ANALYSIS OF FEATURE SELECTION TECHNIQUE FOR HUMAN ACTIVITY RECOGNITION

A thesis is submitted to the faculty of Engineering and Technology, Jadavpur University

in the partial fulfilment of the requirements for the degree of

Master of Technology

in

Computer technology

Submitted by

Suvashree Basu

Roll Number: 001910504027

Registration Number: 149861 of 2019-2020

Under the guidance of

Dr. Chandreyee Chowdhury

Professor, Dept. of Computer Science and Engineering, Jadavpur University

Department of Computer Science and Engineering

Jadavpur University, Kolkata

Declaration of Originality and Compliance of Academic Ethics

I hereby declare that this thesis contains literature survey and original research work done by me as part of my Master of Computer Technology course.

All information in this document have been obtained and presented in accordance with academic rules and ethical conduct.

I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name: **Suvashree Basu**

Class Roll No: **001910504027**

Examination Roll No: **M6TCT22028**

Thesis Title: *Analysis of Feature Selection Technique for Human Activity Recognition*

Signature:

Date:

Department of Computer Science and Engineering

Jadavpur University

To whom it may concern

This is to certify that Suvashree Basu, registration number 149861 of 2019-2020, class roll number 001910504027, a student of department of computer science and engineering, Jadavpur University has done a thesis under my supervision, titled *"Analysis of feature selection technique for human activity recognition"*. The thesis is approved for submission towards partial fulfilment of the requirements for the degree of Master of Technology in computer Technology, Jadavpur University for the session of 2021-2022.

_____

Prof. Chandreyee Chowdhury

Associate Professor,

Department of Computer Science and Engineering,

Jadavpur University

COUNTERSIGNED BY

_____

Prof. Anupam Sinha

Head of the Department of Computer Science and Engineering,

Jadavpur University

COUNTERSIGNED BY

_____

Prof. Chandan Mazumdar

Dean, Faculty of Engineering and Technology,

Jadavpur University

Certificate of Approval

(Only in case of thesis is approved)

The thesis is instance is hereby approved as a creditable study of an engineering subject carried out and presented in a manner satisfactory to warrant its acceptance as a perquisite to the degree for which it has submitted. It is understood that by this approval the undersigned do not necessarily endorse or approve any statement made, opinion expressed or conclusion drawn therein, but approve this thesis only for the purpose for which it is submitted.

_____ _____

Signature of Examiner Signature of Examiner

Date: Date:

ABSTRACT

The process of Human Activity Recognition using mobile phones is quite complicated, with many extracted features, some of which are redundant. Removing redundant features not only reduces the size of the dataset but also saves time. As a result, our key study aimed to identify the most effective and important features. We propose a noble feature selection technique using Filter method and Wrapper method. On the UCI-HAR dataset, we present a new feature selection approach. For classification, we use the Support Vector Machine, Logistic Regression, Random Forest, K-Nearest Neighbour and Decision Tree classifier in original dataset. The goal is to assess each classifier's performance with a reduced feature set and examine the impact of feature selection on model performance. It is observed that SVM and Random Forest show significant gain in accuracy with the reduced feature set.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER ONE

1.INTRODUCTION

Human Activity Recognition (HAR) is a large dimensional machine-learning problem. It can be used to keep track of a person's daily activities. It can also be used for a variety of other applications, such as health care, security, human survey systems, and so on. Human beings involve a variety of activities in their daily lives, such as walking, typing, eating, sitting, standing, jogging and so on. Human activity recognition (HAR) is an interesting but demanding research area. It proposes to use technical analysis to determine a person's activity patterns or categories [1]. This is important for a variety of applications, including long-term healthcare monitoring [2], active and assisted living systems [3], and smart homes [4]. etc. Recognition of human activities is basically a supervised classification problem. The data gathered by the wearable sensors is evaluated and then mapped to a set of activities such as walking, sitting, standing, laying etc. Although a motion sensor may get one's body's local inertial information, the signals are abstract and unintelligible, and there is a gap between the raw data and the activity being performed. Due to the HAR dataset contains numerous complicated signal data and a range of extracted features, it is more practicable and important to remove the irrelevant or redundant features. This research focuses on feature selection in order to extract usable features from the UCI-HAR dataset. However, as a growing research area in healthcare, HAR using smart phones' built-in sensors is attracting the attention of many academics.

In general, the HAR process consists of multiple steps, beginning with the collection of information about human behaviour from raw sensor data and ending with a conclusion on the currently performed activity. These are the steps as follows:

(1) Data collection – Collecting the raw data from sensor device.

(2) Data pre-processing – Pre-processing the raw data from sensor streams, removing noise and redundancy, and performing data aggregation and normalization.

(3) Segmentation – In this phase, the pre-processed data is segmented in windows of particular size, for temporal pattern analysis.

(4) Feature extraction – Extracting the main characteristics of features (i.e., temporal and spatial information) from the segmented data using, for example, statistical moments.

(5) Feature Selection – It is a technique to choose a subset of original feature following a well-defined evaluation criterion, that removes irrelevant and redundant features from the dataset.

(6) Classification – Using different machine classifier determining the given activity.

Figure 1.1: Framework of Human Activity Recognition

Feature selection involves choosing a subset of features from the original features in order to reduce model complexity, improve the computational efficiency of the models generated by noise by irrelevant features. A subset of features from the initial collection of characteristics that best represent the data are obtained through feature selection. Feature selection techniques are categorized into supervised, unsupervised, and semi-supervised models depending on the training data used (labelled, unlabelled, or partially labelled). The following feature selection techniques can be categorized according to their relationship with learning methods: Filter method, Wrapper method, Embedded method.

1.1 Application of HAR

1.1.1 Localization

Activity recognition on smart phones could increase context awareness and, it can be used in localization. One cause for using mobile sensors instead of GPS for location is that GPS signals are generally very low inside buildings and underground. Activity recognition techniques combined with mobile sensors; it could allow in determining the location. A similar approach is used for floor localization without the use of infrastructure. Other reason to using mobile sensors for localization is that GPS accuracy degrades inside cities surrounded by towering structures. In this case, GPS-based localization might distract between a movie theatre and a restaurant, which may be just a few steps apart.

1.1.2 Biometric Signature

The motion pattern of a subject is usually exclusive and especial. When people raise their hands, it's very impossible for two people's hands to have the same motion patterns. Because of the variances in motion-related bones and muscles on human bodies, even in a successful replication, differences still persist. Sensors like accelerometers can detect those variations. Human biometric signatures with patterns in motion/gestures may be solved using activity recognition techniques [5]. Pattern recognition technologies are utilised in these applications to obtain unique motion patterns, which are then kept in a database. Because of the widespread use of mobile devices, it is both handy and feasible.

1.1.3 Observation of Everyday Life

Applications for daily life monitoring are typically designed to serve as a handy reference for activity reporting or to aid in exercising and healthy lives. These gadgets have internal sensors such as an accelerometer, gyroscope, and GPS that track people's steps, stairs climbed, calories burned, hours slept, distance travelled, sleep quality, and other things. Users can review data tracking and visualization in reports using an online service.

1.1.4 Smart healthcare

The healthcare industry is in a poor situation. Healthcare is more expensive than it has ever been, the world population is ageing, and the number of chronic diseases is increasing. The approach is a world where basic healthcare would become available to most people and people would be less prone to chronic disease. An accurate identification of human activities could help us provide better patient recovery guidance, or an early alarm of emergency [6]. These are also useful for mentally challenged people where another person needs to be always with them. Human activity recognition is very much useful in our daily life, to monitoring the kids and elder people. If there is anything happen certainly then immediate actions can be taken. People who live in remote areas, it is very hard to provide them a proper regular treatment in that case smart healthcare using Smartphone act an essential role.

1.2 Motivation

Literature witnesses several works [7], [8], [9] that highlights the importance of Feature Selection (FS) in a benchmark dataset that certainly improves the overall performance of the classifier by accelerating the algorithm, simplifying the model by reducing the complexities incorporated by irrelevant features in the dataset. It is observed that most of the features present in standard benchmark datasets do not contribute much to the classification task. Thus, it is important to come up with a nominal set of useful features especially for real time activity recognition. Further, reduction of dimensionality on dataset can also reduce the cost as well as improve performance on low powered devices like Smartphone or wearable sensors. Needless to mention that for real time human activity recognition, high quality features in both time domain (mean, median, variance) as well as frequency domain (FFT) are essential to deliver high classification accuracy with least computational complexities. Standard techniques like Filter methods use statistical

characteristics of the features to select appropriate features whereas the wrapper methods take the help of learning algorithms to evaluate the solutions at each iteration to find out the best combination of features. Recently, metaheuristics approaches like Genetic Algorithm (GA), have added a new dimension to the feature selection techniques as they can solve optimization problems [10], [11], [12]. GA has gained a lot of attention among researchers to be used for feature selection as it has the ability to reduce the probability to get trapped in local optima which is a major concern for optimization problems. Genetic algorithm proves to be a good optimization technique which iterates across different stages as selection of individuals, and applies genetic operators like crossover and mutation to obtain an optimal solution. This algorithm evaluates the most appropriate features by figuring out the optimum fitness of the population by selecting the most suitable as well as feasible off-springs from a given generation, and uses the genetic information of the same to reproduce the new optimal population of solution. The main motivation behind this work was to bring both statistical FS techniques as well as metaheuristic approaches under one umbrella, on the same benchmark dataset and to analyse their performances across traditional Machine Learning classifiers. It was observed that was able to select optimal features across several generations of a given population by exploration and exploitation of the entire search space of candidate solutions.

1.3 Contribution

As discussed, propose a feature selection algorithm for accelerometer data based HAR applying Genetic Algorithm in this paper may be summarised as follows:

- Selection of a benchmark dataset like UCI-HAR and applying traditional Machine Learning algorithms with all 561 features available in the dataset.
- Application of statistical feature selection techniques such as Filter Method and Wrapper Method on the benchmark dataset.

1.4 Thesis Outline

- A study of relevant works from the past is presented in Chapter 2.
- Steps for genetic algorithms are described in Chapter 3
- The proposed methodology is described in Chapter 4.
- The prediction algorithms, dataset details which are used in this work are described in Chapter 5
- The result and analysis of the techniques which are used in this study are summarized in Chapter 6.
- Conclusion which includes future work.

CHAPTER TWO

2. RELATED WORK

2.1 Framework Description

The activity recognition framework, unlike a classification model that relies on a set of ordered classes, arranges preset activities of interest according to the features of human activities. Due to the complexity of human activities and the various granularities in defining activities, research on activity detection is often focused on specific applications in practice, and it is essentially difficult to consider all activities. Therefore, to explore the effectiveness of the proposed method and encourage discussion, this research covers six common human activities (standing, sitting, lying, walking, walking upstairs, and walking downstairs).

Many researchers have proposed HAR for the recognition of human activities. They cover a wide range of applications, including smart homes, healthcare, security, and surveillance. HAR systems can sense, monitor, and learn from human activities, providing meaningful information that might help people make better decisions about their future requirements or behaviour. The following are the works that will be discussed and analyzed in this survey: Davide Anguita et al. [13] proposed a unique system based on smartphone sensors to track human physical activity. The population of the experiments is made up of 30 participants aged 19 to 48. Each participant did six different actions standing, lying, sitting, walking, walking upstairs, and walking downstairs, while wearing a Smartphone around their waist and using SVM as a classifier. In the experiment, a Samsung Galaxy S2 smartphone was used.

Human activities are often characterized by diversity, ambiguity, concurrency, and overlap due to the intrinsic nature of human behaviour. Some activities can lead the sensors employed in wearable sensor-based approaches to give very identical sensor signals, posing additional hurdles to the activity detection model.

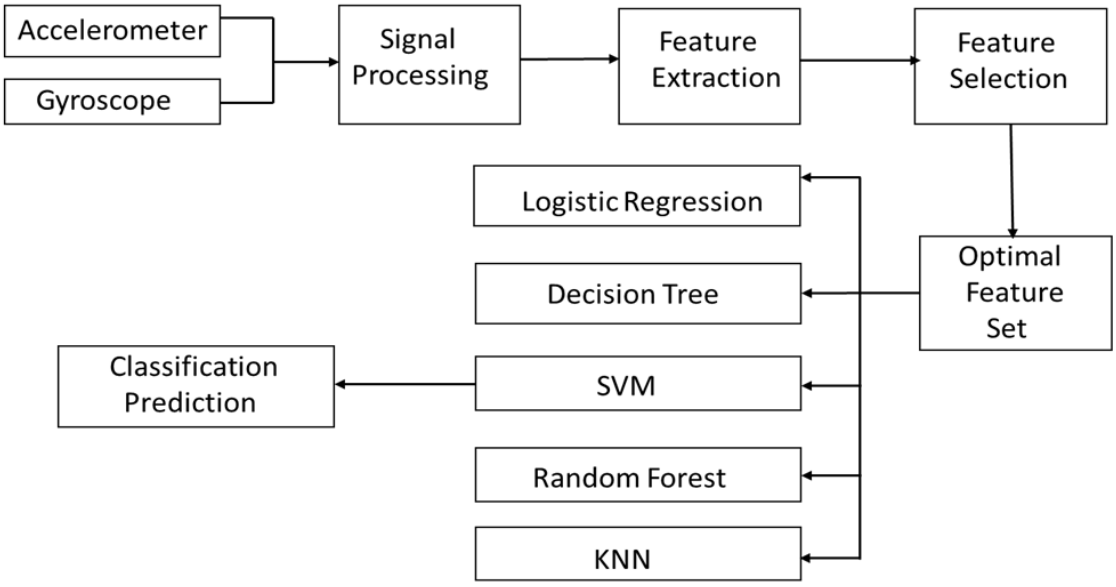

Figure 2.1: Framework description on Human Activity Recognition

2.2 Feature Extraction

Feature extraction is a component in the dimensionality reduction process, which divides and compresses a large set of raw data into smaller groupings. As a result, processing will be simpler. The fact that these huge data sets have a large number of variables is the most crucial feature. To process these variables, a large amount of processing power is required. So, by selecting and merging variables into features, feature extraction helps in selecting the best feature from those large data sets, effectively lowering the amount of data. These features are simple to use while still accurately and uniquely describing the real data set.

The major source of signal for detecting human activity is accelerometer and gyroscope data. The entire procedure is divided into three sections: (1) Data collection, (2) good feature extraction, and (3) classification.

In an activity recognition system, feature extraction is more important. The goal of feature extraction is to recover the common properties of acquired signals that belong to the same activity class [14]. The term "feature" refers to data that is generated using any method based on the sensor's raw signal (e.g., mean value of the acquired signal from the sensor). The difficulty of getting characteristics and the number of features needed for human activity recognition have a direct impact on the system's performance, which is especially important when employing smartphones. Feature extraction is the process of translating raw data into numerical features that may be handled while keeping the information in the original data set. It produces better outcomes than applying machine learning to raw data directly. Feature extraction can be done manually or automatically, as follows:

Identifying the characteristics that are significant for a specific problem, as well as designing a method to extract those features, are all part of manual feature extraction. In many cases, knowing the backdrop or domain can assist in making informed selections about which features can be valuable. Engineers and scientists have created feature extraction methods for images, signals, and text after decades of research. Using specialized algorithms, automated feature extraction extracts feature from signals without the need for human intervention. When you need to move quickly from raw data to constructing machine learning algorithms, this technique can be quite effective.

2.2.1 Feature Extraction Techniques

2.2.1.1 PCA:

PCA is a dimensionality reduction technique that detects relevant correlations in data, modifies existing data, and then quantifies the importance of these relationships. This definition can be divided into 4 steps to make it easier to remember:

i. Using a Covariance Matrix, determine the relationship between features.
ii. Get eigenvectors and eigenvalues by performing a linear transformation or eigen decomposition on the Covariance Matrix.
iii. Then, using Eigenvectors, convert the data into principal components.
iv. Finally, use Eigenvalues to quantify the importance of these associations and keep the most relevant primary components.

2.2.1.2 LDA

LDA is a machine learning classifier and supervised learning dimensionality reduction approach.

LDA tries to maximize the distance between each class's meanwhile minimizing the spread within the class. As a result, LDA uses within-class and between-class measures. This is an excellent choice because, when projecting data in a lower-dimensional space, maximizing the distance between the means of each class can lead to better classification results (thanks to the reduced overlap between the different classes).

When using LDA, it is assumed that the input data follows a Gaussian distribution (as in this case), hence utilizing LDA on data that isn't Gaussian could result in poor classification results.

2.3 A review Feature Selection Method

2.3.1 Filter method

These methods rely on data features to determine variables independently of machine learning algorithms. They are model agnostic as a result of this characteristic. They are less computationally expensive and faster. Features are evaluated based on a set of criteria that are not dependent on feature space. Then, based on the requirement, the highest-ranking

characteristics are chosen. Some examples of filter methods are mutual information, Anova, and Fisher Score.

2.3.1.1 Univariate Selection

A thorough understanding of feature selection and ranking can be quite beneficial to a machine learning practitioner. A solid understanding of these methodologies leads to higher-performing models, a better comprehension of the data's underlying structure and properties, and a better understanding of the algorithms that support in machine learning models. There are two reasons why feature selection is utilized in general:

- Decrease the number of features to reduce overfitting and increase model generalization.
- To learn more about the characteristics and how they relate to the response variables.

Univariate feature selection looks at each feature separately to see how strong of a relationship it has with the response variable. These methods are easy to execute and comprehend, and they are particularly useful for acquiring a deeper knowledge of data in general (but not necessarily for optimizing the feature set for better generalization). For univariate selection, there are numerous alternatives.

2.3.1.2 Pearson Correlation co-efficient

The Pearson correlation coefficient [15] is a key factor in determining similarity. It's the covariance estimated to the standard deviation. It has quite demanding data requirements [24]. The Euclidean distance (the distance between vectors) is commonly used to examine the similarity of vectors; however, it does not account for the differences in values between distinct variables.

To begin, the Pearson Correlation Coefficient formula treats human activity as a vector. Second, the weighted Pearson Correlation Coefficient method is used to calculate the degree of correlation between daily activity variables.

Finally, the relation degree between human activity features removes duplicate features.

Raw sensor data collection, pre-processing and segmentation, feature extraction and selection, classifier training, and data classification are all stages of activity recognition [16]. The human activity features that are extracted and selected determine the performance of activity recognition. However, there have been few studies on feature selection in human activities. We present a feature selection technique based on the Pearson Correlation Coefficient to identify the variables that are truly effective for recognizing users' activities.

The features of human activity are not necessarily related to one another. Most earlier techniques used all available temporal and spatial data. However, this can lead to the activity recognition model being overfitted. Overfitting occurs when a trained model performs well on training data rather than test data. Some learned features are relevant to each other and can suit training data for activity recognition. However, these essential traits may be too inflexible to accurately recognize test data, resulting in overfitting. Reducing significant features aids in the loss of the learned model and, as a result, the elimination of overfitting to some extent. To minimize overfitting, researchers recommend reducing the number of human activity characteristics using the Pearson Correlation Coefficient.

Correlation [17] is a statistical measure of the degree to which two or more quantitative variables are related linearly. The higher the correlation score, the higher degree of relationship. Its value ranges from -1 to 1, with 1 indicating perfect positive correlation or directly proportional correlation, 0 indicating no correlation, and -1 indicating perfectly negative correlation or inversely proportional correlation. The feature subset in which

independent variables are substantially linked with the target variable but not with each other is valuable. One variable can be predicted from the other using correlation. As a result, associated predictor variables are redundant and should be removed. Although correlated features may not always affect model accuracy, reducing dimensionality makes the model easier to understand. As a measure of correlation, we employ the Pearson correlation coefficient [18]. Independent variables should be substantially correlated with the target yet uncorrelated among themselves to construct good machine learning models.

The Pearson Correlation Coefficient can be calculated using the expression given in Eq. (1), which is used to evaluate the linear correlation between two variables X and Y. The covariance of X and Y is represented by the function COV (X, Y). σ_X and σ_Y are the deviations of X and Y, while μ_X and μ_Y are the respective means. $\rho_{X,Y}$ spans from +1 to 1. A value of +1 indicates that X is totally correlated to Y. A score of 0, implies that X is not correlated to Y. Finally, a score of -1 denotes that X is totally negatively linearly correlated to Y.

$$\rho_{X,Y} = \frac{\mathrm{COV}\,(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}$$

2.4 Feature Selection Survey on Human Activity Recognition

Due to its numerous uses in surveillance systems, rehabilitation centers, gaming, and other areas, human activity recognition has piqued the interest of the scientific community in the last two decades. This section covers the most recent research on the subject of sensor-based time-series data-based human activity recognition. A composite activity is made up of many atomic operations. Many existing strategies have focused on identifying simple and basic human movements, but recognizing composite activities remains a challenge.

The identification of high-level human activities, which are further composed of smaller atomic activities, is required for applications in daily life [19].

Human activity recognition has recently attracted a lot of attention because of the numerous possible applications in several research domains, such as home automation [20]. According to the literature review, numerous inertial sensor-based human activity recognition systems have been presented.

Bulling et al. in [21] focused on Human activity detection utilizing on-body inertial sensors, examined the challenges, and gave an overview of Human activity recognition methods in their paper. The Human Activity Recognition Chain was given as a broad framework for designing and evaluating Human Activity Recognition systems. Using a single waist-mounted triaxial accelerometer.

Gupta and Dallas in [17] created an accurate Human activity detection system that recognized six daily living activities and transitional events. They used feature selection algorithms to pick the best features from a set of traditional features as well as the ones they created.

Capela et al. [18] collected AR data from able-bodied, elderly, and stroke patients using the embedded smartphone accelerometer and gyroscope sensors. They generated 76 characteristics and chose subsets, which were then assessed using naive Bayes, support vector machine, and J48 decision tree classifiers. When compared to using the whole feature set, they found that using feature subsets resulted in improved or equivalent accuracies.

Bayat et al. [19] developed an AR system that recognises various common activities Using a single smartphone triaxial accelerometer. To measure recognition performance, they used a new set of features and different classifiers.

Ravi et al. [20] applied an accelerometer connected to an individual's pelvic region to determine eight activities (standing, walking, running, upstairs, downstairs, sitting, vacuuming, and brushing teeth). They then recommended the creation of a metalevel classifier, which outperformed base-level classifiers in terms of performance.

Bao et al. [21] used five biaxial accelerometers that were concurrently mounted to different parts of the human body (the right hip and four limb positions) to characterize twenty daily activities. They enlisted twenty volunteers, who were required to carry out pre-determined tasks in a normal manner in order to acquire experimental sensor data. After that, they used a sliding-window of fixed length to extract a range of time-domain and frequency-domain properties from the raw signals, and then built a classifier to recognize twenty activities.

CHAPTER THREE

3. GENETIC ALGORITHM

A genetic algorithm (GA) is a flexible optimization method. Figure 1 depicts the optimization process of GA, with mating and mutation as the two fundamental processes. Through mating, in which parameter values are transferred between parents to generate offspring, the GA combines the best of the previous generation. Some of the parameters are mutate [22]. The algorithm iterates until it converges, with the objective function judging the fitness of the new sets of parameters. The GA can explore the entire cost surface with these two operators to prevent falling into local minima. Simultaneously, it takes advantage of the best features of the previous generation to converge to increasingly better parameter sets.

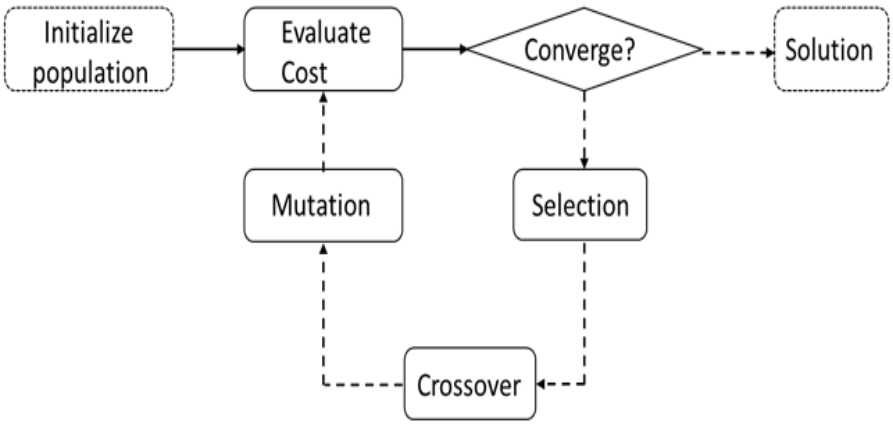

Figure 3.1: Flowchart of optimization with a Genetic Algorithm

3.1 Modelling of Chromosome

The idea behind evolutionary computation was that it could be utilised as an optimization tool and that solutions to problems could be evolved using natural selection operators. Genetic Algorithms (GA), a population-based algorithm. The purpose was to research the evolution process and create a framework that could be applied to a variety of applications [22]. A genetic algorithm (GA) is a heuristic search technique based on natural selection and genetics. The goal is to generate a solution to a problem by mimicking biological

processes such as survival of the fittest. GA is a strategy for developing chromosomal populations into new populations by combining selection with operations like crossover and mutation [22]. Each chromosome consists of genes. Crossover and mutation replicate biological processes that introduce variation to populations, while selection operators choose the fittest individuals from the population. Crossover and mutation are exploration processes, whereas selection is an exploitation process. each chromosome is assigned a fitness value/score, indicating how near the solution represented by the chromosome is to the expected result.

Figure 3.2: Feature Selection using Genetic algorithm

3.2 Population Initialization

There are two primary methods to initialize a population in a GA. They are −

Random Initialization − Populate the initial population with completely random solutions.

Heuristic initialization − Populate the initial population using a known heuristic for the problem.

3.3 Fitness Function and Selection criteria

A fitness function is a function that accepts as input a candidate solution to the problem and produces as output. Each chromosome is assigned a fitness value/score, indicating how near the solution represented by the chromosome is to the expected result. Return the best parents along with their value/score. Evaluates the individual solution for every generation to identify the fittest members. The better a chromosome fits, the more likely it is to be chosen. Selection of the best parents.

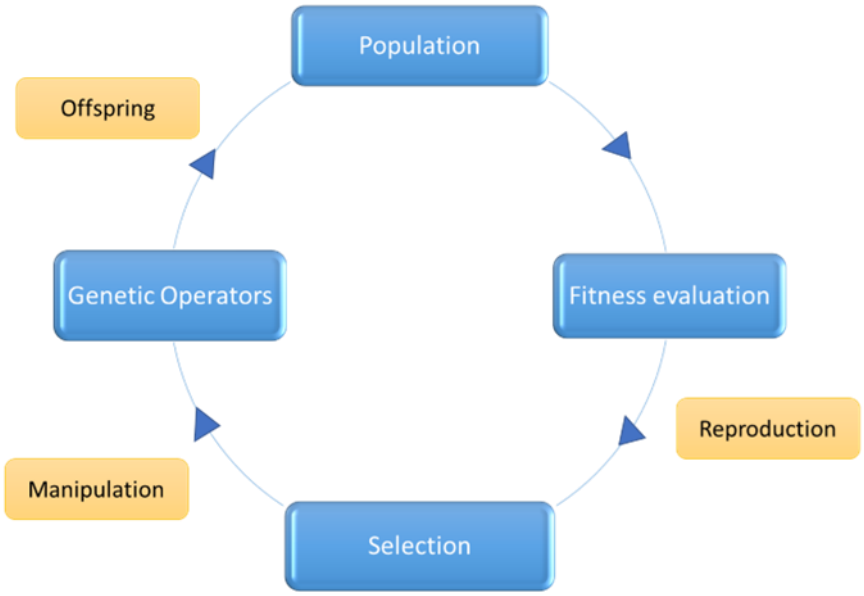

Figure 3.3: Steps involved in Genetic Algorithm

3.4 Introduction to Crossover

Reproduction and biological crossover are analogous to the crossover operator. More than one parent is chosen, and one or more offspring are generated utilising the parents' genetic material. Selects half of the first parent and half of the second parent.

3.4.1 Crossover Operators

We will discuss some of the most commonly utilised crossover operators in this section. These crossover operators are highly generic, and the GA Designer may decide to use a problem-specific crossover operator as well.

3.4.2 Single Point Crossover

In a single-point crossover, a crossover point is created at random, determining the moment at which parents share information to form children.

3.4.3 Multi Point Crossover

In a multi-point crossover, multiple crossover points are created at random, determining the points for information share between parents to form children.

3.5 Mutation

Random bits in the chromosome are flipped to form a new chromosome. Randomly flips selected bits from the crossover child.

3.6 Generations

Executes all the above function for the specified number of generations. Numbers of generations are five.

3.7 Implementation of GA for feature selection

First, run a function to initialize a random population. The randomized population is now run through the fitness function, which returns the best parent. Selection from these best parents will occur depending on the n-parent parameter. After doing the same, it will put through the crossover and mutation function respectively. Crossover is created by combining genes from the two fittest parents by randomly picking a part of the first parent and a part of the second parent. The mutation is achieved by randomly flipping selected

bits for the crossover child. A new generation is created by selecting the fittest parents from the previous generation and applying crossover and mutation. Here Boolean values are used (True represents that the feature has been selected, False represents that the feature has not been selected).

CHAPTER FOUR

4. PROPOSED METHODOLOGY

4.1 Overview

In this work, we have designed a feature selection method for sensor based HAR. The details are discussed in the following sections.

4.2 Modelling GA Based Feature Selection

Figure 4.2 shows the fundamental steps of the proposed GA approach as a flowchart below.

Chromosome Representation: The chromosomes are of size 'C', which is equivalent to the number of features are arranged as a sequence of gene in the chromosome. It is a Boolean list, where True indicates that the specified feature is selected and False indicates that it is not selected. The feature collection is thus represented by the chromosome. Figure 4.1 is a representation of a chromosome which represents that features (1, 4, 7, 8, …., 561) are selected by this chromosome for classification.

1	2	3	4	5	6	7	8	…….	561
True	False	False	True	False	False	True	True	….....	True

Figure 4-1: Representation of a Chromosome

Population: Over the generation, the population has a constant size, P. These are "P" sets of chromosomes. By randomly assigning trues and falses to various chromosome positions, the initial population is generated. In order to allow speedy convergence to the ideal solution, one chromosome with all trues, i.e., all features selected, was kept in the initial population pool. One of the essential genetic algorithmic parameters is the initial population size.

Selection: It specifies how chromosomes are chosen from the population pool. When a specific number of chromosomes from the current population pool are chosen to be included in the following generation, as well as when choosing chromosomes that will cross over to make new chromosomes for the following generation, we apply this selection process twice. The roulette wheel approach is used to choose chromosomes, giving more fit chromosomes (those with a higher fitness value) a larger chance of being chosen.

Crossover: By selecting chromosomes from the chromosome population pool of the prior generation, as described above in the section on "selection," probabilistically, we first establish a group of chromosomes for crossover. Then, to locate the new sets of chromosomes for the following generation, we choose a pair of chromosomes from this set and crossover them using the single point crossover procedure.

Mutation: We choose (M*P) number of chromosomes at random for mutation after getting the complete set of chromosomes for the following generation, where M is the mutation rate. We randomly reverse one of each chromosome's locations for each such picked chromosome, making it true if it is false and vice versa.

Stopping Criteria: GA needs a stopping condition because it is an iterative process. In this instance, GA continues until a stable set of features is obtained.

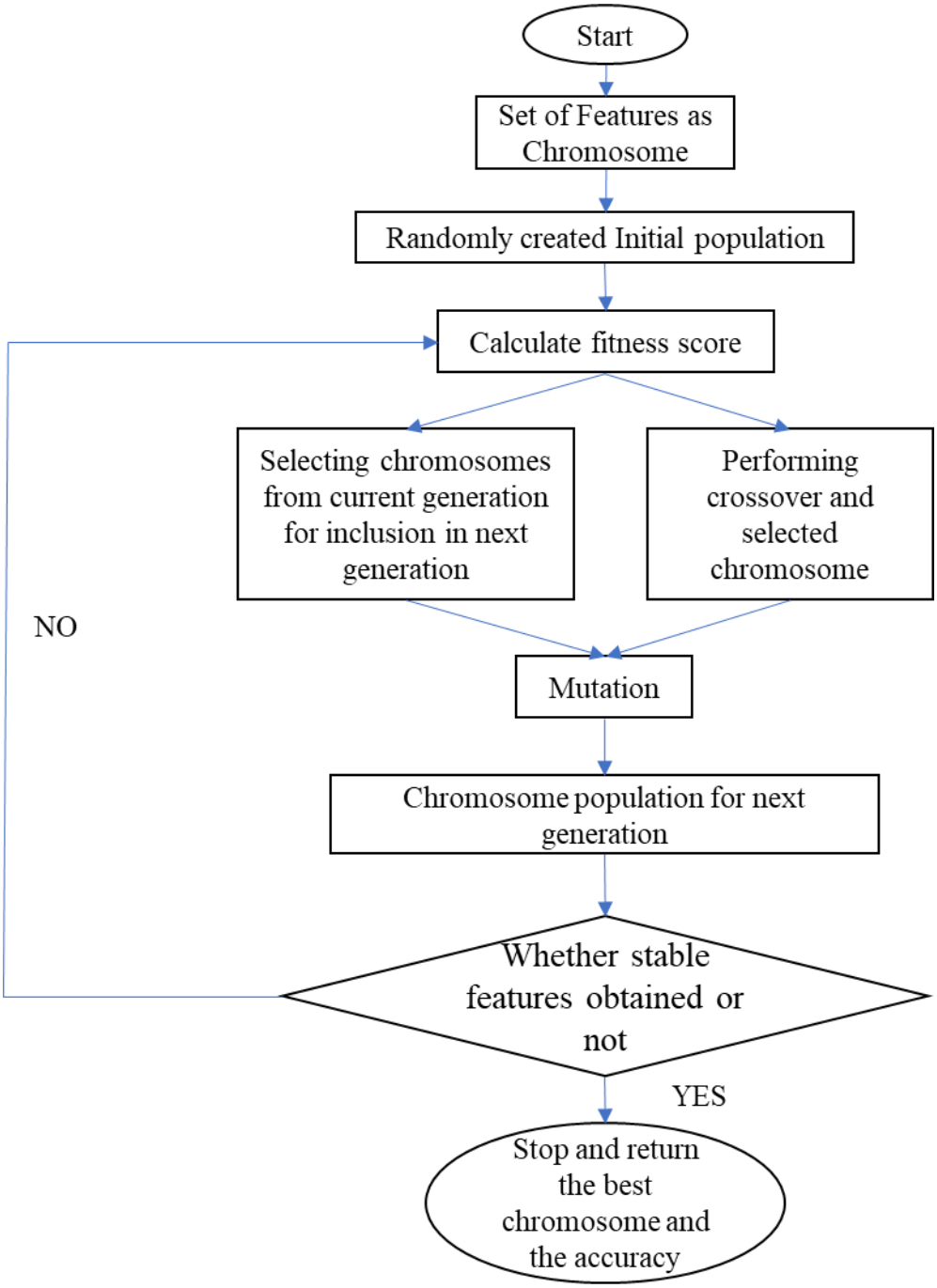

Figure 4.2: Flowchart outlining the steps used in GA

4.3 Fitness function

Different chromosomes are given a fitness value using the fitness function. The average accuracy achieved using the K-Neighbor Classifier (with K=5) is used in this work as the fitness value for each chromosome. The dataset has been split into 75% training and 25%

test set. Rearranging the dataset enhances both the prediction performance and model quality.

4.4 Roulette Wheel selection

In this strategy, the selection probability of a chromosome from the population pool is proportional to its fitness value**.** We can visualise a roulette wheel with various chromosomes filling various spaces according on their fitness scores. When the wheel stops revolving, we discover the chromosome to which the pointer is pointing and choose that chromosome for our mating pool. Currently, we revolve this wheel, which has a pointer attached in the centre. The better chromosome will take up more space on the roulette wheel, increasing its chance of being chosen over other chromosomes with lower fitness values. Roulette-wheel selection and single point crossover are used to choose the chromosomes that will make up the next generation. By using rank selection, where the selection of individuals is only on the basis of their population rank, roulette wheel selection is avoided.

4.5 Summary

This section discussed the suggested algorithm's framework. With respect to the human activity recognition, we examined the implementation of several genetic algorithm i.e., chromosome representation, population, selection criteria, crossover, and mutation.

CHAPTER FIVE

5. EXPERIMENTAL SETUP

5.1 Dataset Details

The UCI-HAR dataset [23] is used to analyse the performance of the proposed approach by performing extensive experiments. This dataset is accessible to the public for research work.

5.1.1 UCI-HAR Dataset

D. Anguita et al. (2013) [23] supplied this dataset. The overall number of instances is 10,299, and there are 561 features in total. This dataset will be used in our research to demonstrate our suggested strategy for deleting irrelevant or redundant features. In this dataset, the researchers looked at five everyday activities of thirty volunteers. The age range of the participants lies from nineteen to forty-eight (19–48). A smartphone worn around the waist provided the needed data on five common human actions. Walking, laying, sitting, standing, and walking stairs were among them (both upstairs and downstairs). Two smartphone sensors accelerometer and gyroscope were used to generate data. With a constant rate of 50 Hz, the accelerometer calculated the triaxial linear acceleration and the gyroscope calculated the triaxial angular velocity. With the help of recorded video, the activities were manually noted.

Activity	Number of Signal
Laying	1944
Sitting	1777
Standing	1906
Walking	1722
Walking Stairs	2950

Table 5.1: Label wise data representation

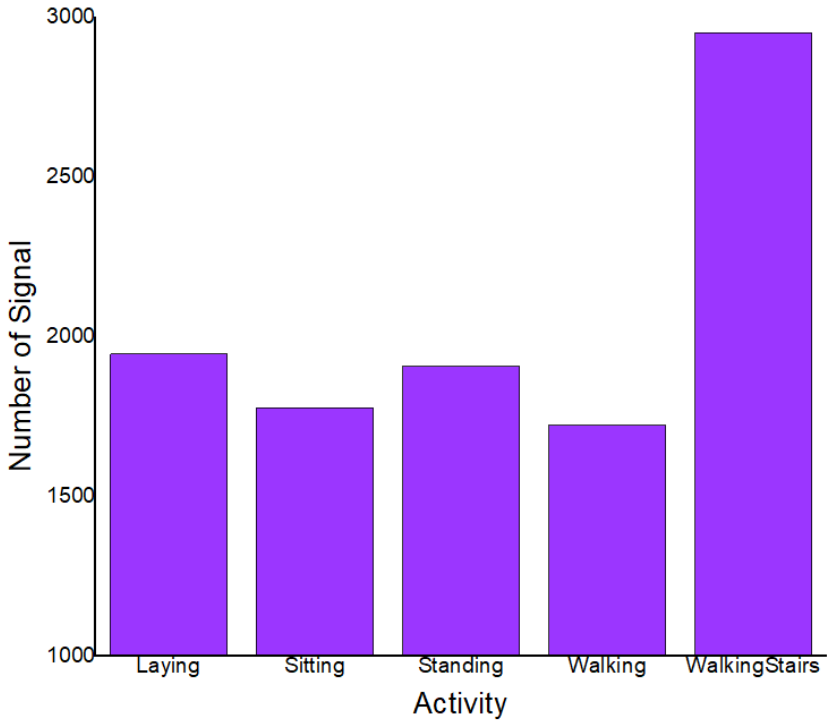

Figure 5.1: Graphical representation of label wise data representation

5.2 Software Tools

All experiments were run on a PC with 1.80 GHz Intel Core i5 -8265U CPU running Windows 10.

Code was written using Python v3.10.0. The following Python libraries were used in the implementation:

- Graph plots - matplotlib v3.5.2, OriginPro v9.0 32bit
- Classification algorithms - scikit-learn v1.1.1

5.3 Classification

5.3.1 Overview

In machine learning, if the predicted variable is of categorical type, it is a classification problem, and if the output value is of continuous type, it belongs to the regression task. To analyze data supervised learning is used. There are two types of Supervised learning. Those are Classification and Regression. Classification is a process of finding the class of new data instance and regression predicts the value of a variable in a specific circumstance. In this study, different classification models are used. Those are:

- K-Nearest neighbor
- Decision tree
- Random Forest
- Support vector machine
- Logistic regression

5.3.1.1 K-Nearest Neighbor

K-Nearest Neighbor (KNN) performs classification as well as regression. Here, the distance between the test instance and the training instances is calculated and then sorted in ascending order. Among all sorted distances, K distances are picked, and finally mode of

the K levels is returned. The algorithm searches through some noted distances. Here Euclidean, Manhattan, and Minkowski distance formulae [24] are used.

Euclidean distance:

$$D = \sqrt[2]{\sum_{i=1}^{K}(X_i - Y_i)^2}$$

Manhattan distance:

$$D = \sum_{i=1}^{K} | X_i - Y_i |$$

Minkowski distance:

$$D = \sqrt[q]{\sum_{i=1}^{K} |X_i - Y_i|^q}$$

Where X is the train set of data, and Y is the test set of data.

5.3.1.2 Decision tree

A Decision Tree (DT) is based on a tree data structure that is used to classify a new instance. A decision tree is made by dividing and conquering the dataset. Leaf nodes indicate the class of the instance. For different outcomes, an instance is partitioned. The class of the subsets depends on heuristics like the Gini index of diversity or information gain ratio. There is a probability of overfitting in the decision tree. To solve this problem pruning is applied to make the accuracy stable [25].

5.3.1.3 Random Forest

Random Forest (RF) was proposed by Breiman [26]. It adds a separate layer of randomness to bootstrap aggregation where the number of features is more. Random forest is a collection of tree classifiers. A vote is generated for each tree for the maximum occurred class to classify an instance a new data instance is passed to each sub tree and one class is

generated for each case, the class which gets the maximum vote is considered. Random forest reduces the pruning problem which is an advantage over the decision tree.

5.3.1.4 Support vector machine

Support Vector Machine (SVM) is a "supervised machine learning algorithm". The classification is implemented by finding the hyper-plane that differentiates the classes. The margin between classes is maximized using a standard Quadratic equation. When the data is not separated by a linear line, different types of the kernel can be chosen such as Gaussian and sigmoid, etc [27]. To decrease the computational cost kernel methods are used where the dimension is high. Support vector machines break the training data to reduce memory space, these points are called support vectors.

5.3.1.5 Logistic regression

Logistic regression (LR) is a statistical method to solve deterministic finite automata [28]. It works well when the number of attributes is more. Logistic regression tries to find out the cooperation of independent variables. To fit the model iteratively maximum likelihood is used. It tries to fit the train data in a logit function and predict test data according to that data. Logit function always lies between 0 and 1.

CHAPTER SIX

6. EXPERIMENTAL RESULTS AND DISCUSSION

6.1 Overview

This chapter includes the performance of GA followed by a comparison of the number of Features required by different labels of experimental setup for human activity. Finally, the accuracies of five generations are discussed.

6.2 Performance of GA on the dataset

Figure 6.1 shows the accuracy levels obtained over different number of generations. The accuracy associated with the best chromosome in the population is presented. In order for the GA to be stable, it was run for a sufficient number of generations. From Figure 6.1, we can see that after 80 generations the accuracies are fairly stabilized. Here, Random Forest classifier is used.

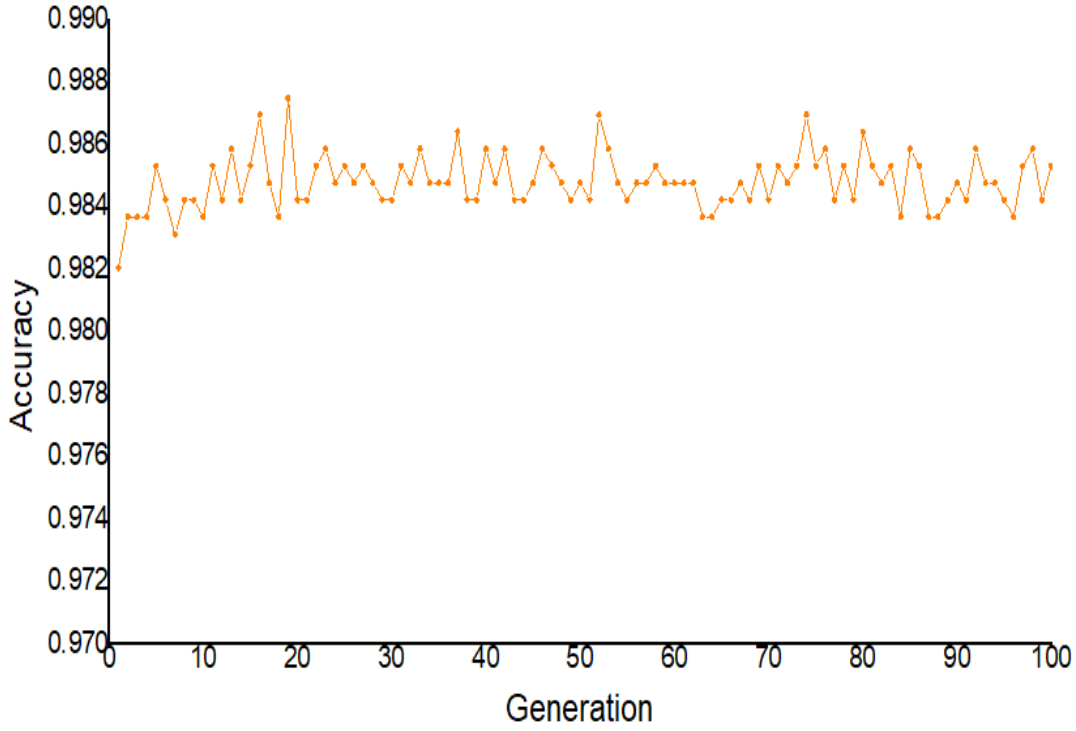

Figure 6.1: Accuracy graph of 100 generations using Random Forest Classifier

6.3 Comparing required number of Features for different labels of experimental setup

Depending on how the experimental set-up moves, a different number of features are needed for human activity. Here, we compare the number of attributes needed for various labels to support human activity. We utilized GA, which was limited by the maximum number of characteristics a given chromosome could have. For instance, if the GA was limited to 10 features, it would indicate that throughout the generations for which the GA was run, no chromosome would contain 10 or more characteristics; nonetheless, the number of features may be less than 10 or equal to 10. If a chromosome has more features than the predetermined number during cross over or mutation, some of the features will be eliminated at random from the chromosome until the number of features in the chromosome is less than or equal to the predefined number.

An experiment is conducted to investigate the role of the different collection of features on the classification of human activity and here, it can be observed from Figure 6.2 that with all features the proposed framework performed best however with Top 50 features the system performed reasonably well while the dimensionality of the problem becomes optimal. In this work, by selecting the best features based on univariate statistical tests, univariate feature selection has been used. It can be seen as a pre-processing step to an estimator. SelectKBest removes all but the K highest scoring features. f_classif used for classification purpose. To choose the top N features from a rank is a typical strategy for choosing features for a model The best score of features has been achieved by performing f_classif and Selectkbest function of univariate method. For ranking approaches, we used f_classif function of univariate method. In Table 6.1 it is seen that the accuracy of Top 50 features are best among others. The dataset has been split into 75% training and 25% test set.

Features	Accuracy
Top 10	0.568553
Top 20	0.607726
Top 50	0.930903
Top 80	0.872144
Top 100	0.541349
Top 150	0.835691
Top 200	0.799782
Top 300	0.807399
Top 400	0.699647
All Features	0.968444

Table 6.1: Accuracy of Features using Support Vector Machine

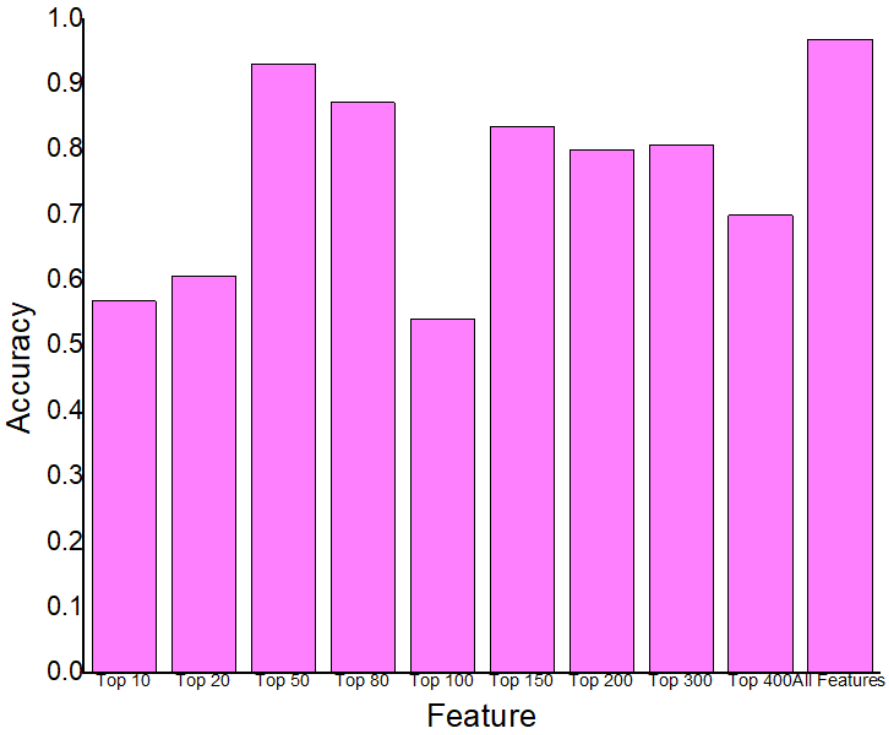

Figure 6.2: Graph of accuracy of the Features using SVM

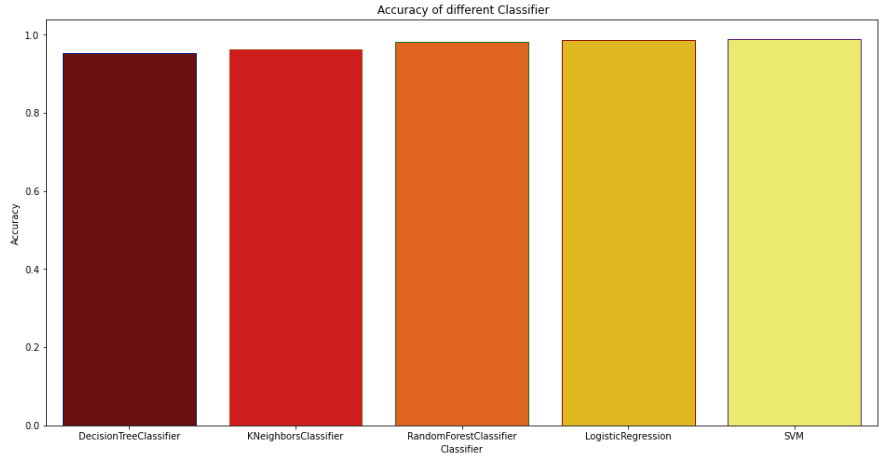

Figure 6.3: Graph of accuracy of the different classifiers

Figure 6.3 shows the graphical accuracy results for various classifiers, including Decision Tree, K-Nearest Neighbor, Logistic Regression, Random Forest & Support Vector Machine (SVM). With a score of 0.989663, SVM has the best accuracy rating, while Decision Tree has the lowest at 0.953210.

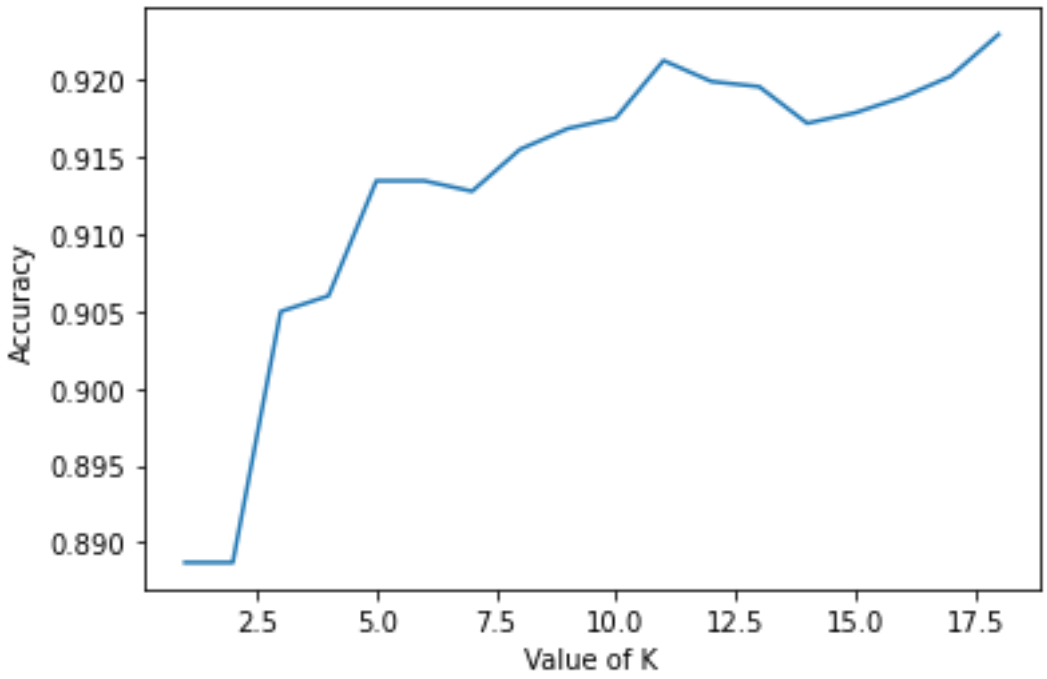

Figure 6.4: Graphical representation of accuracy for different K value

For this graph Figure 6.4, the K value has been tuned for a range of 0-19 by performing KNN classifier on UCI-HAR dataset. Best accuracy of 0.9229725 has been gotten when the K value is 18.

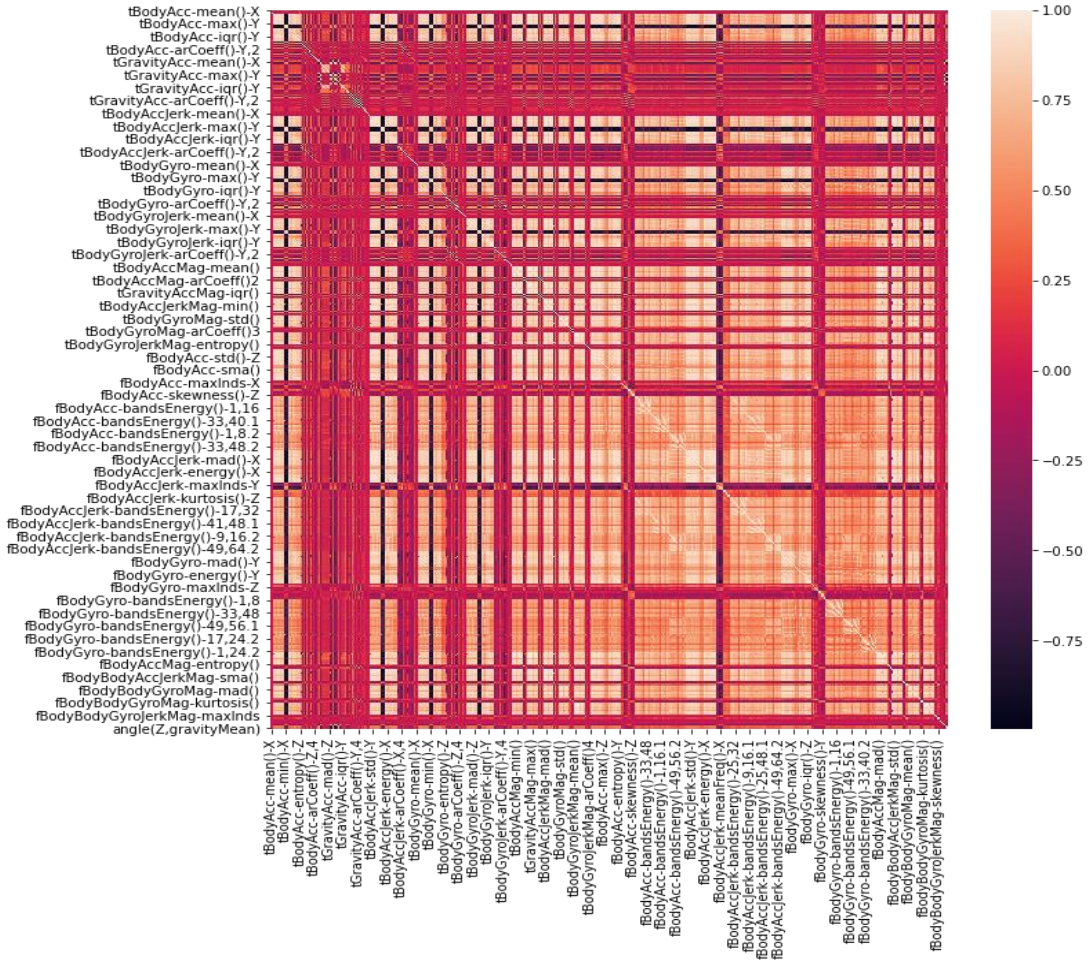

Figure 6.5: Pearson Correlation matrix with Heat map

Figure 6.5 shows a correlation heatmap for the UCI-HAR dataset is a graphical representation of a correlation matrix representing the correlation between different

variables. Correlation can have any value between -1 and 1. To find the correlation between several factors, including predictor and responder variables, the method corr() is used on the Pandas DataFrame. The heat map showing the correlation matrix which is produced using the Seaborn heatmap() function. From the above correlation heatmap, one could get some of the following information: Features such as tBodyAcc-std()-Y & tBodyAcc-std()-X, tBodyAcc-mad()-Y & tBodyAcc-std()-Y are having strong positive correlation. Generally speaking, a Pearson correlation coefficient value greater than 0.7 indicates the presence of collinearity [15]. Features such as angle(X,gravityMean) & angle(Y,gravityMean) is having strong negative correlations. There are several variables that have no correlation and whose correlation value is near 0. Uncorrelated features reveal different aspects of the inter class characteristics to the classifier. However, strongly correlated features can be reduced through feature selection approach as they contribute similar information to the classifier.

Classifier	**Generation 1**	**Generation 2**	**Generation 3**	**Generation 4**	**Generation 5**
Decision Tree	0.941240	0.945593	0.949583	0.950082	0.953210
KNN	0.967900	0.970893	0.970983	0.971844	0.972468
Random Forest	0.982046	0.982590	0.983134	0.983270	0.983351
Logistic Regression	0.979869	0.982046	0.982771	0.982590	0.982698
SVM	0.983134	0.984766	0.986398	0.986126	0.986180

Table 6.2: Accuracy of different classifier for different generations

For this experiment the accuracy of different generations of all the classifiers have been shown in Figure 6.7. Support Vector Machine, out of the five classifiers, has the best

accuracy value throughout all five generations, whereas Decision Tree has the lowest accuracy value. Here, 284 features were utilized to obtain accuracy.

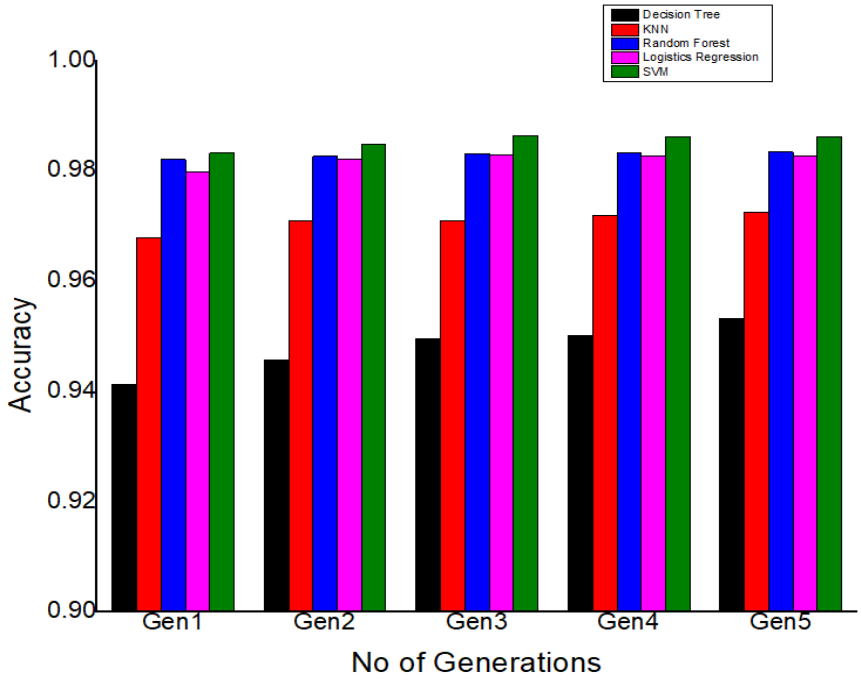

Figure 6.6: Accuracy graph of different classifier for different generations

CHAPTER SEVEN

7. CONCLUSION & FUTURE WORK

One of the main goals of this study is to improve with more accurate recognition of human activity. Despite the fact that HAR using mobile phone sensors is a complex process with a large data dimension, the suggested model can reduce the data dimension size by removing unnecessary and important features from the original dataset. We examined feature selection performance in the HAR domain, including feature selection techniques (i.e., Correlation coefficient, Genetic algorithm, Univariate selection). We studied several metrics on different types of HAR datasets using a large dataset and a suitable software framework that mediates between raw data classification. The feature selection techniques work for many classification methods (Support Vector Machine, K-Nearest Neighbor, Decision Tree, Logistic Regression, and Random Forest). We used a GA to select feature subsets from the complete set of features, by using classifier accuracy as the fitness function, to indicate that feature selection not only improves execution time but also increases classification accuracy. The results of using this reduced feature set for classification demonstrated that feature selection enhances the accuracy of classification in the case of SVM has got the highest accuracy value. GA improves classification performance, while SVM along with feature selection performs best for the classification. As a result, we are able to identify important features from human activity recognition data using this strategy.

This work may evolve the following methods as future work. The prediction can be applied to different datasets like Human Activity data and real-life Healthcare data. Here Filter, Wrapper, correlation-based feature selection methods are applied. To make it more accurate

Embedded methods can be applied. Furthermore, the suggested model might be applied to more complex activities to address other Machine Learning challenges and Human Activity Recognition by assessing it on other public activity dataset.

REFERENCES

[1] E. Kim, S. Helal, and D. Cook, “Human Activity Recognition and Pattern Discovery,” IEEE Pervasive Computing, vol. 9, no. 1, pp. 48–53, Jan. 2010, doi: 10.1109/MPRV.2010.7.

[2] V. Osmani, S. Balasubramaniam, and D. Botvich, “Human activity recognition in pervasive health-care: Supporting efficient remote collaboration,” Journal of Network and Computer Applications, vol. 31, no. 4, pp. 628–655, Nov. 2008, doi: 10.1016/j.jnca.2007.11.002.

[3] F. Al Machot, M. R. Elkobaisi, and K. Kyamakya, “Zero-Shot Human Activity Recognition Using Non-Visual Sensors,” Sensors, vol. 20, no. 3, Art. no. 3, Jan. 2020, doi: 10.3390/s20030825.

[4] S. Ramasamy Ramamurthy and N. Roy, “Recent trends in machine learning for human activity recognition—A survey,” WIREs Data Mining and Knowledge Discovery, vol. 8, no. 4, p. e1254, 2018, doi: 10.1002/widm.1254.

[5] J. T. Sunny and S. M. George, “Applications and Challenges of Human Activity Recognition using Sensors in a Smart Environment,” vol. 2, no. 04, p. 9.

[6] S. Purpura, V. Schwanda, K. Williams, W. Stubler, and P. Sengers, “Fit4life: the design of a persuasive technology promoting healthy behavior and ideal weight,” in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, New York, NY, USA, May 2011, pp. 423–432. doi: 10.1145/1978942.1979003.

[7] M. Zhang and A. Sawchuk, “A Feature Selection-Based Framework for Human Activity Recognition Using Wearable Multimodal Sensors,” presented at the 6th International ICST Conference on Body Area Networks, Beijing, People’s Republic of China, 2011. doi: 10.4108/icst.bodynets.2011.247018.

[8] J. Suto, S. Oniga, and P. P. Sitar, “Comparison of wrapper and filter feature selection algorithms on human activity recognition,” in 2016 6th International Conference on Computers Communications and Control (ICCCC), May 2016, pp. 124–129. doi: 10.1109/ICCCC.2016.7496749.

[9] M. Zubair, K. Song, and C. Yoon, “Human activity recognition using wearable accelerometer sensors,” in 2016 IEEE International Conference on Consumer Electronics-Asia (ICCE-Asia), Oct. 2016, pp. 1–5. doi: 10.1109/ICCE-Asia.2016.7804737.

[10] “User-Independent Activity Recognition via Three-Stage GA-Based Feature Selection - Theresia Ratih Dewi Saputri, Adil Mehmood Khan, Seok-Won Lee, 2014.” https://journals.sagepub.com/doi/full/10.1155/2014/706287 (accessed Jun. 23, 2022).

[11] A. Baldominos, P. Isasi, and Y. Saez, “Feature selection for physical activity recognition using genetic algorithms,” in 2017 IEEE Congress on Evolutionary Computation (CEC), Jun. 2017, pp. 2185–2192. doi: 10.1109/CEC.2017.7969569.

[12] A. M. Abo El-Maaty and A. G. Wassal, “Hybrid GA-PCA Feature Selection Approach for Inertial Human Activity Recognition,” in 2018 IEEE Symposium Series on Computational Intelligence (SSCI), Nov. 2018, pp. 1027–1032. doi: 10.1109/SSCI.2018.8628702.

[13] D. Anguita, A. Ghio, L. Oneto, X. Parra, and J. L. Reyes-Ortiz, “Human Activity Recognition on Smartphones Using a Multiclass Hardware-Friendly Support Vector Machine,” in Ambient Assisted Living and Home Care, Berlin, Heidelberg, 2012, pp. 216–223. doi: 10.1007/978-3-642-35395-6_30.

[14] A. Avci, S. Bosch, M. Marin-Perianu, R. Marin-Perianu, and P. Havinga, “Activity Recognition Using Inertial Sensing for Healthcare, Wellbeing and Sports Applications: A Survey,” in *23th International Conference on Architecture of Computing Systems 2010*, Feb. 2010, pp. 1–10.

[15] “Pearson Correlation Coefficient - an overview | ScienceDirect Topics.” https://www.sciencedirect.com/topics/social-sciences/pearson-correlation-coefficient (accessed Jun. 24, 2022).

[16] K. D. Feuz and D. J. Cook, “Collegial Activity Learning between Heterogeneous Sensors,” *Knowl Inf Syst*, vol. 53, no. 2, pp. 337–364, Nov. 2017, doi: 10.1007/s10115-017-1043-3.

[17] *Noise Reduction in Speech Processing*. Accessed: Jun. 24, 2022. [Online]. Available: https://link.springer.com/book/10.1007/978-3-642-00296-0

[18] M. Hall, “Correlation-Based Feature Selection for Machine Learning,” *Department of Computer Science*, vol. 19, Jun. 2000.

[19] AggarwalJ.K and RyooM.S, “Human activity analysis,” ACM Computing Surveys (CSUR), Apr. 2011, doi: 10.1145/1922649.1922653.

[20] S. T. M. Bourobou and Y. Yoo, “User Activity Recognition in Smart Homes Using Pattern Clustering Applied to Temporal ANN Algorithm,” Sensors (Basel), vol. 15, no. 5, pp. 11953–11971, May 2015, doi: 10.3390/s150511953.

[21] “A tutorial on human activity recognition using body-worn inertial sensors | ACM Computing Surveys.” https://dl.acm.org/doi/abs/10.1145/2499621 (accessed Jun. 23, 2022).

[17] P. Gupta and T. Dallas, “Feature Selection and Activity Recognition System Using a Single Triaxial Accelerometer,” IEEE Transactions on Biomedical Engineering, vol. 61, no. 6, pp. 1780–1786, Jun. 2014, doi: 10.1109/TBME.2014.2307069.

[18] N. A. Capela, E. D. Lemaire, and N. Baddour, “Feature selection for wearable smartphone-based human activity recognition with able bodied, elderly, and stroke patients,” PLoS One, vol. 10, no. 4, p. e0124414, 2015, doi: 10.1371/journal.pone.0124414.

[19] A. Bayat, M. Pomplun, and D. A. Tran, “A Study on Human Activity Recognition Using Accelerometer Data from Smartphones,” Procedia Computer Science, vol. 34, pp. 450–457, Jan. 2014, doi: 10.1016/j.procs.2014.07.009.

[20] N. Ravi, N. Dandekar, P. Mysore, and M. Littman, “Activity Recognition from Accelerometer Data.,” Jan. 2005, vol. 3, pp. 1541–1546.

[21] L. Bao and S. S. Intille, “Activity Recognition from User-Annotated Acceleration Data,” in *Pervasive Computing*, Berlin, Heidelberg, 2004, pp. 1–17. doi: 10.1007/978-3-540-24646-6_1.

[22] M. Melanie, “An Introduction to Genetic Algorithms,” p. 162.

[23] D. Anguita, A. Ghio, L. Oneto, X. Parra, and J. L. Reyes-Ortiz, “A Public Domain Dataset for Human Activity Recognition Using Smartphones,” *Computational Intelligence*, p. 6, 2013.

[24] M. Yesilbudak, S. Sagiroglu, and I. Colak, “A new approach to very short term wind speed prediction using k-nearest neighbor classification,” *Energy Conversion and Management*, vol. 69, pp. 77–86, May 2013, doi: 10.1016/j.enconman.2013.01.033.

[25] S. R. Safavian and D. Landgrebe, “A survey of decision tree classifier methodology,” *IEEE Trans. Syst., Man, Cybern.*, vol. 21, no. 3, pp. 660–674, Jun. 1991, doi: 10.1109/21.97458.

[26] R. L. Lawrence, S. D. Wood, and R. L. Sheley, “Mapping invasive plants using hyperspectral imagery and Breiman Cutler classifications (randomForest),” *Remote Sensing of Environment*, vol. 100, no. 3, pp. 356–362, Feb. 2006, doi: 10.1016/j.rse.2005.10.014.

[27] Q. Chang, Q. Chen, and X. Wang, “Scaling Gaussian RBF kernel width to improve SVM classification,” in *2005 International Conference on Neural Networks and Brain*, Oct. 2005, vol. 1, pp. 19–22. doi: 10.1109/ICNNB.2005.1614559.

[28] D. B. Springer, L. Tarassenko, and G. D. Clifford, “Logistic Regression-HSMM-Based Heart Sound Segmentation,” *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 4, pp. 822–832, Apr. 2016, doi: 10.1109/TBME.2015.2475278.