Prediction of Health Condition using Human Activity Recognition

Abraham Jacob M
abrahamjacob@iisc.ac.in

Alankar Adarsh
alankara@iisc.ac.in

Issaac Kommineni
issaack@iisc.ac.in

Nitin Anand
nitina@iisc.ac.in

***Abstract*—Human Activity Recognition (HAR) is an important task in the field of machine learning and computer vision, which aims to automatically identify and classify human activities based on sensor data collected from wearable devices or smart phones. In recent years, various approaches have been proposed for HAR, including deep learning, feature-based methods, and hybrid methods that combine both. There are various techniques to predict the activity from the sensor information. Our project predicts the Health Index (HI) ranging from 0 to 100 using duration of each activity performed by a person.**

***Index Terms*—HAR, Classification, Regression, MSE,** R^2

I. Introduction

HAR is the process of interpreting human activities using computer and machine vision technology. HAR is a classification activity using sensors (or gadgets attached with sensors) that can sense various human movements. HAR has many potential applications, such as healthcare monitoring, sports performance analysis, and personalized coaching.

An important field of recognizing human activity is in the field of health. With the evolution of wearable devices that includes sensors like accelerometer and gyroscope, it is easy to record the movements of a person periodically throughout the day. These movements can be mapped to various human activities of daily living (ADL) like standing, sitting, laying, walking, walking up the stairs, walking down the stairs and so on. Upon analysing the duration of the activities performed by a person, we can define his lifestyle (physically active, sedentary, etc.). This lifestyle is a vital measure of the health of a person.

II. Related Works

There are many available projects that implements HAR. One such study was done where movements or actions of a person based on sensor data were identified. Movements are captured while carrying a waist-mounted smartphone with embedded inertial sensors.

- Accelerometers is an electronic sensor that measures the acceleration forces acting on an object, in order to determine the object's position in space and monitor the object's movement
- Gyroscope is a device that can measure and maintain the orientation and angular velocity of an object. These are more advanced than accelerometers. These can measure the tilt and lateral orientation of the object whereas accelerometer can only measure the linear motion.

Using embedded sensors, data captured included 3-axial linear acceleration and 3-axial angular velocity at a constant rate of 50Hz. Based on triaxial acceleration from the accelerometer (total acceleration) and the estimated body acceleration and triaxial angular velocity from the gyroscope, A 561 feature vector with time and frequency domain variables was treated as feature variables. Target variable was the one of the activities mentioned above. Various modelling algorithms have been used and was compared among them. This study doesn't predict or conclude on the health aspect of a person and the study was performed on the database for a very small duration for each activity. Also, each activity is captured for same duration, making it difficult to judge the overall HI of a person.

III. Our Approach

Our main aim is to predict the HI of a person based on the data captured for his/her activities over a span of average 24 hours. As per the previous study (referred in previous section), activities can be classified for a person based on existing data. In order to address our main problem statement i.e to predict HI, we formulated an equation to establish a relation between HI and each activity based on it's relevance in overall health of a person. Further to it, we generated a synthetic dataset which includes sensor measurements of all activities for one day. This synthetic data has been generated using our sample c-program which actually does it on purely random basis. This random data generation helped us in generating all possible combinations of activities performed by a person in a day. In below section, We first explain the details of data munging and activity class predictions based on existing data. Then we explain our work done on new synthetic data.

A. Dataset

Dataset consists of features which are combination of signals from sensors. Below are 3-axial signals in the X, Y and Z directions:

- tBodyAcc-XYZ
- tGravityAcc-XYZ
- tBodyAccJerk-XYZ
- tBodyGyro-XYZ
- tBodyGyroJerk-XYZ
- fBodyAcc-XYZ
- fBodyAccJerk-XYZ
- fBodyGyro-XYZ

Here, X, Y and Z axis usually aligned with the direction of forward movement, lateral movement and the direction of gravity.

Other signals generated from sensors involves rotation, jerk, gravitation acceleration etc. of the body, are as below:

- tBodyAccMag
- tGravityAccMag
- tBodyAccJerkMag
- tBodyGyroMag
- tBodyGyroJerkMag
- fBodyAccMag
- fBodyAccJerkMag
- fBodyGyroMag
- fBodyGyroJerkMag

These signals are used to estimate attributes like mean, standard deviation, entropy etc. Combination of above mentioned signals and attributes define various features of dataset for the classification model to identify the activity. Assuming that above dataset is extrapolated for the entire duration of a day, duration of each activity can be estimated and HI is calculated based on the domain knowledge using the formula

$$\begin{aligned} HI = B + W_1 * LayingTime + \\ W_2 * StandingTime + W_3 * SittingTime + \\ W_4 * WalkingTime + W_5 * WalkingUpStairsTime + \\ W_6 * WalkingDownStairsTime \end{aligned} \quad (1)$$

Where B is the Bias and W_i are the weights for the duration of each activity.

In order to emulate the deviation in values, a random noise is added to HI. Following features are considered for regression problem.

- LayingTime
- StandingTime
- SittingTime
- WalkingTime
- WalkingUpStairsTime
- WalkingDownStairsTime

A dataset consisting of 7,93,865 rows and 8 columns are generated using a program.

B. Data Pre-processing

There were some special characters in column names which were removed to avoid any undesired behaviour.

1) Data distribution with respect to each class: Imbalance in data may result in incorrect classification hence, balanced data is important for good model accuracy. We can observe that data is well balanced across activities as shown in Fig. 1

2) Principal Component Analysis: Sensor data is often high-dimensional and contains large number of features, which can make it difficult to process and analyse. PCA involves transforming the sensor data into a new set of variables, called principal components, which are a linear combination of the original features. Applying PCA on HAR dataset had reduced the number of features from 561 to 157.

3) Data cleaning: Data cleaning involves removing any noise or outliers from the sensor data, correcting missing or incorrect values, and transforming the data to be in a consistent format. This ensures that the data used in the HAR model is accurate and reliable. For the dataset which is considered, there are no outliers and can be observed from the box plot as shown in Fig. 2

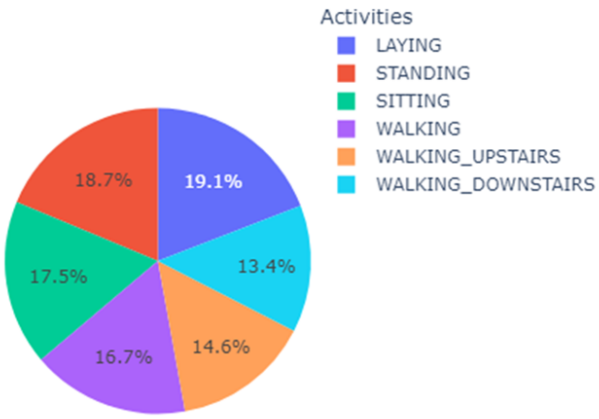

Fig. 1. Distribution of activities across data

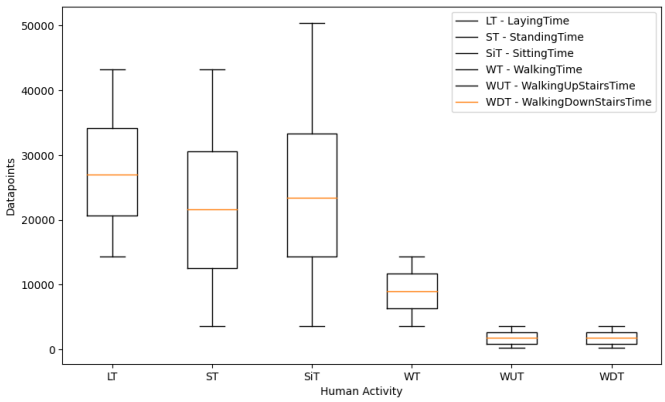

Fig. 2. Distribution of activities across data

C. Model

In the current project, we established a linear relationship between the duration of each activity and HI. Models like Linear regression, Bayesian ridge, Ridge, Lasso, ElasticNet, ElasticNetCV and Huberregressor are used.

- **Linear Regression** - A linear approach to modeling the relationship between a dependent variable and one or more independent variables.
- **BayesianRidge** - Incorporates Bayesian regularization to prevent overfitting by assigning a prior probability distribution to the regression coefficients.
- **Ridge regression** - Incorporates L2 regularization to prevent overfitting by adding a penalty term to the sum of squared coefficients.
- **Lasso regression** - Incorporates L1 regularization to prevent overfitting by adding a penalty term to the absolute sum of coefficients.
- **ElasticNet** - Combines L1 and L2 regularization to prevent overfitting by adding both penalty terms to the sum of squared coefficients.
- **ElasticNetCV** - A model with built-in cross-validation for hyperparameter tuning.

TABLE I
MODEL AND HYPERPARAMETERS

Model	Hyperparameter1		Hyperparameter2	
	Name	Value	Name	Value
BayesianRidge	alpha_1	1e-06	lambda_1	1e-06
Ridge	alpha	1.0	solver	auto
Lasso	alpha	1.0	selection	cyclic
ElasticNet	alpha	1.0	l1_ratio	0.5
ElasticNetCV	eps	0.001	l1_ratio	0.5
Huberregressor	epsilon	1.35	alpha	0.0001

TABLE II
MSE AND R^2 ACROSS DIFFERENT MODELS

Model	MSE		R^2	
	Training	Testing	Training	Testing
Linear regression	0.122	0.121	0.991	0.991
Bayesian ridge	0.122	0.121	0.991	0.991
Ridge	0.122	0.121	0.991	0.991
Lasso	0.122	0.121	0.991	0.991
ElasticNet	0.122	0.121	0.991	0.991
ElasticNetCV	0.124	0.123	0.990	0.990
Huberregressor	0.122	0.121	0.991	0.991

- **HuberRegressor** - A model that is robust to outliers by minimizing the sum of absolute errors for small errors and the sum of squared errors for larger errors.

Fig. 3 provides the prediction results of HI with respect to actual data with linear regression.

IV. RESULTS

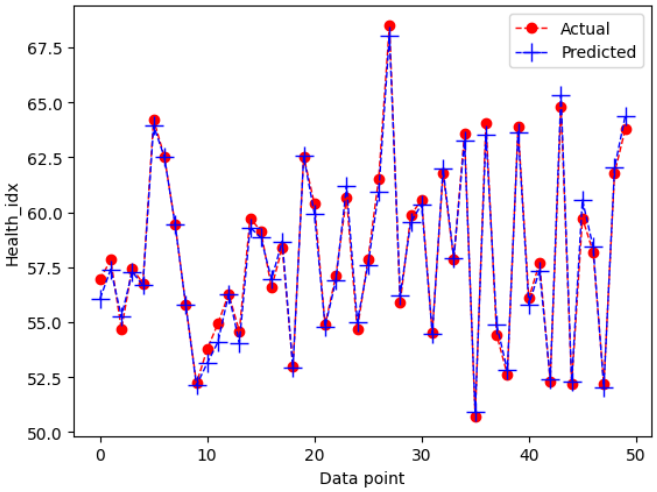

Fig. 3. Linear Regression (Predicted vs. Actual)

As shown in the feature importance plot Fig. 5, stationary activities like laying, standing, and sitting have a negative correlation but activities involving physical activities like walking, walking up stairs and walking down stairs have a positive correlation to the HI. Note that for any person, overall time duration of physical activities is less compared to the overall time duration of stationary activities.

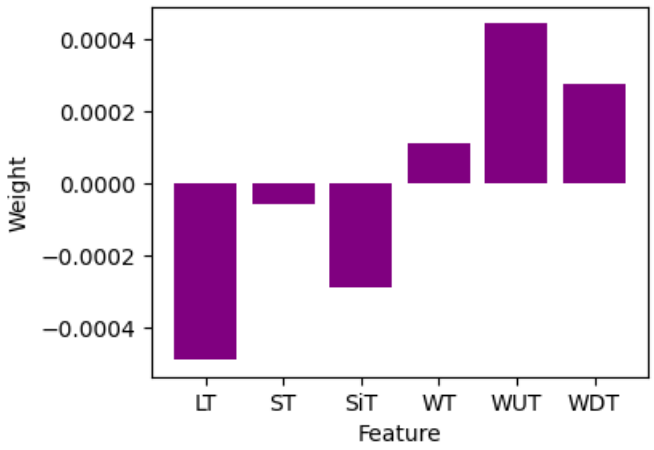

Fig. 4. Linear Regression - Feature Importance Plot

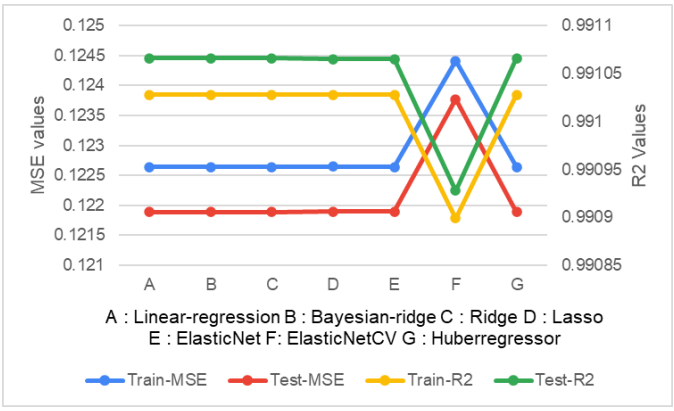

Fig. 5. Error comparison across models

V. CONCLUSIONS AND FUTURE WORK

In this project, we have predicted the health condition of a person based on the duration of each human activity performed in a day. Human activity is classified based on the sensor readings. HI is inferred from duration of each activity using regression model. Results of the prediction are promising and this model can be extended further to include new features like eating habits, smoking and drinking habits. Including more of these features make the prediction much closer to the reality. We can also add features like heart rate monitor, BP monitor etc. to make prediction much more sophisticated.

REFERENCES

[1] https://www.mdpi.com/1424-8220/10/2/1154/htm
[2] https://www.kaggle.com/code/fahadmehfoooz/human-activity-recognition-with-neural-networks
[3] https://www.kaggle.com/datasets/arashnic/har-1
[4] https://www.kaggle.com/code/fahadmehfoooz/human-activity-recognition-with-neural-networks/input