthe Paul Kogeda
University of Free State, Bloemfontein, South Africa

Abstract—The availability of Smartphones has increased the possibility of self-monitoring to increase physical activity and behavior change to prevent obesity. Self-monitoring on Smartphone comes with some challenges such as unavailability of lightweight classification algorithm, personalized dataset to capture bodily postures, subject sensitivity, limited storage and computational power. However, most classification algorithms such as Support Vector Machines, C4.5, Naïve Bayes and K Neighbor relies on largest dataset to accurately predict human activities. In this paper, we present top-k of compressed personalized dataset collected from 13 participants to reduce computational cost with increased accuracy. We benchmarked our dataset and found that is suitable for tree-oriented algorithm, such as Random Forest, C4.5 and Boosted tree with accuracy and precision of 100% except for KNN, Support Vector and Naïve Bayes.

Keywords—Top-k personalized dataset, gravimeter filtering technique, tree oriented

1 Introduction

It is important in Human Activity Recognition (HAR) to select suitable classification algorithm for collected dataset to improve classification accuracy and precision, versus computational cost. Normally, HAR is divided into three steps: firstly sensory data collection from accelerometer and gyroscope sensors, and labeling of collected sensor data as human activity; secondly feature extraction to remove noise and missing values from dataset, also known as pre-processing; and third, classification based on machine learning algorithms such as K Nearest Neighbor (KNN), Support Vector Machine (SVM), C4.5, Naïve Bayes, Random Forest, Boosted Tree algorithms [1][2]. Most algorithm (KNN, SVM, C4.5 and Naïve Bayes) relies on larger dataset with implication on resource constraint environment such as Smartphone. Availability of sensory dataset for benchmarking real time HAR systems activity is still lagging behind due to the shortage of personalized dataset [3][4][5].

Recently, some HAR researchers collected and published their datasets online to allow other researchers to benchmark their HAR systems [6][7][8]. However, some of the datasets lack personal attributes (e.g., age, height, weight, body mass index) and are

in high dimensional space, which requires high computational power and large memory [4]. The latter poses a challenge in providing HAR systems on resource-constrained platforms such as Smartphone. Recent HAR models are mindful of resource constrained Smartphone *vis-à-vis* classification requirements [2]. However, most of the dataset lacks a common feature to completely capture bodily posture to accurately classify static and dynamic human activities, hence subject-sensitivity challenge still persist. In this paper, we present a compressed top-k personalized dataset based on harmonic motion principle to mimic walking transition to augment Signal Magnitude Vector and Tilt Angle proposed in [9] and [10]. In this paper our contribution is three-folds, firstly a novel gravimeter filtering technique to select usable best classification feature. Our unique usage of compressed top-k dataset lead to optimal usage of Smartphone limited resources. Thirdly, the inclusion of harmonic motion to augment signal magnitude vector (SMV) opens-up the possibility of using single HAR training dataset for multiple-subjects. The rest of this paper is structured as follows: In Section II, we present related work. In Section III, conceptual model to collect top-k personalized dataset is presented. The selection and benchmarking are presented in section IV. Lastly the conclusion and future work is presented in Section V.

2 Related Work

HAR has been studied for decades, but there is still limited number of publicly available HAR dataset for machine learning. Researchers in [6] investigated and established that there is a lack of baseline HAR dataset. To close the gap, they collected and published dataset called Physical Activity Monitoring for Aging People dataset (PAMAP2), obtained from 9 subjects. They used 3 Calibri wireless Inertial Measurement Unit (IMU) devices attached to the dominant arm wrist, ankle and one on the chest of each participant at a frequency rate of 100 hertz. However, researchers in [7] collected and published their dataset called University of Southern California Human Activity dataset (USC-HAD) from 14 subjects wearing Motion Node device on their waist, which is connected to the laptop at a frequency rate similar to that in [7]. On the other hand, the authors in [8] and [11] found that public dataset is incrementally introduced, however, there is still a lack of personalized Smartphone dataset. To address this problem, the researchers in [8] and [11] published datasets collected using Samsung Galaxy S II Smartphone from 30 subjects at a frequency rate of 50 hertz, respectively. However, their dataset suffers from dimensionality limitation and thus inappropriate for resource constrained Smartphone platform [4]. A Similar personalized dataset was proposed and published on UCI of machine learning by [12].

The dataset is named dataset-har-PUC-Rio-Ugulino (PUCRO) and collected dataset from 4 healthy subjects consisting of 5 classes (sitting-down, standing-up, standing, walking, and sitting). Four sensors were used to collect 165634 rows of data consisting 6 personal attributes (name, gender, age, bmi, height, weight) and 13 group of tri-axial values (X, Y, Z). A more representative personalized dataset called Wireless Sensor Data Mining (WISDM) was proposed by [13]. The WISDM dataset was donated and published by [14]. They used Smartphone and Smartwatch two sensors (accelerometer

and gyroscope) to collect the dataset from 51 recruited subjects performing 18 different human activities using Samsung Galaxy S5, Google Nexus 5 and LG G Watch. Each of 51 collected files consist of 6 attributes (subject code, activity label, time-stamp, X, Y and Z tri-axial values) and partitioned as (Non-hand-oriented activities: {walking, jogging, stairs, standing, kicking}, Hand-oriented activities (General): {dribbling, playing catch, typing, writing, clapping, brushing teeth, folding clothes} and Hand-oriented activities (eating): {eating pasta, eating soup, eating sandwich, eating chips, drinking}). However, it lacks features capable to capture bodily postures to mimic walking patterns completely [10]. Similar to our study in [9] is the work of [15], they found that learning new activities to adapt to new users' needs is challenging due to shortage of annotated dataset. They proposed Feature-Based and Attribute-Based learning to leverage the relationship between existing and new activities to compensate for shortage of dataset. Radial Basis Function was used to feed the SVM algorithm with 11 attributes for classification [15]. They evaluated their technique and found it outperforms other traditional HAR models in recognizing new activities using limited training dataset. However, the technique does not address shortage of personalized dataset; it only detects new activities from existing dataset. More recently is the technique proposed in [16] and [17], it allows each subject to enter height, weight and BMI and employs machine learning algorithm to filter social media using height, weight and BMI to recommend physical activity plans. We present personalized top-k dataset incline on harmonic motion to capture bodily posture to address the shortage of personalized.

3 A Conceptual Model to Collect Top-K Personalized Dataset

We present our personalized model given in Fig 1 to collect top-k personalized dataset, convenience sampling was followed to recruit 13 staff and students to participate in dataset collection process.

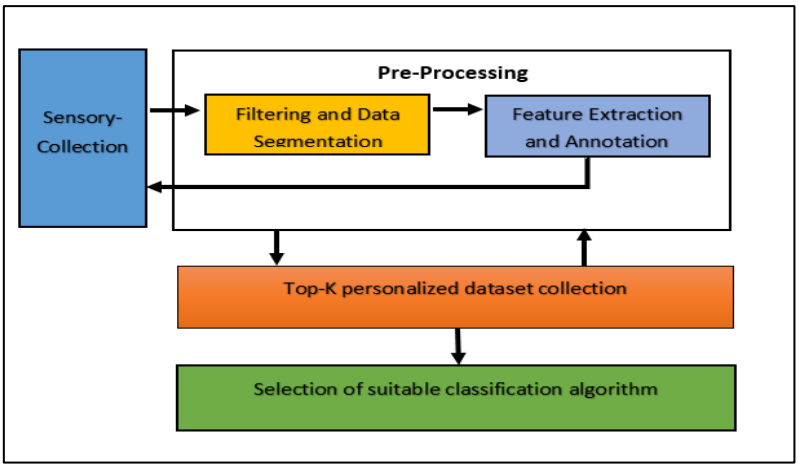

Fig. 1. Personalized model to collect personalized Top-K dataset

The model consists of (a) sensory data collection, (b) pre-processing (Filtering and data segmentation, and Feature Extraction and Annotation), (c) Top-K personalized data collection, and (d) Selection of suitable classification algorithm.

3.1 Sensory data collection process

In this process, we present sensory data collection using cost effective Smartphone accelerometer available through the day at closer proximity of subjects [10]. All subjects are expected to carry Smartphone inside front pocket similar to [13][14] as portrayed in Fig 2 and Fig 3 whilst real-time accelerometer tri-axial (X, Y, Z) values are being generated, computed and automatically annotated [10].

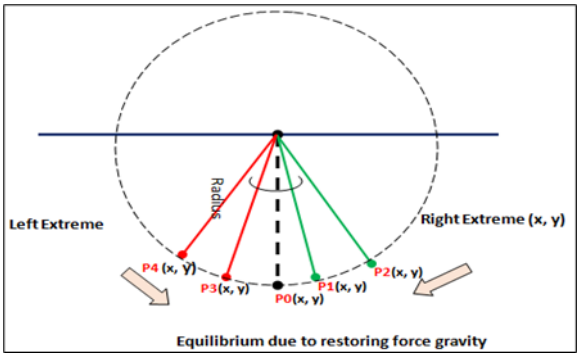

Fig. 2. Simple pendulum movements

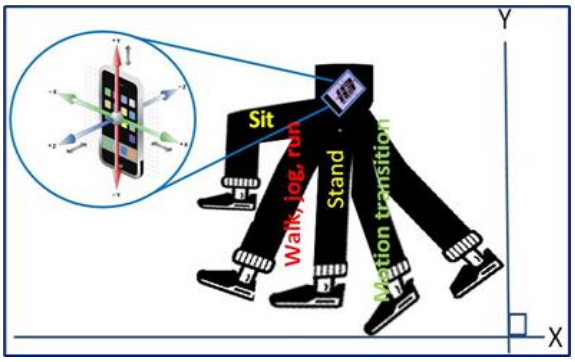

Fig. 3. Harmonic motion of legs transition position

Using front pocket, we can properly capture human legs postures during standing, swinging back and forth of human legs just like a simple harmonic motion [10] as shown in Fig 2 and Fig 3.We describe harmonic motion as a uniform projection of circular motion along a diameter of circle as a center of mass, given as pivot point of a

hanging blob relative to restoring gravitational force. Hence, we fix Smartphone orientation based on tilting angle (TA) by treating Smartphone as mass hanging in weightless front pocket.

3.2 Pre-processing

In this process, we solve Smartphone orientation (Rest or Portrait or Landscape) to extract signal values using signal magnitude vector (SMV) and TA, because SMV is inadequate to capture different bodily postures [9][10]. Based on size of arc due to restoring gravity force at the time we determine Smartphone orientation. The gravity force causes blob shifts of (*x*, *y*) from equilibrium (P0) to and from extremes (P1:P2; P3:P4) positions due to legs transitions depicted in Fig. 3 and Fig. 4. We compute radius as magnitude using SMV equation (1) based on accelerometer tri-axial as point *R* (*x*, *y*, *z*).

$$\text{SMV} = \sqrt{\mathrm{r_x}^2 + \mathrm{r_y}^2 + \mathrm{r_z}^2} \tag{1}$$

The accelerometer measures acceleration due to gravity about 9.8 m/s^2 when phone is in portrait position around y-axis [19]. Thus, the angle of interest (α, β and γ) is determined as largest theta to find phone orientation associated with SMV and gravity similar to Euler triple angles shown in Fig.4[19][20].

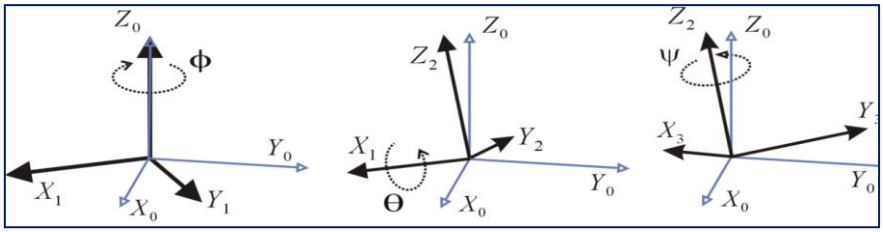

Fig. 4. Euler Angle Axes, rotation by Yaw (γ), Pitch (β) and Roll (α)

TA is defined as angle between positive z-axis and gravitational vector g if phone is in Portrait. Thus, the largest theta is given as size of arc in radians less than $\frac{\pi}{2}$ =1.570 [19][20] defined by:

$$Tilt\ Angle\ around\ \ \alpha \text{ or } \beta \text{ or } \gamma = \left| tan^{-1}\left(\frac{axis}{\sqrt{radius}}\right)\right| \times \frac{180}{\pi} \tag{2}$$

Where axis is either X, Y and Z depending on orientation of tri-axial values around α, β and γ angles. That is, if the absolute value around alpha (α) approximates 1.570, the orientation is Landscape, else if beta (β) approximates to 1.570, the orientation is Portrait otherwise is in Rest around Gamma(γ).

a) **Filtering and data segmentation**: Raw tri-axial features from accelerometer are not without restoring gravity force [3][20]. Based on restoring force relative to maximum TA around (α, β and γ) [19], we propose unique gravimeter filtering technique associated with harmonic motion period defined by:

$$period \ = 2\pi\sqrt{\frac{L}{gravity}} \ \ where\ \theta \ll 1 \tag{3}$$

Where L, is a magnitude expressed as $L = \frac{gravity}{\pi^2} \times \frac{period^2}{4}$ in equation (3) in the presence of gravity approximation of 9.8 m/s^2 as $\frac{g}{\pi^2} \approx 1$ (0.994), thus we rearranged equation (3) to equation (4):

$$graviMeter \ = 4\pi^2 \times (\frac{L}{period^2}) \tag{4}$$

Therefore, if gravimeter approximates to 9.8 m/s^2 as $\frac{g}{\pi^2} \approx$ 1g (0.994), *x* and *y* extremes positions of harmonic motion are sinusoidal in time *t* (period) computed by equation (5) and (6) [10].

$$x = A\,sin(\omega t + \varphi) \tag{5}$$

$$y = A\ \ cos(\omega t + \varphi) \tag{6}$$

Where (ωt + φ) is maximum TA as theta, A is magnitude (L) given by SMV, ω is omega the angular velocity rotation from equilibrium (P0) to/from positions (P1:P2) and (P3:P4) (see Fig. 3) defined by:

$$\omega = (\frac{2\,Theta}{period}) \tag{7}$$

Therefore, we collect confident harmonic motion (period, omega, x and y) as confident features into matrix $V^{N\times4}$, where all null values are discarded using equation (4). All equations from 2 to 7 are combined into our unique gravity filtering technique equation defined by:

$$V^{N\times4} = \begin{bmatrix} \text{period}_{i,1} & \text{omega}_{i,2} & \text{x}_{i,3} & \text{y}_{i,4} \\ \text{period}_{i,1} & \text{omega}_{i,2} & \text{x}_{i,3} & \text{y}_{i,4} \\ \text{period}_{N,1} & \text{omega}_{N,2} & \text{x}_{N,3} & \text{y}_{N,4} \end{bmatrix} \tag{8}$$

Expanded to

$$V^{N\times4} = \sum_{i,j}^{N} f(period_{i,j}) = \left\{ \begin{matrix} \text{period}_{i,j} \\ \text{omega}_{i,j} \\ \text{x}_{i,j} \\ \text{y}_{i,j} \end{matrix} \textbf{ if } \text{period}_{i,j} \text{approximate (0.994)} \right\} \tag{9}$$

Where $f(period_{i,j})$ is gravimeter filtering approximation function based on gravity relative to maximum TA at specific period, *N* is window segment to store group of confident features ($\text{period}_{i,j}$, $\text{omega}_{i,j}$, $\text{x}_{i,j}$ and $\text{y}_{i,j}$) in row *i* until maximum *N* Hertz

(Hz) is reached. We combine equation (8) and equation (9) into Gravimeter Filtering Technique Algorithm 1, which require accelerometer sensor (accel.x, accel.y, accel.z) values as input in line 1. During the sensory data collection, the while loop is tested whether row count reached the window size or not in line 5 of Algorithm 1. Therefore, if the window size is not reached, the SMV, TA around (α, β, γ), the gravity pull and theta are calculated based on maximum TA in line 11 to 21. The harmonic motion attributes (period, omega, X and Y) are computed from tri-axial accelerometer from line 6 to 10 to capture human legs postural features. The harmonic motion attributes are stored in matrix $V^{50\times4}$ only if their gravimeter approximate to $\frac{\text{gravity}}{\pi^2}$ otherwise are discarded in line 29 to 34 of Algorithm 1.

Algorithm 1: Gravimeter Filtering Technique Algorithm

1. **Require** : accelerometer set to 50 Hertz as input accel.x, accel.y, accel.z
2. **procedure** extractRealTimeFeature(accel.x, accel.y, accel.z)
3. *Set widowSize* ← *50 //set the widow size segment to 50*
4. ***Output***: $V^{50\times4}$
5. *Set row* ← 0
6. ***while*** *row* ≤ *windowSize* ***do***
7. *x = accel.X, y=accel.Y, z = accel.Z generate accelerometer values*
8. $radius \leftarrow \sqrt{x^2 + y^2 + z^2}$
9. ***if*** computeTiltAngle(x, radius) $approx\ \frac{\pi}{2}$ ***then***
10. $gravity \leftarrow tan^{-1}\left(\frac{\sqrt{x^2+z^2}}{y}\right)$
11. $theta \leftarrow$ computeTiltAngle(x, radius)
12. ***end if***
13. ***if*** computeTiltAngle(y, radius) $approx\ \frac{\pi}{2}$ ***then***
14. $gravity = tan^{-1}\left(\frac{\sqrt{x^2+y^2}}{z}\right)$
15. $theta =$ computeTiltAngle(y, radius)
16. ***end if***
17. ***If*** computeTiltAngle(z, radius) $approx\ \frac{\pi}{2}$ ***then***
18. $gravity \leftarrow tan^{-1}\left(\frac{\sqrt{y^2+z^2}}{x}\right)$
19. $theta \leftarrow$ computeTiltAngle(z, radius)
20. ***end if***
21. $period \leftarrow 2\pi\sqrt{\frac{radius}{gravity}}$
22. $graviMeter \leftarrow 4\pi^2 \times (1/period^2)$
23. $omega = (\frac{2\times theta}{period})$
24. $xPosition = radius \times sin\ (omega \times period)$
25. $yPosition = radius \times cos\ (omega \times period\ )$
26. ***If*** $graviMeter$ $\boldsymbol{approx}$ $\frac{\text{gravity}}{\pi^2}$ **Then**
27. $V_{row,0} \leftarrow period,\ V_{row,1} \leftarrow omega$
28. $V_{row,2} \leftarrow xPosition\ , V_{row,3} \leftarrow yPosition$
29. $row \leftarrow row + 1$
30. ***end if***

31. **end while**
32. **end procedure**
33. **return**
34. **Procedure computeTiltAngle(axis,radius)**
35. $Tilt\_around\_\alpha \leftarrow \left| tan^{-1}\left(\frac{\textbf{axis}}{\sqrt{radius}}\right)\right| \times \frac{180}{\pi}$
36. **end procedure**

Filtering and data segmentation: We automatically assign lightweight time-domain features a class label and associated Metabolic Equivalent Task (MET) used to measure daily energy expenditure based on related human activity performed.

Table 1. Predefined list of daily human activities table

Human Activity	Physical Activity Intensity Level (MET)
Laying	0.9
Standing	1.3
Sitting	1.8
Walking slowly	2.3
Walking downstairs	2.5
Walking upstairs	4.0
Normal walking pace	2.9
Cycling	3.0
Jogging	8.0
Rope jumping	10.0
Running	10.0
Brisk walking	3.3

It is found that the minimum, maximum and mean features improves the classification of static and dynamic human activities than other time-domain features [3][21][22][23][24][25]. Hence, we extracted them from $V^{50\times4}$ for each feature (period, omega, X, Y) as valley, peak and average [24]. We compress every 50 instances in *K* iterations as *top-k* of best classification features, because reduced dataset improves training and classification accuracy [6][8][10]. We set the size of *K* to 20 instead of storing every 50 by 4 real-time instances in *K* iterations. We extract *K* time-domain features as Human Activity Model (HAM), consequently producing *top-k* of compressed features per human activity as matrix $HAM^{K\times13}$ expanded to:

$$HAM^{K\times13} = \begin{bmatrix} min_{1,1}, max_{,1,2}\ mean_{1,3} min_{1,4}\ max_{1,5}\ mean_{1,6} \ldots\ldots & min_{1,10}\ max_{1,11}\ mean_{1,12}\ MET_{1,13} \\ \ldots\ldots\ldots & \ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ min_{K,1} max_{K,2} mean_{K,3} & min_{K,4}\ max_{K,5} mean_{K,6} \quad \ldots\ldots\ min_{K,10}\ max_{K,11}\ mean_{K,12}\ MET_{K,13} \end{bmatrix}$$

We also included the physical activity intensity level for each human activity known as MET published by [26]. The $HAM^{K\times(13)}$ is automatically annotated with commonly used predefined static and dynamic human activities listed in Table 1 using Algorithm

2, which requires a set of human activity labels with corresponding MET given as $dailyActivities^{12\times2}$. The subject will select each human activity he/she wants to create in line 4 as HAM. Subsequently, the subject will slide in the Smartphone in front pocket and perform the selected human activity for 2 minutes. In line 7 Algorithm 1 is called and returns matrix $V^{50\times4}$ of harmonic attributes, subsequently the maximum, minimum and mean are extracted per K iteration and stored to $HAM^{K\times13}$ in line 8 to 19, again the MET value is included as part of HAM.

Algorithm 2: Feature Extraction and data annotation algorithm

Require : daily human activities set as $dailyActivities^{12\times2}$

1. **procedure** ANNOTATEData()
2. *Set JSONTraining empty, append*
3. *Set* $K \leftarrow 20$
4. ***Output***: $HAM^{K\times13}$
5. *Select* ***humanActivity*** *and* ***MET*** *from* $dailyActivities^{12\times2}$
6. ***While*** *humanActivity exist* ***from*** $dailyActivities^{12\times2}$ ***do***
7. ***for*** $i = 1$ *to* K ***do***
8. $vectorFeature \leftarrow$ extractRealTimeFeature(accel.x, accel.y, accel.z)
9. $HAM_{i,0} \leftarrow max(vectorFeature_{PERIOD})$
10. $HAM_{i,1} \leftarrow min(vectorFeature_{PERIOD})$
11. $HAM_{i,2} \leftarrow mean(vectorFeature_{PERIOD})$
12. $HAM_{i,3} \leftarrow max(vectorFeature_{OMEGA})$
13. $HAM_{i,4} \leftarrow min(vectorFeature_{OMEGA})$
14. $HAM_{i,5} \leftarrow mean(vectorFeature_{OMEGA}$
15. $HAM_{i,6} \leftarrow max\ vectorFeature_{X})$
16. $HAM_{i,7} \leftarrow min(vectorFeature_{X})$
17. $HAM_{i,8} \leftarrow mean(vectorFeature_{X}$
18. $HAM_{i,9} \leftarrow max(vectorFeature_{Y})$
19. $HAM_{i,10} \leftarrow min(vectorFeature_{Y})$
20. $HAM_{i,11} \leftarrow mean(vectorFeature_{Y}$
21. $HAM_{i,12} \leftarrow MET$
22. ***end for***
23. *JSONTraining.append(* $HAM^{20\times13}$*, human Activity)*
24. *Select next* ***humanActivity*** *and* ***MET*** *from* $dailyActivities^{12\times2}$
25. ***end for***
26. **Write** *JSONTraining to JSONT file*
27. **end procedure**

Then, in line 22 the collected $HAM^{K\times13}$ features are stored to JavaScript Object Notation (JSON) collection. The next human activity and MET are selected, then the while loop is repeated until all human activities in $dailyActivities^{12\times2}$ are exhausted. Finally, the JSON collection is written to JSON file locally on Smartphone secure digital card in line 25.

3.3 Top-K personalized dataset collection

We randomly selected 13 subjects to collect personalized HAM every 2 minutes similar to [6][7]. We implemented Algorithm 1 and Algorithm 2 as part of our already developed training prototype published in [10]. The training prototype User Interface (UI) are sequenced from number 1 to number 7 to guide a subject to collect little personalized dataset (see Fig. 5). We installed our user-friendly personalized training prototype on Samsung Galaxy Grand Prime+ Smartphone. All the commonly used human activities listed on Table 1 are preloaded as dropdown menu as shown in UI-1 and UI-4 in Fig 5.

Fig. 5. Top-k data collection prototype developed by [10]

We implemented 10 seconds delay indicated in UI-4 in Fig 5 before recording each HAM activities. Every time, the selected and performed human activity is removed from the pre-loaded menu to prevent replication. The subject must place the Smartphone inside his / her right front-pocket under the supervision of researcher (see Fig. 6).

Fig. 6. Right Front pocket location

The On every occasion, the researcher selects a specific human activity from dropdown-menu. The subject will slot in the Smartphone inside the front right pocket (see Fig 9), and then the start sound will be triggered after 10 seconds. A subject is given 4 minutes breaks in between each human activity.

The subject starts to perform a selected activity to collect HAM features until the stop sound is triggered after 2 minutes. The subject will stop and give the Smartphone to the researcher, then our prototype prompts the researcher for the next human activity as indicated in Fig 8 UI-5. The researcher selects the next human activity until all the pre-loaded human activities are exhausted. Thereafter, all generated HAM features with labels are automatically written on SD card (see Appendix A). All 13 collected *top-k* datasets files were transferred and merged into a single Comma Separated Values (CSV) file called Real-time Personalized dataset with 2860 rows as listed in Table 2.

Table 2. Top-k Real-time Personalized Dataset

Physical Activity	Total Activities
Laying	260
Standing	260
Sitting	260
Walking slowly	260
Walking downstairs	260
Walking upstairs	260
Normal walking pace	260
Jogging	260
Rope jumping	260
Running	260
Brisk walking	260

4 Selection of Suitable Classification Algorithm

In this Section, we selected 6 state of the art classification algorithms (C4.5, KNN, Support Vector Machines (SVM), Boosted Trees (BT), Random Forest (RT) and Naïve Bayes (NB)) to benchmark our proposed *top-k* dataset in terms of accuracy and precision to select the most suitable classification algorithm similar to [6] The preliminary results presented by [6] using own personalized PAMAP2 dataset on 4 algorithms (C4.5, KNN, SVM and Naïve Bayes) revealed the accuracy of 85.03%, 87.62%, 62.31% and 74.14% respectively. In this study, we selected *R* programming languages because it is free, easy to use, allows researchers to use predefined algorithms and confusion matrix [26][27] listed in Table 3.

Table 3. Selected Algorithm for Simulation [27][28]

Algorithm	R Package used	Method
C4.5	RWeka- is Weka package to allow R to use Weka methods.	J48
Naïve Bayes	Caret	nb
Support Vector Machine	Caret	svmLinear
K Nearest Neighbor	Caret	Knn
Boosted Trees	Caret	gbm (BT)
Random Forest	Random Forest	rf
Confusion Matrix	e1071	Confusion Matrix ()

The benchmarking *R-Source-Code G.1* algorithm proposed by [27][28] is used to simulate and compare existing *R* machine learning algorithms presented in Table 3. Firstly, we loaded different libraries in line 3 to line 7 of implemented benchmarking required by classification models described in Table 3.

R-Source Code 1:Benchmarking Algorithms

```
1.  # Using K fold Cross Validation
2.  #Load library
3.  library(caret)
4.  library(klaR)
5.  library(e1071)
6.  library(RWeka)
7.  #load testing dataset
8.  setwd("C:\\dataset\\ ")
9.  JSONData=  read.csv ("Top-k Real-time Personalized.csv" ,header =TRUE)
10. #cross validation chosen
11. k_fold  = 10
12. k  = 1
13. #Train the model
14. train_control = train(method="cv", number =k_fold, repeats = k)
15. #train the model to predict Human activities on personalized dataset
16. model= train(label~., data= JSONData, trControl=train_control,
    method="knn")
17. #Print the results of the prediction model
18. confusionMatrix(model)
```

We used 10-fold cross validation since the dataset is limited to 260 per human activity to determine the reliability of each 6 selected model [4]. For reproducibility and future comparison, we implemented *k-fold* cross validation given as *R-Source-Code G.1* to conduct comparison to select best classification algorithm, where our collected personalized dataset is partitioned into *K* equal sub-sets; such that all simulated algorithms can use *K-1* as training dataset and *1* as testing dataset. The cross-validation results are given in precision and accuracy and summarized in accuracy and precision graphs. We ran the implemented *R-Source-Code G.1* for each of the 6 algorithms (SVM, C4.5, KNN, NB, RF and BT) listed in Table 3 similar to [18][27][28]. The simulated *R-Source-Code G.1* is used to benchmark our *top-k* personalized dataset. The results presented in [6][7][29][30] indicates that tree-oriented algorithms perform far

better than simpler algorithm with accuracy and precision of 98% using little personalized dataset [29]. The results are summarized in Precision and Accuracy in Fig. 7 and Fig. 8.

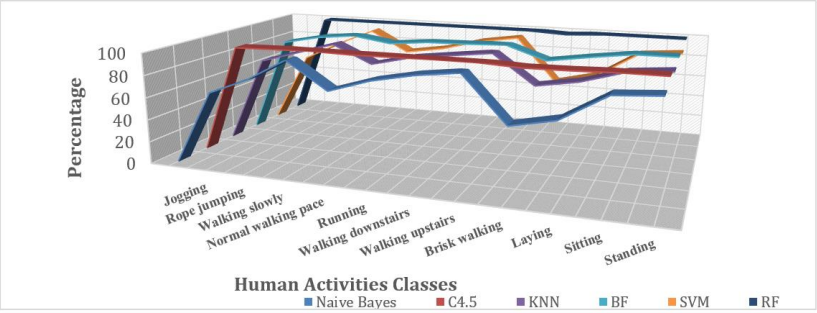

Fig. 7. Precision results per human activity classes using Real-time Personalized Training Master

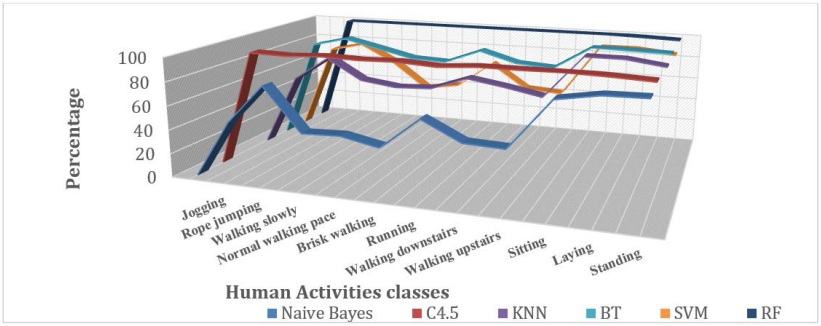

Fig. 8. Accuracy results per human activity classes using Real-time Personalized Training Master

The results presented in Fig.7 and Fig.8 show improved accuracy and precision of 100% with RF and C4.5 algorithms in all static and complex human activities with 100% reported TP in all human activities with no TN, FP and FN, RF outperformed its predecessor BT, which scored precision and accuracy of 80% and 90% respectively. Our results confirm that tree-oriented algorithms (RF, C4.5 and BT) are suitable for smallest training dataset as compared to SVM, KNN and Naïve Bayes. The results as compared to [6][23][29] shows significant drop from 87.62% to 40% and 74.14% to 60% in accuracy and precision on simpler algorithms with little dataset, and confirm existing gap between tree-oriented and simpler algorithms (SVM, KNN and Naïve Bayes) reported in [6][23][29].This results are similar to the results reported by [23],

where Naïve Bayes scored accuracy and precision of 42.30% and 47.61% respectively using smaller and reduced dataset. However, Fig.7 and Fig.8 shows improved accuracy and precision above 90% on static human activities in all simpler algorithms due to regularized harmonic features. Hence, personalized static human activities can be replicated to multiple subjects.

5 Conclusion and Future Work

In this paper, we presented a personalized model to collect *top-k* personalized dataset to select a suitable classification algorithm. Harmonic motion based on simple pendulum was used to augment Signal Magnitude Vector to capture human bodily postures. We proposed novel filtering technique based on gravimeter to remove noise in order to personalized dataset from 13 subjects; the dataset was benchmarked using state of the art machine learning algorithms. We found that our dataset is suitable for tree-oriented algorithms such as RF, C4.5 and BT, because each feature creates as tree-like hierarchical structures. However, the dataset is not suitable for simpler algorithms such as KNN and Naïve Bayes. In future we intend to propose a hybrid model combining KNN and Naïve Bayes for smallest datasets.

6 Acknowledgement

We would like to thank the Faculty of Information Communication Technology and department of Computer Science, Tshwane University of Technology for financial and other logistical support for the success this project. This project is ethically approved by the Faculty Committee of Research Ethics of Tshwane University of Technology (FCRE ICT Ref#2016=06=001(2) = GadebeML).

7 References

[1] Lockhart, Jeffrey W., Tony Pulickal, and Gary M. Weiss. "Applications of mobile activity recognition." In Proceedings of the 2012 ACM Conference on Ubiquitous Computing, pp. 1054-1058. ACM, 2012. https://doi.org/10.1145/2370216.2370441

[2] Su, Xing, Hanghang Tong, and Ping Ji. "Activity recognition with smartphone sensors." Tsinghua science and technology 19, no. 3 (2014): 235-249. https://doi.org/10.1109/tst.2014.6838194

[3] Slim, S.O., Atia, A., Elfattah, M.M. and Mostafa, M.S.M.,Survey on Human Activity Recognition based on Acceleration Data.International Journal of Advanced Computer Science and Applications (IJACSA). 2019 Vol 10:3 pp: 84:98. https://doi.org/10.14569/ijacsa.2019.0100311

[4] Lockhart, Jeffrey W., and Gary M. Weiss. "Limitations with activity recognition methodology & data sets." In Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication, pp. 747-756. ACM, 2014. https://doi.org/10.1145/2638728.2641306

[5] Vo, Quang Viet, Minh Thang Hoang, and Deokjai Choi. "Personalization in mobile activity recognition system using K-medoids clustering algorithm." International Journal of Distributed Sensor Networks 9, no. 7 (2013): 315841. https://doi.org/10.1155/2013/315841
[6] Reiss, Attila, and Didier Stricker. "Introducing a new benchmarked dataset for activity monitoring." In 2012 16th International Symposium on Wearable Computers, pp. 108-109. IEEE, 2012. https://doi.org/10.1109/iswc.2012.13
[7] Zhang, Mi, and Alexander A. Sawchuk. "USC-HAD: a daily activity dataset for ubiquitous activity recognition using wearable sensors." In Proceedings of the 2012 ACM Conference on Ubiquitous Computing, pp. 1036-1043. ACM, 2012. https://doi.org/10.1145/2370216.2370438
[8] Anguita, Davide, Alessandro Ghio, Luca Oneto, Xavier Parra, and Jorge Luis Reyes-Ortiz. "A public domain dataset for human activity recognition using smartphones." In Esann. 2013. https://doi.org/10.1016/j.neucom.2015.07.085
[9] Gadebe, Moses L., and Okuthe P. Kogeda. "Personification of Bag-of-Features Dataset for Real Time Activity Recognition." In 2016 3rd International Conference on Soft Computing & Machine Intelligence (ISCMI), pp. 73-78. IEEE, 2016. https://doi.org/10.1109/iscmi.2016.27
[10] Gadebe, M.L., Kogeda, O.P. and Ojo, S.O., 2018, November. Personalized Real Time Human Activity Recognition. In 2018 5th International Conference on Soft Computing & Machine Intelligence (ISCMI) (pp. 147-154). IEEE https://doi.org/10.1109/iscmi.2018.8703240
[11] Reyes-Ortiz, Jorge-L., Luca Oneto, Albert Samà, Xavier Parra, and Davide Anguita. "Transition-aware human activity recognition using smartphones." Neurocomputing 171 (2016): 754-767. https://doi.org/10.1016/j.neucom.2015.07.085
[12] Ugulino, W.; Cardador, D.; Vega, K.; Velloso, E.; Milidiu, R.; Fuks, H. 2012. Wearable Computing: Accelerometers' Data Classification of Body Postures and Movements. Proceedings of 21st Brazilian Symposium on Artificial Intelligence. Advances in Artificial Intelligence - SBIA 2012. In: Lecture Notes in Computer Science, pp. 52-61. Curitiba, PR: Springer Berlin / Heidelberg, 2012. ISBN 978-3-642-34458-9. https://doi.org/10.1007/978-3-642-34459-6_6
[13] Kwapisz, J.R., Weiss, G.M. and Moore, S.A., 2011. Activity recognition using cell phone accelerometers. ACM SigKDD Explorations Newsletter, 12(2), pp.74-82. https://doi.org/10.1145/1964897.1964918
[14] Gary M. Weiss, Kenichi Yoneda, and Thaier Hayajneh. Smartphone and Smartwatch-Based Biometrics Using Activities of Daily Living. IEEE Access, 7:133190-133202, Sept. 2019. https://doi.org/10.1109/access.2019.2940729
[15] Nguyen, Le T., Ming Zeng, Patrick Tague, and Joy Zhang. "Recognizing new activities with limited training data." In Proceedings of the 2015 ACM International Symposium on Wearable Computers, pp. 67-74. ACM, 2015. https://doi.org/10.1145/2802083.2808388
[16] Harous, Saad, Mohamed Adel Serhani, Mohamed El Menshawy, and Abdelghani Benharref. "Hybrid obesity monitoring model using sensors and community engagement." In 2017 13th International Wireless Communications and Mobile Computing Conference (IWCMC), pp. 888-893. IEEE, 2017. https://doi.org/10.1109/iwcmc.2017.7986403
[17] Harous, Saad, Mohamed El Menshawy, Mohamed Adel Serhani, and Abdelghani Benharref. "Mobile health architecture for obesity management using sensory and social data." Informatics in Medicine Unlocked 10 (2018): 27-44. https://doi.org/10.1016/j.imu.2017.12.005
[18] Zhang, Z., 2016. Naïve Bayes classification in R. Annals of translational medicine, 4(12).
[19] Palais, B. and Palais, R., 2007. Euler's fixed point fixed-point the axis of a rotation. Journal of fixed-point theory and applications, 2(2), pp.215-220. https://doi.org/10.1007/s11784-007-0042-5
[20] Pedley, M., 2013. Tilt sensing using a three-axis accelerometer. Freescale semiconductor application note, 1, pp.2012-2013.

[21] Keogh, E., Chu, S., Hart, D. and Pazzani, M., 2001, November. An online algorithm for segmenting time series. In Proceedings of IEEE international conference on data mining (pp. 289-296). IEEE. https://doi.org/10.1109/icdm.2001.989531
[22] Keogh, E., Chakrabarti, K., Pazzani, M. and Mehrotra, S., 2001. Dimensionality reduction for fast similarity search in large time series databases. Knowledge and information Systems, 3(3), pp.263-286. https://doi.org/10.1007/pl00011669
[23] Kose, M., Incel, O.D. and Ersoy, C., 2012, April. Online human activity recognition on smart phones. In Workshop on Mobile Sensing: From Smartphones and Wearables to Big Data 16(2012), pp. 11-15).
[24] Al Huda, F., Tolle, H. and Asmara, R.A., 2017. Realtime online daily living activity recognition using head-mounted display. International Journal of Interactive Mobile Technologies (iJIM), 11(3), pp.67-77.39 https://doi.org/10.3991/ijim.v11i3.6469
[25] Rivero-Rodriguez, A., Pileggi, P. and Nykänen, O.A., 2016. Mobile context-aware systems: technologies, resources and applications. International Journal of Interactive Mobile Technologies (iJIM), 10(2), pp.25-32. https://doi.org/10.3991/ijim.v10i2.5367
[26] Ainsworth, B.E., Haskell, W.L., Whitt, M.C., Irwin, M.L., Swartz, A.M., Strath, S.J., O Brien, W.L., Bassett, D.R., Schmitz, K.H., Emplaincourt, P.O. and Jacobs, D.R., 2000. Compendium of physical activities: an update of activity codes and MET intensities. Medicine and science in sports and exercise, 32(9; SUPP/1), pp. S498-S504. https://doi.org/10.1097/00005768-200009001-00009
[27] Kuhn, Max. "A Short Introduction to the caret Package." R Found Stat Comput (2015): 1-10.
[28] Kuhn, M., 2008. Building predictive models in R using the caret package. Journal of statistical software, 28(5), pp.1-26.
[29] Weiss, G.M and Lockhart, J.W., 2012. The impact of personalization on Smartphone-based Activity Recognition. Association for the Advancement of Artificial Intelligence, Technical Report, pp 98-104.
[30] Christiana, A.O., Gyunka, B.A. and Oluwatobi, A.N., 2020. Optimizing Android Malware Detection Via Ensemble Learning. International Journal of Interactive Mobile Technologies, 14(9). https://doi.org/10.3991/ijim.v14i09.11548

8 Authors

Moses L. Gadebe obtained master's degree in Computer Science at Tshwane University of Technology, Pretoria, South Africa in 2013. He is currently Lecturer at department of Computer Science, faculty of Information and Communication Technology at Tshwane University of Technology, Pretoria, South Africa. Previously he was a Web Developer in from 2000 – 2003 in Arivia.kom. He has published 3 journals and 3 conference papers and one patent.

Okuthe P. Kogeda obtained a doctorate degree in Computer Science from University of the Western Cape in Cape Town, South Africa in 2009. He is currently Associate Professor at Department of Computer Science & Informatics, Natural and Agricultural Sciences faculty at University of the Free State, Bloemfontein, South Africa. Previously he was a Senior Lecturer, Chair of Departmental Research & Innovation Committee, and Head of Postgraduate Section in the Computer Science Department, ICT Faculty at Tshwane University of Technology, South Africa from 2011 to 2019. He was a Senior Lecturer in the Computer Science Department at University of Fort Hare in Eastern Cape, South Africa from 2009 to 2011. He was a Lecturer in the Computer Science

Department at University of the Western Cape in Cape Town, South Africa from 2004 to 2009. He was a Lecturer at University of Nairobi in Nairobi, Kenya from 1999 to 2000. He is a member of IITPSA, IEEE and IAENG. He is NRF rated researcher since 2015. He has published three books, 26 journal articles, 7 Chapters in books, over 60 refereed conference papers, and one patent.

Article submitted 2020-07-12. Resubmitted 2020-07-30. Final acceptance 2020-07-31. Final version published as submitted by the authors.

9 Appendix: Top-k Personalized Dataset

{"Standing":[[14,15,14,46,47,46,6,33,6,4,17,4,1.3], [14,14,14,46,47,46,6,7,6,4,5,4,1.3], [17,24,14,46,47,46,5,7,4,3,5,3,1.3], [14,14,14,46,47,46,6,7,6,4,4,4,1.3], [14,14,14,46,47,46,6,6,6,4,4,4,1.3], [14,14,14,46,46,46,6,6,6,4,4,4,1.3], [14,14,14,46,47,46,6,6,6,4,4,4,1.3], [14,14,14,46,47,46,6,7,6,4,4,4,1.3], [14,14,14,46,47,46,6,6,6,4,4,4,1.3], [14,14,14,46,47,46,6,6,6,4,4,4,1.3], [14,14,14,46,47,46,6,6,6,4,4,4,1.3], [14,14,14,46,47,46,6,6,6,4,4,4,1.3], [14,14,14,46,47,46,6,7,6,4,4,4,1.3], [14,14,14,46,47,46,6,7,6,4,4,4,1.3], [14,14,14,46,47,46,6,6,6,4,4,4,1.3], [14,14,14,46,47,46,6,6,6,4,4,4,1.3], [14,14,14,46,47,46,6,6,6,4,4,4,1.3], [14,14,14,46,47,46,6,6,6,4,4,4,1.3], [14,14,14,46,47,46,6,6,6,4,4,4,1.3], [14,14,14,46,47,46,6,6,6,4,4,4,1.3]	"Sitting":[[19,19,13,38,38,37,26,27,25,18,18,18,1.8], [19,19,19,38,38,38,26,26,26,18,18,18,1.8], [19,19,19,38,38,38,26,26,26,18,18,18,1.8], [19,19,19,38,38,38,26,26,26,18,18,18,1.8], [19,19,19,38,38,38,26,26,26,18,18,18,1.8], [19,19,19,38,38,38,26,26,26,18,18,18,1.8], [19,19,19,38,38,38,26,26,26,18,18,18,1.8], [19,19,19,38,38,38,26,26,26,18,18,18,1.8], [19,19,19,38,38,38,26,26,26,18,18,18,1.8], [19,19,19,38,38,38,26,26,26,18,18,18,1.8], [19,19,19,38,38,38,26,26,26,18,18,18,1.8], [19,19,19,38,38,38,26,26,26,18,18,18,1.8], [19,19,19,38,38,38,26,26,26,18,18,18,1.8], [19,19,19,38,38,38,26,26,26,18,18,18,1.8], [19,19,19,38,38,38,26,26,26,18,18,18,1.8], [19,19,19,38,38,38,26,26,26,18,18,18,1.8], [19,19,19,38,38,38,26,26,26,18,18,18,1.8], [19,19,19,38,38,38,26,27,26,18,18,18,1.8]	,"Normal walking pace":[[21,38,13,53,62,30,10,26,3,5,23,2,2.9], [23,33,13,50,117,19,16,34,8,12,23,5,2.9], [24,47,13,51,93,23,15,42,1,8,18,1,2.9], [22,37,13,54,65,36,12,27,4,5,10,2,2.9], [18,30,12,46,102,20,24,52,9,15,21,7,2.9], [20,32,13,50,79,41,7,13,3,4,7,1,2.9], [21,27,15,52,61,37,11,23,3,6,10,1,2.9], [22,56,13,41,100,17,31,58,15,22,31,10,2.9], [20,36,13,46,59,33,11,33,4,7,20,3,2.9], [19,28,12,44,69,20,16,25,4,13,28,3,2.9], [25,45,12,51,87,30,21,31,6,12,20,4,2.9], [18,24,15,53,68,45,8,19,4,4,10,1,2.9], [17,28,13,42,86,24,24,54,7,16,23,6,2.9], [20,36,13,45,86,32,16,54,3,10,22,1,2.9], [21,29,12,50,70,19,14,26,2,9,31,1,2.9], [24,44,13,40,72,16,28,49,12,19,32,13,2.9], [18,22,15,51,64,31,9,22,4,5,13,2,2.9], [24,43,14,35,58,18,19,33,7,18,32,8,2.9], [25,43,17,43,58,27,28,46,7,16,27,4,2.9], [17,23,14,54,66,42,11,28,4,5,10,2,2.9]	,"Walking downstairs":[[24,34,15,50,80,25,14,54,5,9,22,2,2.5], [18,26,13,38,82,18,22,34,9,20,28,2,2.5], [19,24,14,43,68,20,15,34,5,12,23,2,2.5], [21,35,13,40,82,18,19,34,11,17,28,3,2.5], [23,37,15,47,82,28,13,19,8,9,17,3,2.5], [23,34,15,49,82,28,12,22,4,8,19,4,2.5], [21,52,16,43,87,26,23,50,16,15,22,8,2.5], [26,52,13,43,87,31,15,50,7,9,16,6,2.5], [20,28,17,50,94,28,21,40,10,11,22,8,2.5], [25,42,11,40,55,12,22,32,11,17,37,8,2.5], [22,35,11,38,63,12,11,18,4,13,37,2,2.5], [21,30,15,43,65,23,21,39,6,16,27,3,2.5], [23,41,16,49,72,23,17,22,11,11,27,6,2.5], [21,39,13,47,78,32,14,27,1,9,13,1,2.5], [22,37,17,41,78,29,15,28,11,12,19,7,2.5], [22,25,17,56,79,37,17,33,4,8,18,3,2.5], [21,29,16,44,77,32,15,31,4,9,12,3,2.5], [18,24,12,34,62,12,20,51,4,19,40,3,2.5], [22,31,16,49,80,30,17,36,7,11,22,2,2.5], [19,27,16,35,69,24,20,29,7,18,25,5,2.5]
]"Laying":[	]"Walking upstairs":[	],Brisk walking":[	], "Running":[

[14,14,13,48,48,42,3,14,3,2,10,2,0.9],	[20,38,12,44,65,22,18,32,5,10,20,7,4.0],	[25,44,11,63,100,26,16,38,2,6,24,2,3.3],	[21,43,13,47,190,23,34,133,15,17,26,8,10],
[14,14,14,48,48,48,3,3,3,2,2,2,0.9],	[26,41,12,58,81,33,11,21,4,6,19,1,4.0],	[26,62,11,45,124,12,27,71,9,19,33,9,3.3],	[28,58,15,79,159,23,47,96,13,12,24,4,10],
[14,14,14,48,48,48,3,3,3,2,2,2,0.9],	[18,23,12,35,50,16,21,33,9,19,25,13,4.0],	[20,26,13,51,81,16,21,37,4,13,30,2,3.3],	[30,58,15,69,160,17,43,96,7,14,34,5,10],
[14,14,14,48,48,48,3,3,3,2,2,2,0.9],	[19,21,13,44,56,34,17,33,6,11,23,5,4.0],	[23,32,13,51,109,16,32,68,12,18,30,6,3.3],	[21,35,15,57,130,17,38,84,18,20,34,9,10],
[14,14,14,48,48,48,3,3,3,2,2,2,0.9],	[23,42,13,52,73,20,11,20,6,7,18,2,4.0],	[17,23,9,45,81,8,17,33,7,17,46,4,3.3],	[21,35,16,54,141,26,31,60,14,16,24,7,10],
[14,14,14,48,48,48,3,3,3,2,2,2,0.9],	[17,30,10,43,98,20,17,55,3,11,19,6,4.0],	[25,42,9,52,81,8,36,68,7,14,46,8,3.3],	[25,37,16,74,162,24,48,91,16,17,28,6,10],
[14,14,14,48,48,48,3,3,3,2,2,2,0.9],	[21,24,16,56,75,30,10,18,6,5,9,3,4.0],	[20,28,14,58,69,37,16,30,5,8,18,2,3.3],	[28,57,11,65,137,25,32,61,18,14,24,6,10],
[14,14,14,48,48,48,3,3,3,2,2,2,0.9],	[17,21,14,32,51,18,14,21,7,16,32,8,4.0],	[16,22,10,41,64,11,15,30,8,17,36,4,3.3],	[27,54,15,72,187,26,37,126,12,13,23,6,10],
[14,14,14,48,48,48,3,3,3,2,2,2,0.9],	[17,25,11,40,75,27,9,17,2,7,17,3,4.0],	[17,23,13,51,80,18,14,30,7,11,32,3,3.3],	[28,62,11,80,187,19,40,126,15,16,28,4,10],
[14,14,14,48,48,48,3,3,3,2,2,2,0.9],	[24,44,15,56,67,24,10,20,5,5,9,3,4.0],	[20,27,14,40,66,15,20,40,4,15,26,8,3.3],	[25,62,6,67,187,6,31,73,5,22,54,5,10],
[14,14,14,48,48,48,3,3,3,2,2,2,0.9],	[21,28,13,35,54,24,16,24,5,14,27,8,4.0],	[25,50,12,48,78,15,18,45,2,12,29,1,3.3],	[27,38,16,84,178,22,32,46,18,14,28,3,10],
[14,14,14,48,48,48,3,3,3,2,2,2,0.9],	[22,44,14,45,60,34,10,20,5,7,18,5,4.0],	[16,27,9,41,82,11,16,46,2,18,35,1,3.3],	[25,36,17,72,150,30,32,70,7,11,19,2,10],
[14,14,14,48,48,48,3,3,3,2,2,2,0.9],	[22,26,16,59,82,34,9,13,5,5,10,2,4.0],	[19,23,14,50,65,32,22,46,8,11,19,4,3.3],	[23,37,13,56,146,14,36,95,7,20,37,2,10],
[14,14,14,48,48,48,3,3,3,2,2,2,0.9],	[19,30,12,37,62,26,16,32,7,14,21,6,4.0],	[21,28,14,50,70,20,19,26,9,12,25,3,3.3],	[22,40,11,56,196,13,27,49,6,17,33,2,10],
[14,14,14,48,48,48,3,3,3,2,2,2,0.9],	[17,23,12,48,68,33,8,32,2,5,13,2,4.0],	[24,35,14,48,79,21,35,64,7,16,25,10,3.3],	[24,40,10,74,196,18,40,117,17,17,30,2,10],
[14,14,14,48,48,48,3,3,3,2,2,2,0.9],	[21,42,12,48,73,28,15,25,6,11,24,2,4.0],	[23,31,16,57,71,26,16,26,5,7,10,5,3.3],	[19,28,9,49,91,14,31,64,9,18,33,7,10],
[14,14,14,48,48,48,3,3,3,2,2,2,0.9],	[24,37,10,54,84,25,18,53,4,9,20,2,4.0],	[24,34,11,59,110,13,40,102,5,14,22,5,3.3],	[24,41,16,64,175,23,46,118,12,17,29,8,10],
[14,14,14,48,48,48,3,3,3,2,2,2,0.9],	[20,26,10,50,70,26,7,14,1,4,9,1,4.0],	[23,47,14,58,79,48,15,29,3,7,13,2,3.3],	[22,38,13,62,153,26,38,96,12,15,25,8,10],
[14,14,14,48,48,48,3,3,3,2,2,2,0.9],	[17,29,12,34,59,15,17,23,8,17,26,6,4.0],	[19,35,8,50,140,13,33,68,8,21,32,7,3.3],	[29,43,14,83,178,11,23,48,4,10,23,3,10],
[14,14,14,48,48,48,3,3,3,2,2,2,0.9]]	[18,25,13,46,76,25,12,23,4,8,26,4,4.0]]]	[23,38,15,51,84,25,20,57,5,10,15,2,3.3]]	[26,42,10,64,189,23,32,127,10,15,24,5,10]]