DEVELOPMENT OF FALL RISK CLUSTERING ALGORITHM IN OLDER PEOPLE

WONG KAM KANG

UNIVERSITI TUNKU ABDUL RAHMAN

DEVELOPMENT OF FALL RISK CLUSTERING ALGORITHM IN OLDER PEOPLE

WONG KAM KANG

A project report submitted in partial fulfilment of the requirements for the award of Bachelor of Engineering (Honours) Mechatronics Engineering

Lee Kong Chian Faculty of Engineering and Science
Universiti Tunku Abdul Rahman

September 2020

DECLARATION

I hereby declare that this project report is based on my original work except for citations and quotations which have been duly acknowledged. I also declare that it has not been previously and concurrently submitted for any other degree or award at UTAR or other institutions.

Signature :

Name : WONG KAM KANG

ID No. : 16UEB06589

Date : 12 September 2020

APPROVAL FOR SUBMISSION

I certify that this project report entitled **"DEVELOPMENT OF FALL RISK CLUSTERING ALGORITHM IN OLDER PEOPLE"** was prepared by **WONG KAM KANG** has met the required standard for submission in partial fulfilment of the requirements for the award of Bachelor of Engineering (Honours) Mechatronics Engineering at Universiti Tunku Abdul Rahman.

Approved by,

Signature :

Supervisor : DR. GOH CHOON HIAN

Date : 12 September 2020

Signature :

Co-Supervisor : IR. DR. CHUAH YEA DAT

Date : 12 September 2020

ACKNOWLEDGEMENTS

I would like to thank everyone who had contributed to the successful completion of this project. I would like to express my gratitude to my research supervisor, Dr. Goh Choon Hian and Ir. Dr. Chuah Yea Dat for their invaluable advice, guidance and enormous patience throughout the development of the research.

In addition, I would also like to express my gratitude to my loving parents and friends who had helped and given me encouragement.

ABSTRACT

Falls are serious problem which lead to negative consequences on the quality of life especially for older people. Most falls are caused by the interaction of multiple risk factors. However, manual analysis in big and complex medical data to analyse the fall risk factor are time consuming with high processing cost. Therefore, the aim of this study is to develop a clustering-based fall risk algorithm which can provide assistances for clinician in management of falls. The proposed algorithm consists of several stages, includes data pre-processing, feature selection, feature extraction, clustering and characteristic interpretation. This study employed Malaysian Elders Longitudinal Research (MELoR) dataset. A total of 1279 subjects and 9 variables from dataset (1411 subjects and 139 variables) are selected for clustering. t-Distributed Stochastic Neighbour Embedding (t-SNE) for feature extraction and K-means clustering algorithm achieved the highest performance in clustering, which grouping the subjects into Low (13%), Intermediate A (19%), Intermediate B (21%) and High (31%) fall risk group. In comparison, older people with higher fall risk have slower gait, imbalance, weaker muscle strength, with cardiovascular disorder, poorer performance in cognitive test, and advancing age. This is supported by the finding in literature review. To concluded, the proposed fall risk clustering algorithm is capable to group those subjects that have similar features. It presents a potential as assessment tool in management of falls.

TABLE OF CONTENTS

DECLARATION **iii**

APPROVAL FOR SUBMISSION **iv**

ACKNOWLEDGEMENTS **vi**

ABSTRACT **vii**

TABLE OF CONTENTS **viii**

LIST OF TABLES **xi**

LIST OF FIGURES **xii**

LIST OF SYMBOLS / ABBREVIATIONS **xiv**

LIST OF APPENDICES **xvi**

CHAPTER

1 INTRODUCTION 1

1.1 General Introduction 1

1.2 Importance of the Study 3

1.3 Problem Statement 3

1.4 Aim and Objectives 4

1.5 Scope and Limitation of the Study 4

1.6 Contribution of the Study 4

1.7 Outline of the Report 5

2 LITERATURE REVIEW 6

2.1 Introduction 6

2.2 Risk Factor 6

2.2.1 Gait and Balance 7

2.2.2 Muscle Strength 10

2.2.3 Cardiovascular Disorder 12

2.2.4 Cognitive Impairment 14

2.2.5 Demographic 16

2.2.6 Other Risk Factors 18

2.3 Current Fall Risk Predictors 20

2.4 Dimensionality Reduction 21

2.4.1 Feature Selection 21

2.4.2 Feature Extraction 23

2.5 Clustering Techniques 25

2.5.1 Partitional Clustering 26

2.5.2 Hierarchical Clustering 27

2.5.3 Density Based Clustering 29

2.5.4 Grid Based Clustering 30

2.5.5 Model Based Clustering 30

2.6 Clustering Validation Measures 31

2.7 Summary 32

3 METHODOLOGY AND WORK PLAN 33

3.1 Introduction 33

3.2 Equipment 33

3.3 Proposed Clustering Algorithm 33

3.3.1 Input Data 35

3.3.2 Data Pre-processing Methods 36

3.3.3 Feature Selection Methods 37

3.3.4 Feature Extraction Methods 39

3.3.5 Clustering Methods 41

3.3.6 Cluster Evaluation Methods 44

3.3.7 Characteristic Interpretation 45

3.4 Project Planning 45

3.5 Problems Encountered and Solutions 47

3.6 Summary 48

4 RESULTS AND DISCUSSION 49

4.1 Introduction 49

4.2 Data Pre-processing 49

4.2.1 Missing Data Handling 50

4.2.2 Feature Categorization 50

4.2.3 Normality Testing 51

4.3 Feature Selection 52

4.3.1 Hypothesis Testing 52

4.3.2 High Correlation Filter 53

4.3.3 Feature Importance 53

4.3.4 Discussion of Feature Selection Methods 54

4.4 Feature Extraction 55

4.4.1 Principal Component Analysis (PCA) 56

4.4.2 t-Distributed Stochastic Neighbour Embedding (t-SNE) 57

4.4.3 Discussion of Feature Extraction Methods 58

4.5 Clustering Algorithm 60

4.5.1 K-means Clustering 60

4.5.2 Hierarchical (Agglomerative) Clustering 61

4.5.3 Fuzzy C-mean Clustering 62

4.5.4 Discussion of Clustering Methods 63

4.6 Characteristic Interpretation 65

4.6.1 Timed Up and Go Test (TUG) 69

4.6.2 Functional Reach (FR) 69

4.6.3 Hand Grip Strength (HGS) 69

4.6.4 Cardiovascular Variability Ratio (Systolic Blood Pressure, Diastolic Blood Pressure and RR Interval) 70

4.6.5 Cognitive Assessment (MoCA questionnaire) 70

4.6.6 Demographic 70

4.6.7 Height, Body Mass Index and Waist to Hip Ratio 71

4.7 Summary 71

5 CONCLUSIONS AND RECOMMENDATIONS 72

5.1 Conclusions 72

5.2 Recommendations for Future Work 72

REFERENCES 74

APPENDICES 89

LIST OF TABLES

Table 4.1: Measurements Between Normally and Non-Normally Distributed Data. 52

Table 4.2: Number of Variables After Each Stage in Feature Selection. 55

Table 4.3: The Comparison of Performance Among Six Different Combination. 65

Table 4.4: Summary of Each Generated Cluster. 67

Table 4.5: Variables That Chosen for Characteristic Interpretation. 67

Table 4.6: Characteristic of Clustered Groups. 68

LIST OF FIGURES

Figure 2.1: Types of Clustering Techniques. 25

Figure 2.2: Types of Hierarchical Clustering. 28

Figure 3.1: Overview of Proposed Clustering Algorithm 34

Figure 3.2: Probability Density Function of T-test. 37

Figure 3.3: Steps Involved in PCA Feature Extraction. 40

Figure 3.4: Steps Involved in t-SNE Feature Extraction. 41

Figure 3.5: Flow Chart for K-means Clustering Algorithm. 42

Figure 3.6: Flow Chart for Agglomerative Clustering Algorithm. 43

Figure 3.7: Flow Chart for Fuzzy C-means Clustering Algorithm. 43

Figure 3.8: Gantt Chart. 46

Figure 3.9: FYP Project Timeline. 47

Figure 4.1: Summary of Data Set (1411 Subjects, 139 Variables). 49

Figure 4.2: Number and Category of Variables. 51

Figure 4.3: Heat Map for Correlation Test. 53

Figure 4.4: Feature Importance for Cardiovascular Variables in ‘RR’ Group (Left) and Selected Variables after Combining Cardiovascular and Common Variables (Right) 54

Figure 4.5: Summary of Data Set after Z-score standardization (1279 Subjects, 9 Variables). 56

Figure 4.6: The Variance Ratio Explained by Each Principal Components (Left) and Total Explained Variance by Number of Principal Components (Right). 57

Figure 4.7: 2D Representation of 1st and 2nd Principal Components. 57

Figure 4.8: 2D Representation for t-SNE with Perplexity 10 (Left) and 180 (Right). 58

Figure 4.9: Analysis and Results in K-means Clustering Algorithm. 61

Figure 4.10: Analysis and Results in Hierarchical (Agglomerative) Clustering Algorithm. 62

Figure 4.11: Analysis and Results in Fuzzy C-mean Clustering Algorithm. 63

Figure 4.12: Different Fall Risk Group. 67

LIST OF SYMBOLS / ABBREVIATIONS

$\bar{x}$	means
α	significance level
σ	standard deviation
%	percentage
cov	covariance
cm	centimetre
kg	kilogram
m	meter
s	second

ABC	Activities-specific balance confidence
AST	Active standing test
BMI	Body mass index
CO	Central obesity
DBP	Diastolic blood pressure
DBSCAN	Density-based spatial clustering of applications with noise
EF	Executive function
FCM	Fuzzy C-means
FR	Functional reach
FRID	Fall risk increasing drug
HFRMII	Hendrich II fall risk model
HGS	Hand grip strength
HUTT	Head-up tilt test
ID	Identity
LDA	Linear discriminant analysis
MCI	Mild cognitive impairment
MELoR	Malaysian Elders Longitudinal Research
MFS	Morse fall scale
MIS	Memory impairment screen
MMSE	Mini-mental state exam
MoCA	Montreal cognitive assessment
OH	Orthostatic hypotension

OR	Odd ratio
PCA	Principal component analysis
RR	RR interval
SBP	Systolic blood pressure
SOM	Self-organising map
STING	Statistical Information Grid-Based Clustering
STRATIFY	St. Thomas risk assessment tool
STS	Sit to stand
t-SNE	t-Distributed Stochastic Neighbour Embedding
TUG	Timed ‘Up & Go’ test

LIST OF APPENDICES

APPENDIX A: Clustering Algorithm (Python Codes) 89

CHAPTER 1

INTRODUCTION

1.1 General Introduction

Fall is described as an incident that unintentionally causes a person to come to rest on the ground or floor (World Health Organization, 2018). According to studies, about one third of people who older than sixty-five years old had fall at least once in the past twelve months (Sieri and Beretta, 2004; Stevens and Sogolow, 2005; Rubenstein and Josephson, 2006). In fact, falls are the second most common source of accidental or unintentional injury in the world. The injury caused by fall can associated with fractures, disability and even mortality (Pfortmueller et al., 2014). Thus, falls in older people are considered as a major public health issue (Gale, Cooper and Aihie Sayer, 2016).

In general, the cause of fall is complex, so it is difficult be analysed if only depend simple diagnosis results. To deal with this, fall interpretation by fall risk assessment was suggested. Fall risk is simply used to describe the possibility of falling (Horton, 2007). The most common alternative for fall risk assessment is implementation of multiple risk factors intervention. There are many studies have used risk factors to determine the fall risk for older people (Tromp et al., 2001; Whitney et al., 2012). Common factors including but not limited to, advanced age, muscle weakness, medications and gait imbalance. In addition, environment from house or hospital also considered as a factor which can directly affect the incidence of fall (Letts et al., 2010). In short, the possibility of fall can be linked with number of fall risk factors and their strength of association towards older people.

Based on the evaluated fall risk, fall prevention strategies can be proposed to create safer environment with reduced fall risk (Elliott, Painter and Hudson, 2009; Gillespie et al., 2009; Van Vost Moncada and Mire, 2017). An effective prevention programme should explore multiple risk factors and prevent fall from it. Consideration would probably be targeted only individuals at high risk of falling due to feasibility and cost effectiveness. However, it is still challenging to decide the major fall risk factors among older people because it can vary according different population and scenarios.

In recent years, machine learning and data mining are commonly applied in medical field (Polat and Güneş, 2007; Dağ et al., 2012; Tran et al., 2014). Data mining technique provides a user oriented approach to the novel and hidden pattern in various medical data (Nithya, Duraiswamy and Gomathy, 2013). This is because hand picking features which depend on expertise and experience may not that efficient for medical analysis. It cannot guarantee that all important information in the existing data are included. Moreover, it is time consuming and expensive if complex data are presented. Thus, data mining is useful to generate new information from large databases.

Machine learning techniques can be classified into supervised and unsupervised learning. Supervised methods like classification infers function from labelled training data. To illustrate this, the diseases symptoms are tagged with label and trained by classification algorithm. Eventually, this classification algorithm can identify the class of a patient based on symptoms (Saxena et al., 2017). Therefore, the supervised classification is used as a tool for prediction.

On the other hand, unsupervised learning is self-learning technique that without the labelling. Unsupervised clustering is used to find a structure in a collection of unlabelled data. The outcome of clustering is a data definition, where a cluster describes a set of objects that are identical to them and are separate from objects belonging to other clusters. In the medical field, cluster analysis offers a standardized, formalized approach for analysing data and identifying clinical similarity groups (Kalyani, 2012). Clustering techniques are typically more demanding than supervised solutions because it offer greater insight into complex medical results (Khalid and Prieto-Alhambra, 2019). Learning from data can help to know the disease evolution and personalise treatments according the need (Álvarez et al., 2019).

In this study, a clustering algorithm model is proposed for fall risk clustering in older people. The core idea is using the clustering strength to discover the major risk factors and characteristics for falls in older cohort. From fall dataset, the clustering algorithm should be able to cluster those subjects with similar characteristic into same group. All clustered groups are identified with different fall risk. By conducting analysis, the association of the risk factors and fall risk can be revealed.

1.2 Importance of the Study

Falls are serious problem which lead to negative consequences on the quality of life especially for older people. To illustrate this, falls and consequent injuries are major public health problems that often require medical attention. Older people constitute a large and increasing proportion of the population. The cost arising from falls represent a large proportion of healthcare spending (World Health Organization, 2007). Direct cost encompasses health care such as medication and adequate services. Indirect cost are productivity losses and disability caused by fall-related injuries. These economic impacts of falls are critical to family and society. Therefore, identify of relevant fall risk factors to prevent fall is of major importance for community (Todd and Skelton, 2004).

In fact, fall data analytics can effectively reduce the cost due to falls in hospital (Bill, 2007). The benefit of big data is the ability to look at thousands of factors at the same time, including those seemingly 'extrinsic' to the problem at hand. Nevertheless, it is not convenient to test such big data manually. With the development of clustering algorithm, a huge number of independent risk factors can be efficiently evaluated. It can help clinicians deal with the abundance of knowledge and improve the accuracy of diagnosis. Clustering observations can be used to examine the correlation or independence of features to offer a deeper insight. All these advantages lead the importance to develop a clustering models for fall risk assessment in older people. Besides that, this clustering algorithm may not be limited to fall risk assessment but also usable for other similar or related problems. All the findings in this study can provide valuable insight for future research.

1.3 Problem Statement

Falling among the older people is not a new problem, but one of the most complicated and high-cost unresolved problems concerned by the healthcare system. Problem statement for the current study is summarised as below:

- Although there are numerous studies have investigated the fall risk factors, it is still challenging to identify which group of factors contribute higher risk in older people and should specifically prevent from fall.

- Big and complex medical data analysis are time consuming with high processing cost if conducted manually.
- Although various machine learning techniques have applied in medical field, it is still lack of research and algorithm that specifically for the domain of fall risk clustering.

1.4 Aim and Objectives

The main aim of this study was to propose a clustering-based fall risk algorithm as assessment tool which can provide assistances for clinician in management of falls. The specific objectives of this research were:

- To identify the major risk factors for falls in older cohort.
- To identify dimensionality reduction and clustering techniques that can efficiently partition objects into number of clusters from a large dataset.
- To study the characteristics between higher and lower fall risk group.

1.5 Scope and Limitation of the Study

This study provides literature review on major risk factors of falls in older people. Apart from that, this study is focus on developing working algorithms that able to perform clustering in dataset. A final thesis included the development, flow, and performance of the algorithm are documented.

The study is limited to only Malaysian Elders Longitudinal Research (MELoR) dataset. Besides that, the number of faller and non-faller subjects are not balance in this dataset. Therefore, it may cause some bias in analysis. In addition, not all the potential fall risk factors are available in this dataset. The association of fall and risk factors are analyzed only for those provided in dataset.

1.6 Contribution of the Study

This study makes the following contributions:

- Provide summary of major fall risk factors from literature review of recent twenty years.
- Provide clustering algorithm that potential as assessment tool in fall analysis. It simplifies the process of data analysis to discover useful information if there exist.

- Clustering analysis in MELoR dataset. The characteristics that contribute higher fall risk are analysed and discussed.

1.7 Outline of the Report

This report covers a total of five chapters. Chapter 1 discusses the Introduction which consisted of background study, the importance of study, problem statement, aim and objectives, the limitation and scope and contribution of study.

Chapter 2 is about the Literature Review. It is conducted on fall risk factors, current fall risk predictors, dimensionally reduction techniques and clustering techniques.

Chapter 3 describes the Methodology in this study. The proposed clustering algorithm is explained from initial phase to final phase. The methods applied in data pre-processing, feature selection, feature extraction, clustering and characteristic interpretation are illustrated. The Gantt chart and milestone are also included. The problems encountered and solutions are discussed at the end of this chapter.

Chapter 4 includes the results and discussion. The results that generated from each phase are described follow by discussion. Tables and figures are provided to illustrate the findings.

Chapter 5 discusses the conclusion and recommendations for future works.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

In this chapter, literature review that covering falls in older people and clustering techniques will be conducted. All relevant theories, statements and gaps in existing research are identified through this. In order to ensure the information obtained are up to date, only relevant studies in recent twenty years were included.

This chapter can be generally divided into two parts. The first part was discussing about fall risk factor in older people. Major fall risk factors followed by assessment tools were identified from existing researches. Besides that, current fall risk assessments that widely used in hospital were illustrated. The second part was focus on dimensional reduction and clustering techniques. Different methods that are useful in dimensionality reduction were reviewed. After this, various clustering approaches as well as clustering validation methods were also explored.

2.2 Risk Factor

Fall risk factor is condition or characteristic that increase the likelihood of fall. Absolutely, there are various factors that cause older people to fall. It can further classify into intrinsic or extrinsic factors. Intrinsic factors include individual characteristics such as demographic, fall history and health status. On the other hand, extrinsic factors refer environmental factors which cause slipping or loss of balance. In general, fall may occur as the result of independent or complex interaction among fall risk factors. It can be confirmed that fall risk will greater with the increase number of risk factors. However, the correlations between each risk factor and fall is different. Therefore, it is necessary to identify the relationship between each risk factor and fall in older people.

2.2.1 Gait and Balance

Gait is a person’s manner of walking whereas balance is an even distribution of body weight. In this case, gait and balance disorders are identified as major fall risk factor in older people, lead to serious injuries and even mortality.

Normal walking among human are achieved by two legs that provide both support and balance. Gait cycle is used to describe the unique and repeatable motion that used for gait analysis. It can be illustrated from two major phases which are stance phase and swing phase. Stance phase start from heel strike then pass through list of motions and end with terminal swing in swing phase (Lakany, 2008). The movement in each of the leg and body are varying in different phases. It is important to ensure continual interchange between each phase because it enables balance while walking. If there is disorder in any segment of body or alter timing of muscle action, it may cause abnormal gait pattern and loss of balance. Thus, fall happen as consequence.

2.2.1.1 Current Researches Findings

Several studies had proved that gait and balance disorder were significant risk factor for fall in older cohort. To illustrate this, Rubenstein and Josephson (2006) provided the important insight that the risk for fall was nearly threefold increase for older people who had gait and balance impairments. Although the study has examined among multiple risk factors, gait and balance deficit still contributes higher relative risk ratio compared to others. It shown that this risk factor able to predict possibility of future fall with more consistency and precise. To support such statement, Ganz DA, Bao Y and Shekelle PG (2007) also reported that presence of gait or balance abnormalities increased risk of fall (1.4 to 2.6 odd ratio (OR) range) after conducted fifteen studies with relevant information. Besides that, ten out of fifteen studies have reported statistically significant results on this. All these evidences are indicating that gait and balance is an acknowledgement risk factor. Therefore, evaluation on this risk factor is essential step to identify fall risk.

There are two studies have analysed the diseases that associated with gait and balance disorder (Duxbury, 2000; Salzman, 2011). Duxbury (2000) reported that the linkage of gait disorders with either diseases in musculoskeletal, cardiovascular and nervous system. In like manner, Salzman (2011) listed down various medical conditions associated in a table. It is noteworthy that musculoskeletal, cardiovascular and neurological disorders are included in both studies. This indicates the gait disturbance is likelihood caused by combinations of one or more diseases under these three categories. Pain, imbalance, restricted range of motion and poor posture may be induced by these diseases to gait and balance.

As supportive study, Chaiwanichsiri, Janchai and Tantisiriwat (2009) reported that foot pain (OR = 2.5) and knee osteoarthritis (OR = 3.2) in foot musculoskeletal disorder were identified as fall risk factors. However, foot pain is found that has only little effect on gait in this study. This may be limited to differences in population in the settings. Moreover, Sinaki et al. (2005) concluded that women who had osteoporosis with hyperkyphosis resulted in slower gait and poorer balance will increased the risk of fall.

In summary, gait and balance is the major fall risk factor and it related to several diseases. However, the effect of some diseases on gait can be obvious while some are hard to identify and often discovered only after fall. Therefore, early detection on disease that cause gait and balance disorder is crucial for fall prevention.

2.2.1.2 Assessment Tools

The simplest way to identify the abnormal gait is through clinical evaluation for common patterns of abnormality. Two studies have listed common gait disorder patterns, associated characteristics and possible causes for each type of gait (Duxbury, 2000; Salzman, 2011). This assessment method does not involve complicate setup, but it required understanding mechanisms of each gait patterns and such characteristics may varies from person to person. In this case, development and use of tools for gait disorder assessment can provide more reliable results.

Among of screening tools, Timed ‘Up & Go’ test (TUG) is widely used. It measures total time taken (second) to rise from chair, walked three meters with usual gait speed, turned around, back to chair and sit. Cut-off times based on category of testers are used to evaluate functional mobility. Another common tool is Activities-specific Balance Confidence (ABC). ABC is 16 items scale which testers rate their own confidence interval when performing daily living activities. This rating scale ranges from zero (no confidence) to hundred (complete confidence) and overall score is obtained from average score of all sixteen items. Functional reach (FR) is measure of distance for maximal forward reach exceed arm's length with fixed base of support maintaining. A threshold distance is defined, and the tester will be predicted has low balance if cannot exceed it. Other tests as Dynamic Gait Index (DGI) appeared low sensitivity to risk indication and thus not be reviewed (Wrisley et al., 2003).

Several studies have proved TUG as a reliable measure to identify between fallers and non-fallers (Shumway-Cook et al., 2000; Lin et al., 2004; Alexandre et al., 2012). Shumway-Cook et al. (2000) reported that 13.5 second as threshold value had prediction rate of 90 % of faller classifying. This threshold value is not consistent with Alexandre et.al (2012), who found that predictive value of 12.47 second. Due to time of published, the latter result is more persuasive. However, this threshold value is likely to vary in different age group because the gait speed decrease with advanced age. Instead of support TUG as valid tool, Lindsay, James and Kippen (2004) stated TUG was poor in assessing fall risk but it may not that reliable because the data only collected from medical records of 160 patients (mean age = 81). Another study reports on accuracy of TUG rely on individual's ability to complete the test instead of TUG time (Large et al., 2006). Also, Schoene et al. (2013) concluded that TUG was more useful in frailty group instead of examined healthy old population. Nevertheless, TUG still the popular assessment tool in gait and balance.

There are two studies have reported that ABC scores related to fall (Hatch et al, 2003; Huang and Wang's, 2009). On the other hand, Hotchkiss et al. (2004) showed the ABC scale had no ability to identify people who had falling history. After systematic review, Stasny et al. (2011) concluded that there were insufficient researches and evidences to prove ABC scale can predicted falls. Apart from that, the ABC-6 which consists of 6 chosen activities in ABC only indicates stronger relationship to falls and useful in assessment of fear of falling (Peretz et al., 2006; Schepens, Goldberg and Wallace, 2010).

Johnsson, Henriksson and Hirschfeld (2003) found that FR is weak in stability measure as it may influenced by movement of trunk during testing. In contrast, one study found that FR has high reliability in balance measurement (Lin et al., 2004). Overall, FR still valid as a simple balance assessment.

2.2.2 Muscle Strength

Skeletal muscles assist in body support and movement through muscle contraction and relaxation. Muscle strength determines the amount of force that used to maintain balance. Weak muscle strength can cause imbalance of body due to insufficient force support. As muscle strength and mass reduced with age, the gait and functional performance are also be affected. Thus, muscle strength is evaluated as major risk factor for falls in older people.

2.2.2.1 Current Researches Findings

Many studies have reported muscle weakness is associated with fall in older adults. To illustrate this, Ding and Yang (2016) suggested older people with knee muscle strength around 1.05 to 1.10 Nm/kg were susceptible to high fall risk. Muscle weakness especially lower knee joint muscle can lead to slip-related fall. The odd ratio (OR) between lower extremity weakness and fall risk reported from 1.2 (Tromp et al., 2001) to 4.4 (Rubenstein and Josephson, 2006). Besides that, Moreland et al. (2004) also indicated lower extremity weakness was significant fall risk factor after constructed meta-analysis in thirteen studies.

During walking, the most activated muscle group that control body's anteroposterior equilibrium is plantar flexors. Other than that, knee extensors provide stability by support the weight of body. Hence, reduction in plantar flexor and knee extensor strength are correlate with falls in older people (Borges et al., 2017). Other than that, hip muscle strength can significantly disrupt comfort and balance of body movement. According to Neumann (2010), hip extensor muscle produced torque when body accelerated upward and forward. Reduced of hip extensor muscle may causes difficulty to climb step. Moreover, Rogers and Mille (2003) proposed that the sideways postural balance may impacted by interlimb hip abductor-adductor. Therefore, the strength of hip extensor, abductor and adductor is related to fall (Morcelli et al., 2014).

Horlings et al. (2008) found that eight studies reported increased falling risk with reduced muscle strength. The muscle strength measure not only from lower limb but included upper limb. Moreland et al. (2004) also reported combined odd ratio for upper extremity weakness to fall was 1.53, which consider significant correlation. There may some conflicts that whether hand grip strength is consider as efficient measure for muscle weakness with its relationship to fall. To clarify this, rapid arm movement and grasping are effective defence against sudden fall (Bateni et al., 2004; Allum et al., 2002). Hence, weak muscle strength on upper limb can increase the fall risk. In another way, hand grip strength may reflect strength of lower limb thus related to fall.

2.2.2.2 Assessment Tools

There are many tools used to measure muscle strength. Direct measures are straightforward which directly test the manual muscle strength while indirect measures examine through functional performance such as ability to get up from chair. These two measures cannot compare directly with each other but one suggests the use of direct measure instead of indirect (Horlings et al., 2008).

Among direct measures, hand grip strength (HGS) measurement is the simplest method by using digital dynamometer. Testers are requested to grip and squeeze the dynamometer as hard as possible for 3 seconds. After three trials for each hand, the score with highest value will be recorded. As stated in Akbar and Setiati (2018), the standard threshold of HGS for male was 26 kg, while female was 18 kg. This standard is slightly different based on nation and age groups. Testers will be classified as low muscle strength if record lower than specified standard threshold. However, HGS tests can be influenced by body size so Maranhao Neto et al. (2017) suggested allometric normalization of HGS with body height can provided more reliable result. This idea is accepted by Sevene et al. (2017) with D. Belka and DeBeliso (2019). In fact, HGS is proved highly correlate with functional mobility in many studies (Pijnappels et al., 2008; Wang et al., 2016; Akbar and Setiati, 2018). Therefore, it is a valid assessment tool to examine muscle strength and identify faller from non-faller.

Sit to stand (STS) is one of indirect measures. Testers are requested to stand from chair. The measurements are time taken per repetition or number of repetitions completed within 10 or 30 second. Performance on this will determines lower limb strength. This measure is proved that had moderate association (OR = 1.2) with falls (Tromp et al., 2001). The alternative view is that STS result affected by balance and other multiple factors instead of represent lower limb strength only (Lord et al., 2002). Therefore, it may effective in fall risk assessment but not for muscle strength.

2.2.3 Cardiovascular Disorder

Cardiovascular refers circulatory system which consists of heart and blood vessels. Its primary function is carrying oxygen and nutrient from heart to whole body. Disorder in cardiovascular may lead to insufficient supply of oxygen, loss of consciousness and then fall. Among of cardiovascular disorders, prevalence of fall risk increases with orthostatic hypotension. Orthostatic hypotension (OH) refers significant blood pressure reduction within 3 minutes of standing which systolic blood pressure decreases ≥ 20 mmHg or diastolic blood pressure ≥ 10 mmHg (Schatz et al., 1996). Low blood pressure causes slow transportation of oxygen to body parts especially brain which will easily lead to syncope and subsequent fall.

2.2.3.1 Current Researches Findings

Shaw and Claydon (2014) had expressed the relationship between falls and OH in flow chart. It also concludes that OH is associated with falls in older people. This is supported by several studies which report odd ratio (OR) of 1.7 to 2.5 (Heitterachi et al., 2002; Van Der Velde et al., 2007c). In contrast, Tromp et al. (2001) reported that OH was not associated with fall. This may because it is general study for all potential risk factors but not specifically for OH. Besides direct mechanism, which is reduced blood pressure, OH can associated with falls through impairments. To demonstrate this, diabetes older people without OH has better balance compare with those with OH (Cordeiro et al., 2009). Hohler et al. (2012) also reported that Parkinson’s patient with OH had higher level of disability. In summary, OH represents an intrinsic fall risk factor.

Apart from OH, hypertension is also one of cardiovascular disorders. It happens when diastolic blood pressure ≥ 90 mmHg or systolic blood pressure ≥ 140 mmHg and the duration exceed two occasions. Hypertension can increase the risk of OH and cause fall. Gangavati et al. (2011) reported that older people with uncontrolled hypertension had highest possibility of OH and greater risk of fall (hazard ratio = 2.5).

Moreover, cardiovascular drugs can have effect towards fall. Although there are insufficient data to show the ways that cardiovascular drugs cause fall, one studies reported that withdrawal of cardiovascular drugs can significantly decrease fall (Van Der Velde et al., 2007a). Common cardiovascular drugs include diuretics, beta-blockers, digitalis and statins. De Vries et al. (2018) had analysed sixteen drug class among 131 articles. According to study, medication of loop diuretics (OR = 1.36) and digitalis (OR = 1.60) may increase fall risk.

2.2.3.2 Assessment Tools

Commonly, manual sphygmomanometer is used as blood pressure measuring tool, but it can only provide instant result per measurement. The precise timing is required to capture the transient alternative blood pressure. In this case, beat-to-beat monitoring of blood pressure can provide more frequent and precise result. Hemodynamic system such as Finapres system and Task Force® Monitor are used to estimate beat-to-beat finger blood pressure (Van Der Velde et al.,2007b). However, blood pressure selection as interpretation result between lowest single beat or average over period was still inconsistent in studies.

Assessment in OH is done with blood pressure recording from two different body positions. It can further be classified into active and passive testing. Active tests involve muscles contraction when change of positions while passives tests do not. Active standing test (AST) involved lying-to-standing procedure which participants need five to ten minutes of rest at supine position and then stand upright. Measurement is conducted to check whether blood pressure decrease significant within three minutes according to definition of OH. On the other hand, standardised tilt table is used in passive head-up tilt test (HUTT). Testers still follow procedures as AST but raised upright at 60° to 80° of head-up tilt instead of active standing.

AST and HUTT have been used in many studies (Tromp et al, 2001; Weiss et al., 2002; Gangavati et al., 2011). There is disagreement regarding which assessment tool is more suitable as standard. According to Aydin, Soysal and Isik (2017), HUTT had higher sensitivity and specificity than AST. In contrast, several studies indicates the limitation of HUTT. Heitterachi et al. (2002) proposed that positioning finger on chest which above heart level in HUTT could affected the accuracy. Tan and Kenny (2006) also stated that more exaggerated response produced in HUTT which lead to misinterpreted. Besides that, there are several variability indices that computed from AST result can increase accuracy of fall prediction (Goh et al., 2016). In overall, AST which not required tilt table is more accurately to measure OH occurs in real life.

2.2.4 Cognitive Impairment

Cognitive impairment describes condition when person faces difficulty in concentrating, memorising and making decisions. Cognitive impairment is very common in older people which dementia, Alzheimer's diseases or stroke. There are some signs of cognitive impairment such as loss of memory, fail of recognition and vision problems. All these can affect the sensory and motor systems of human. As the result, people with cognitive impairment are difficult to regulate their gait, balance and response with environment changes. In other words, fall risk will be greater in this group.

2.2.4.1 Current Researches Findings

Cognitive impairment is known as fall risk factor in many studies (Sieri and Beretta, 2004; Rubenstein and Josephson, 2006). Vassallo et al. (2009) reported that the risk of fall for patients with cognitive impairment were higher. After conducted meta-analysis for twenty-six studies, Muir, Gopaul and Montero Odasso (2012) estimated the OR for cognitive impairment to any fall was 1.32 and serious injury fall was 2.33 among community-dwelling older people. All these had shown association of increase fall risk with cognitive impairment.

Cognitive impairment can be classified to disease-specific diagnosis or specific cognitive domain. Mild cognitive impairment (MCI) and dementia are inside category of disease-specific diagnosis. MCI is earliest sign of dementia which describes the state of cognitive functioning lower than ordinary (Feldman and Jacova, 2005). It can causes gait dysfunction in older people (Verghese et al., 2008). Impaired gait has greater incidence of fall. This is supported by Delbaere et al. (2012), it suggested the risk of multiple fall was twice in people associated with MCI compare to those without MCI. Apart from that, dementia is proved an independent risk factor for falling (Doorn et al., 2003). Dementia is described as severe or persistent disorder characterized by decline in memory and thinking skill. In fact, dementia are associated with other risk factors include impaired vision and motor impairment (Härlein et al., 2009). Different types of dementia can have different fall risk pattern. However, current researches still unable to provide adequate findings on this.

Executive function (EF) disorder is one example of cognitive impairment in specific cognitive domain. This disorder leads problems such as hard to concentrate and unable control self behavior. As stated by Herman et al. (2010), healthy older people with poorer EF had higher risk of fall.

2.2.4.2 Assessment Tools

There are several cognitive screening tools available. Although these tools are not diagnostic, it can still provide useful analysis in detection of cognitive change and possible underlying dementia. Sidal-Gidan (2013) provided good summary of various types of assessment tools with explanation.

Montreal Cognitive Assessment (MoCA) is primary examining short-term memory, executive function and concentration. Participants required to complete the test within ten minutes. The total mark for this assessment is thirty points and a score of equal or less than twenty-five point is considered subnormal. MoCA has excellent sensitivity to identify mild cognitive impairment and its short assessment duration useful in busy clinical setting (Nasreddine et al., 2005; Harkness et al., 2011).

Mini-Mental State Exam (MMSE) is generally applied assessment to examine attention, language and short-term memory. The maximum score is thirty marks and less than twenty-five marks is impairment suspected. It is different from MoCA as it takes 15 minutes to complete and executive function is excluded. Its primary assessment includes early dementia and Alzheimer’s disease. MMSE has high specificity but not very sensitive (Larner, 2012).

Many studies have compared the performance of MoCA and MMSE. According to Dong et al. (2010) was that MoCA had higher sensitivity than MMSE in vascular cognitive impairment detection after acute stroke. A similar view is held by Gluhm et al. (2013). This study found that MoCA is better cognitive impairment predictor than MMSE. Its finding also indicates that the mean MoCA score is lower than MMSE. This shows that MoCA is more challenging so it can distinguish cognitive impairment more accurately. On the other hand, MMSE is less capable to determine complex cognitive impairment.

Apart from MoCA and MMSE, there are also Memory Impairment Screen (MIS), Clock Drawing Test and Mini-Cog Test used for assessment tool. Each of the cognitive screening tools has its strength in specific clinical setting. However, one with high sensitivity and specificity should be chosen as ideal assessment tool. In this case, MoCA was highly prefered.

2.2.5 Demographic

Demographics describe the population based on factors. In this case, age and gender are identified as major fall risk factors. It has no direct relationship with fall. However, different groups of age and gender tend to have different intervention with each risk factor and contribute to fall.

2.2.5.1 Current Researches Findings

Many studies show fatal fall rates increase with age and gender. Age of 65 years old above has higher fall risk (World Health Organization, 2018). One study shows that OR for fall is increasing with age in both men and women (Gale, Cooper and Aihie Sayer, 2016). However, among the fallers, women are more significant associated with fall (Stevens and Sogolow, 2005). This is supported by Stevens et al. (2012).

Verghese et al. (2006) found that high incidence of abnormal gait increased with age. Normal human tends to walk slower when getting older. This may because of lesser energy and body strength due to biological factors. Besides that, a low speed may help to maximise balance and stability (Duxbury, 2000). Apart from that, increased stance width, period of double support phase and change of bent posture are characteristics of gait that may varied with aging (Salzman, 2011). Therefore, both speed and stability decrease when age increase (Schrager et al., 2008). However, there are still inadequate of accepted standards that clearly define a normal gait pattern in older people. Therefore, it is quite challenging to identify abnormal gait pattern in different age groups. According to Verghese et al. (2006), women had higher incidence of non-neurological gait abnormal compared to men. This may due to foot problems or medical risk factors. To support this, foot degeneration is more severe for women (Chaiwanichsiri, Janchai and Tantisiriwat, 2009). Hence, women tend to have slow walking speed and weaker balance.

Many studies have set age as inclusion criteria when examine the relationship between muscle strength and fall (Moreland et al., 2004; Borges et al., 2017). This shows that age is a factor that interferes with muscle strength in indirect way. To support this, Keller and Engelhardt (2013) proved that aging process had caused reduction of muscle mass and muscle strength. This may due to reduced number of muscle fibre and its size. Moreover, Allum et al. (2002)

proposed that balance correcting muscle responses were altered with age so older people are unable reacted immediately when sudden fall happened. Apart from that, the pattern of deficit in lower extremity strength is different for gender. Sieri and Beretta (2004) proposed that male faller had deficit in ankle plantar-flexion strength while female faller had lower knee extension strength. These differences in muscle strength can contribute to different risk for fall.

Orthostatic hypotension (OH) is also influenced by age. Low (2008) had proved that prevalence of OH increased with age. This may because the aging causes physiological changes which lead to orthostatic problems. As body's homeostatic mechanism, baroreflex helps to maintain blood pressure level through heart rate control. When blood pressure is decreasing, baroreflex will come out with a feedback loop so the heart rate will be faster to restore it back. In other words, blood pressure level will not fluctuate significantly if baroreflex is effective functioning. Furthermore, compliance describes blood vessel wall's capacity to actively expand and contract with changes in pressure. When such mechanism is not working properly, OH can easily happen. To illustrate this, older people associated with reduced of baroreflex responsiveness and cardiac compliance have higher risk for OH (Shibao et al., 2007). By the way, there are insufficient studies to show the OH prevalence with different gender.

With aging, brain processing speed and sensory perception are decreasing. Cognitive abilities also will decline as degenerate in brain structure (Murman, 2015). Besides that, brain damaged or degenerative dementias with age can lead to cognitive impairment. Therefore, risk for cognitive impairment is associated with age (Feldman and Jacova, 2005). In order to prove this, major study samples with cognitive impairment report a mean age above seventy years old (Muir, Gopaul and Montero Odasso, 2012). In addition, reduction in executive function is also associated with age. Such deficits can impair the ability of an older adult to compensate for age-related gait and balance changes (Herman et al., 2010). However, there are insufficient studies to show the cognitive impairment prevalence with different gender.

2.2.5.2 Assessment Tools

In this section, questionnaire is implemented as assessment tool to record down the age and gender.

2.2.6 Other Risk Factors

Instead of the major fall risk factors that had mentioned, there are also other minor factors that determine fall risk in older people. It is important to explore the effect of each towards the fall risk. The factors included falling history, fear of falling, medication, visual impairment and obesity.

2.2.6.1 Current Researches Findings

Falling history is the major factor for recurrent fall. Individuals with history of falls are threefold increased risk for falling again (Rubenstein and Josephson, 2006). It similar view, Dhargave and Sendhilkumar (2016) identified falling history had strong association with falling. Recurrent fall can be caused by same underlying fall factor as previous or associated with new fall risk factors. Therefore, it is important to ask whether patient have fallen before, the number of falls and its causes. Although falling history cannot directly linked to first fall, it can be useful information when screening for risk of future fall.

Fear of falling describes the psychological fear that can affect balance and functional performances. Subsequent falls can happen indirectly through fear of falling. According to Jung (2008), there were many modifiable risk factors related to fear of falling. History of fall is one of the modifiable risk factors. Individuals who have previous falls are more easily feel anxiety and depression associated with fear of falling. Denkinger et al. (2015) also concluded that walking ability and mobility disability were associated with fear of falling. However, there is lack of robust evidence shows fear of falling will cause falls as isolation factor. It is commonly together with other fall risk factors.

Major medication can influence the central nervous system and increase the fall risk. Van Vost Moncada and Mire (2017) proposed a table that listed common medication that associated with falls. These drugs are called fall risk increasing drug (FRID). Polypharmacy which consume high number of different drugs will cause higher risk for fall (Pfortmueller et al., 2014). This is because of different side effects and interactions between these drugs. However, the side effect of medication for each individual is non-identical.

Visual impairment also leads to increment of fall risk (Rubenstein and Josephson, 2006). Stimuli from visual and vestibular system can affect the balance of body. Therefore, individuals that have visual impairment may have lower sense of direction and muscle response. The relative risk for falls will be higher if visual impairment paired with other sensory impairment (Dhital, Pey and Stanford, 2010). Despite that, it still needs some researches to explore more in this area.

Obesity can increased the fall risk in older adults. Himes and Reynolds (2012) indicated that weight is linear proportional to fall risk. In other words, greater risk of falling for those individuals of obesity. As body weight rises, the balance control mechanism becomes less prone to controlling body sway oscillations. Therefore, there will be greater balance instability (Hue et al., 2007).

2.2.6.2 Assessment Tools

Questionnaire is used to record down of falling history. Number of fallen, risk factors associated with previous falling, injuries caused and difficulty after previous fallen should answered in detail (Arnold and Faulkner, 2007). Besides that, questionnaire should be conducted to record the medication review for types and total numbers of medication used.

Fear of falling can be assessed by single question "Are you afraid of falling?" or Fall Efficacy Scale (Denkinger et al., 2015). Fall Efficacy Scale consists of ten questions which total hundred marks to examine the level of fear of falling. However, the questions only evaluate indoor activities but not included outside activities. Therefore, Activities Specific Balance Confidence Scale which has more specific questions is more preferred (Jung, 2008).

There are some studies have used Snellen eye chart as assessment of visual impairment (Van Helden et al., 2007; Herman et al., 2010). Participants are requested to stand three meters away from eye chart and read. The participant is considering visual impairment if visual acuity is less than 0.40.

Body mass index (BMI) is a common tool that assess obesity. It is generated by participant's weight and height through simple calculation. Participant will classified as obesity when BMI is more than thirty. Cho et al. (2018) suggested that central obesity (CO) was accurate way to assess obesity with fall. It is assessed by using waist circumference. Participant will be classified as CO if waist circumference is more than 88 cm for women or more than 102 cm for men. Combined both BMI and CO measurement can provide more accurate result compared with BMI alone.

2.3 Current Fall Risk Predictors

Fall risk predictors are tools used to evaluate a patient's risk for falling. It includes relevant risk factors in a structured format. By answering the questions, patient can be identified whether he has high possibility to fall or not. This is quick and cost-effective method to facilitate busy hospital and clinical setting.

Hendrich II Fall Risk Model (HFRM II) is a common standard to predict fall risk. The latest version is developed in year 2003. HFRM II provides quick assessment on eight identified risk factors. It included confusion, depression, alter elimination, dizziness, gender, administration of antiepileptics and ability to rise in single movement. Each factor will be assigned specific score after evaluated by nurses. Individual who accumulates five point or above out of total sixteen points is determined as high fall risk. HFRM II can provides 74.9 % sensitivity and 73.9 % specificity of predictive result (Hendrich, Bender and Nyhuis, 2003). This proves that HFRM II is a validate tool to examine fall.

Morse Fall Scale (MFS) is also a tool to measure likelihood of falling. Six risk factors which included history of falling, gait, mental status, heparin lock, use of ambulatory aid and secondary diagnosis will be examined. Each factor will be assigned specific scores after evaluation. If the participant scores more or equal than forty-five points, his fall risk level is high. According to Morse, Morse and Tylko (1989), MFS had 78 % sensitivity and 83 % specificity.

St. Thomas Risk Assessment Tool (STRATIFY) is used to identify clinical characteristics and falling risk of older people. Oliver et al. (1997) reported that STRATIFY can predict fall risk at 93 % of sensitivity and 88 % of specificity for investigation in local cohort. It has five variables for assessment which are falling history, mental status, toileting frequency, visual impairment and mobility. Participants need to answer yes or no to each question. One mark will be assigned if the answer is yes and vice versa. Five questions contribute five marks in total. A score of above two will considered high falling risk.

After conducted meta-analysis among fourteen related studies, Aranda-Gallardo et al. (2013) had summarised the diagnostic odd ratios and likelihood ratio which represent the global performance ratio of each assessment tool. In comparison, STRATIFY has the best performance in assessing fall risk. However, the included fall risk factors also not exactly same in all predictors. Thus, the performance can changes depend on type of predictor and population.

2.4 Dimensionality Reduction

In machine learning, dimensionality is defined as the quantity of features inside input dataset. To deliver a reliable analysis, the amount of data that required for learning algorithm will increase if dimensionality is higher. In other words, more data are needed if number of features is larger. However, some algorithms are difficult to train an effective model in problem with huge features number but small sample size because it prone to overfitting (Hira and Gillies, 2015). To overcome this, dimensionality reduction included feature selection and feature extraction are proposed to preserve only significant features.

Dimensionality reduction is important technique in many automation applications especially medical field (Khalid, Khalil and Nasreen, 2014). To illustrate this, test results after various diagnoses can act as different type of features to assess the fall risk of patient. However, the analysis may not that meaningful because some irrelevant features are associated within existing data. Therefore, features selection and extraction are essential in this case. It can be used in isolated or combination to reduce dimension of feature sets and improve performance for subsequent processing stages (Motoda and Liu, 2002).

2.4.1 Feature Selection

Feature selection is useful to reduce size of search space by selecting subset from existing features. A brute force feature selection is assessing all possible relationship of underlying features by experience and expertise. However, this is not a reliable way to assess large dataset so it usually done by automatic feature selection (Krakovska et al., 2019). According to Cheng, Wei and Tseng (2006), feature selection algorithm was able to remove irrelevant attributes in medical data. Different from feature extraction, no new features are created after feature selection. This does not make the interpretation of features complicated for human comprehension. Therefore, feature selection is more widely used to analyse medical data due to this advantage (Samant and Rao, 2013).

2.4.1.1 Filter Type Feature Selection

Filter model selects features based on information content which are interclass distance or statistical dependence. Different from wrapper, it does not involves learning techniques. Most filter type feature selection techniques include feature ranking which determined by cross-validation (Santos, Datia and Pato, 2014). In univariate method, every feature is evaluated separately. Apart from that, multivariate method assesses the relationships among features.

Independent T-test feature selection is a general used method. It computes the statistical information and examine which group are statistically different from each other. To illustrate this, features with maximum inter-group mean value and minimal intra-group variability will be searched and used (Hira and Gillies, 2015). Correlation-Based Feature Selection (CFS) is also another common method to search feature subset corresponding to the degree of connectivity between the features. It finds the features that strongly correlated to class but uncorrelated to each other. CFS is proven as effective selection method to boost learning algorithm efficiency (Chormunge and Jena, 2018).

2.4.1.2 Wrapper Type Feature Selection

Wrapper method involved training and testing phases to evaluate which feature is meaningful. The wrapper approaches are good with precision because it chooses the best features but come with price of computational complexity. Sequential search is a heuristic based algorithm that find the features with the highest classification accuracy when every new function is added. This search is terminated when a new inserted feature does not improving selected feature criterion (Dy and Brodley, 2004). Besides that, genetic algorithm (GA) is a randomized approach which find the smaller set of features through the uses of evolutionary biology technique. However, the generation and population size of GA must be quite large to obtain an effective result.

2.4.1.3 Embedded Type Feature Selection

Embedded method is efficient because it integrates selection of features as part of the training process and is typically unique to the learning algorithms provided. Random forest is set of classifiers which use different samples of the original data to construct a variety of decision trees and compute the importance of each feature (Hira and Gillies, 2015). The feature of lowest importance may be excluded out. Another method is Least Absolute Shrinkage and Selection Operator (LASSO) which builds a linear model that sets multiple feature coefficients to zero and the non-zero ones is classified as chosen features.

2.4.2 Feature Extraction

Feature extraction is general method that developing a transformation from original features to a smaller set features while preserves most of relevant information at the same time (Chumerin and Van Hulle, 2006). Unlike feature selection, feature extraction produces new features through merging (Hira and Gillies, 2015). As medical data sets commonly small and high dimensionality, Li, Liu and Hu (2011) suggested feature extraction can improved analytical efficiency after extracted the optimal subset. In similar view, Tran et al. (2014) had proposed a framework for feature extraction which useful in risk prediction.

2.4.2.1 Linear Feature Extraction

According to Hira and Gillies (2015), data which transformed to lower dimensional space through linear mapping was represented as linear feature extraction. Principal Component Analysis (PCA) is the most well-known linear algorithm. It is proven as effective dimensionality reduction in medical data sets (Polat and Güneş, 2007; 2008). PCA aims to detect the correlation between variables and convert those data which have correlated features into linearly uncorrelated. Covariance matrix or correlation matrix are used to compute the covariance or correlation between two features. Based on eigen-decomposition, the eigenvector and eigenvalue of covariance matrix can determine new feature space directions and its magnitude. Most of the information about data set distribution are carried by the eigenvectors with highest eigenvalues. Besides that, eigenvalues can used to calculate the variance which determine total features number along the new feature axes. In other words, PCA uses covariance measure for redundancy minimisation and variance measure for information maximization (Khalid, Khalil and Nasreen, 2014).

Linear Discriminant Analysis (LDA) is used widely in dimension reduction which involving high-dimensional data. Its working principle is subspace selection with maximum discriminant power. LDA maps the data onto a lower-dimensional vector space in such a way that the ratio of the distance between the class and the distance within the class is maximized and thus maximizes discrimination. In past twenty years, LDA was developed and applied as pre-processing step. De La Torre and Kanade (2006) had expressed LDA in matrix factorization which more convenient to understand. As LDA has similar properties with K-means clustering, Ding and Li (2007) proposed LDA used as subspace selection before K-means. This study shows high clustering accuracy with this approach. However, LDA has singularity problem that affect the its performance in certain applications. To solve this, an intermediate stage by using PCA before LDA can be used. Besides that, two-dimensional LDA is proposed to overcome limitation in classical LDA and thus improve efficiency (Ye, Janardan and Li, 2005).

2.4.2.2 Non-Linear Feature Extraction

In real life, there may have non-linear relationship exits in linking variables. These non-linear dependencies can increase the difficulty in correct dimensionality reduction as many linear methods can fail to adequate identify them (Krakovska et al., 2019).

In this case, the use of kernel function provides a powerful and principled way of detecting non-linear relations (T. SenthilSelvi and R. Parimala, 2018). It is usually combined with linear algorithm. For example, Kernel Principal Component Analysis (KPCA) maps the data into a high dimensional feature space by using nonlinear mapping first then apply PCA to extract the optimal feature subspace (Li, Liu and Hu, 2011). Jade et al. (2003) proved that good performance of KPCA as features extraction and denoising method.

In fact, t-Distributed Stochastic Neighbor Embedding (t-SNE) is also a popular method that used to map high dimensional data to only two or three dimensions. It achieved the better visualization result compare to other non-parametric visualization methods (Van Der Maaten and Hinton, 2008).

Moreover, non-linear feature extraction can be accomplished by neural network approach. The basic idea is using feedforward neural networks along with single hidden layer as newly extracted feature (Motoda and Liu, 2002). This neural network is initiated by one hidden unit and its predictive accuracy is estimated. Then, the network is enhanced by adding additional units until it fully connected. At this stage, irrelevant or redundant network will be removed. In short, this approach is designed to find minimum number of hidden units that contains most of the information. The hidden units represent features extracted from original data set. Autoencoders and Self Organizing Map (SOM) are examples of this (Hira and Gillies, 2015).

2.5 Clustering Techniques

Data clustering defined as unsupervised classification which partition objects into different groups without class labels. The primary objective of clustering is to explore series of underlying patterns from natural grouping which useful for anomalies detection (Oyelade et al., 2019). An efficient clustering should have the maximum similarity for intra-cluster while minimum for inter-cluster. In recent years, clustering is adopted as machine learning technique to identify patterns of various diseases and develop risk predictive model for patients (Álvarez et al., 2019). Although there are numerous studies available for comparison of different clustering techniques, but it still lacks of empirical result to decide which clustering approaches can obtain the most reliable and accurate results (Saxena et al., 2017; Rodriguez et al., 2019). This is because different approaches have its own strength in specific input data and applications. The major types of clustering techniques that were summarised in Figure 2.1.

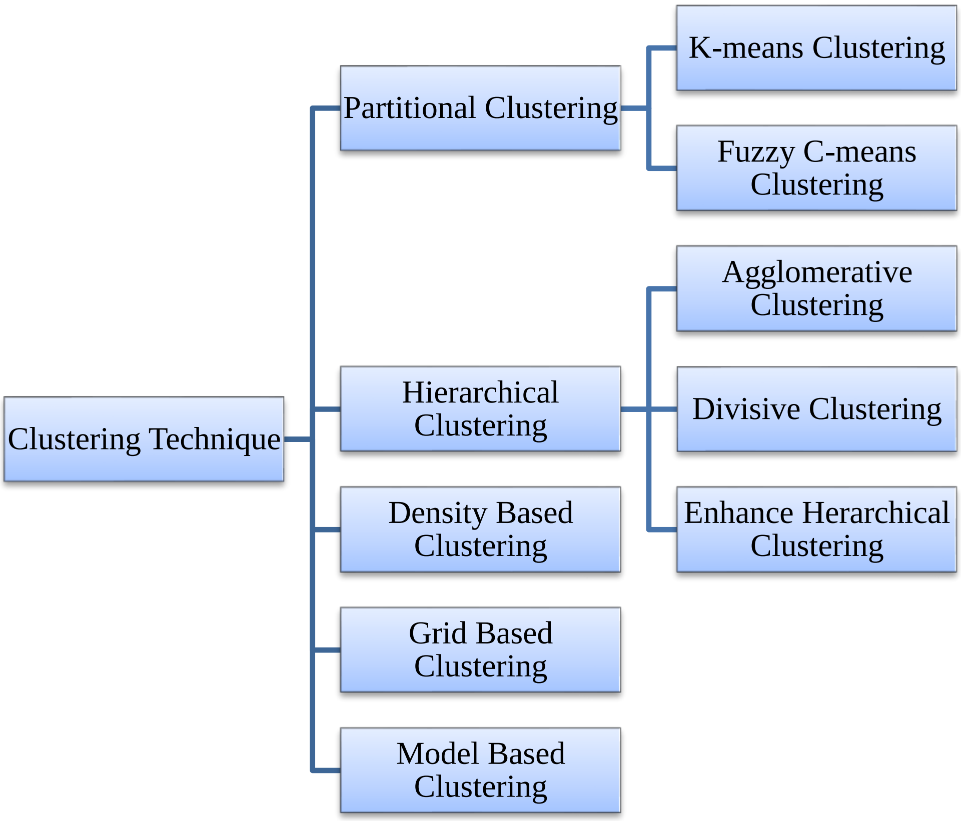


Figure 2.1: Types of Clustering Techniques.

2.5.1 Partitional Clustering

Partition clustering is widely used technique due to its usability and easiness of execution. Its working principle is decomposing objects of a dataset into different clusters based on predefined objective function and improve iteratively for partition efficiency until possible optimisation made (Saxena et al., 2017).

2.5.1.1 K-means Clustering

K-means clustering is popular algorithm among partitional approaches (Valarmathy and Krishnaveni, 2019). This clustering algorithm requires defined number of cluster (k) and centroid for each cluster. According distance to centroid, each data point is initially associated with nearest cluster. New centroid is computed based on associated data point and classification process repeated until convergence criterion happen in new iteration. In fact, K-means clustering applies the objective function of Sum Squared Error (SSE) which measure of variation within a cluster. The SSE is decrease with each iteration so that grouping can identified more correctly.

K-means clustering is simple and efficient method used in medical diagnosis (Nithya, Duraiswamy and Gomathy, 2013). According to Escudero, Zajicek and Ifeachor (2011), K-means clustering was applied to integrate information from diverse variables into relevant disease pattern. To illustrate this, Guo et al. (2017) divided participants into specific groups based on diagnostic features and identified underlying risk factors with K-means analysis.

The performance of K-means clustering is affected by initial centroids chosen and number of clusters. This is due to the final classification can rely heavily on these factors. Therefore, several modifications or enhancement of this algorithm are proposed. (K-means ++) initialization follows weighted probability score to select the first centroid. Malarvizhi and Ravichandran (2018) proposed that (K-means ++) had lesser computed time and better accuracy compared with traditional K-means algorithm in clustering of medical datasets. A similar view is held by Kalyani (2012) which stated enhanced K-means algorithm achieved better performance.

Instead of K-means clustering, K-median and K-modes also can produce reliable results according different scenario in dataset. Other than that, Partitioning Around Medoid (PAM) creates cluster by making use of medoid.

The obtained medoids are highly resistant to outliers and noise (Oyelade et al., 2019). However, it is high cost algorithm compared with K-means clustering.

2.5.1.2 Fuzzy C-means Clustering

Hard assignment of cluster points is not feasible in complex datasets where clusters overlap. To solve this, a fuzzy clustering algorithm can be used to extract such overlapping structures. In fuzzy C-means (FCM) clustering, the membership of each point to multiple clusters may range from zero to one but the weighted sum must be equal to one. Then, the membership and centroid are updated after each iteration. In other words, this approach allows two or more clusters have similar point at the same time. However, the objective of FCM still same which find centroids that minimize a dissimilarity function.

Ramya (2018) had proposed disease prediction system by using FCM. In this system, the membership degree is associated with the value of features in clusters so that the cluster is not affected much by noise. Apart from that, Rustempasic and Can (2013) reported combined FCM with pattern recognition systems were useful in diagnosis of Parkinson’s disease. In addition, FCM provides better result than hard-K-means algorithm to cluster thyroid gland data (Albayrak and Amasyalı, 2003). All these shown that membership function can improves the clustering performance especially in medical diagnosis system.

2.5.2 Hierarchical Clustering

Hierarchical clustering creates a nested sequence of clusters. Different from partitional clustering, predefined number of clusters is not required in hierarchical clustering. This algorithm will decide the appropriate clusters or groups in the end of process. This approach allows a more heuristic and robust process for clustering data objects. Hierarchical clustering is an useful clustering technique in medical domain (Nithya, Duraiswamy and Gomathy, 2013). This technique will uncover trends using either a top-down or a bottom-up strategy. Therefore, it can be categorised into agglomerative (bottom-up) and divisive (top-down) clustering methods which illustrated in Figure 2.2 (Pawan, 2019).

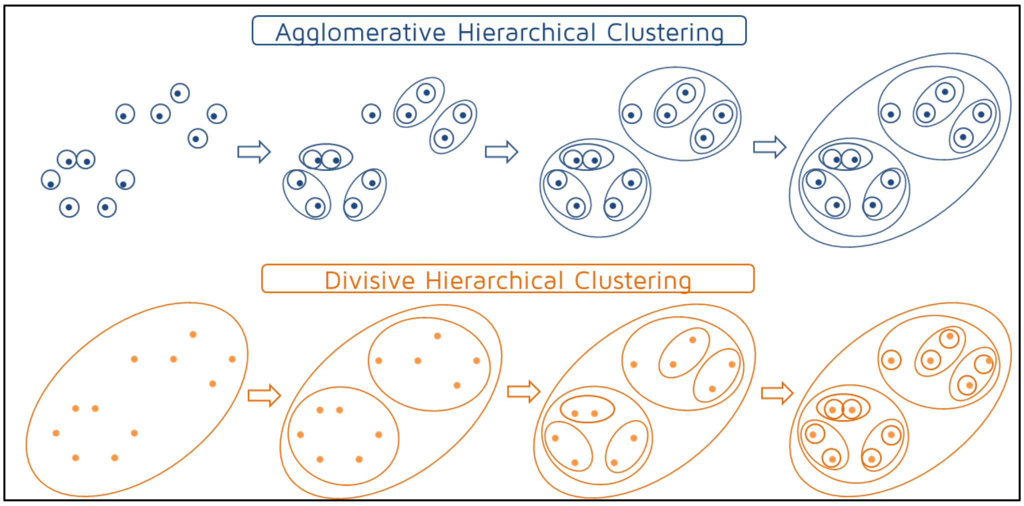

Agglomerative Hierarchical Clustering
Divisive Hierarchical Clustering

Figure 2. 2: Types of Hierarchical Clustering.

2.5.2.1 Agglomerative Clustering

Agglomerative clustering begins with a singleton cluster having just one data object per cluster. All clusters are now uniquely depicted at the base of the dendrogram. Then, the nearest cluster sets begin to merge at a time to create a bottom-up cluster hierarchy. This process is terminated when final cluster which contain all data objects achieved (Murtagh and Contreras, 2012).

Agglomerative clustering can further breakdown to three categories of clustering based on linkages (Saxena et al., 2017). The first is a single-relation clustering, the relation between the two clusters is created by a single entity pair. The distance between two clusters in this clustering is measured by shortest distance from either member of one group to any member of other group. Complete link clustering tests the resemblance between two clusters as their nearest dissimilar members are identical. It is similar as choosing the pair of clusters whose merger has the smallest diameter. The last one is the clustering of average-linkage also known as the form of minimal variance. Average distance from either member of one cluster to any member of the other cluster determines the distance between two clusters.

2.5.2.2 Divisive Clustering

On the other hand, divisive approaches begin with all the data entities in a large macro-cluster and break it continually into two classes, creating a top-down structure of clusters (Rodriguez et al., 2019). This approach has the benefit of being more powerful in contrast with agglomerative clustering particularly when there is no need to produce a full hierarchy all the way down to the individual leaves. However, there are several factors to affect the performance of this algorithm. The primary factors are the splitting criterion and method used. K-means square error standard may be used to get effective division here. Selecting the cluster chosen to split might not be as relevant as the first two reasons but selecting the most suitable cluster to split further while the aim is to create a compact dendrogram may also be beneficial. An easy way of picking the cluster to be further separated may be achieved by simply testing the cluster's square errors and separating the one with the greater value. According to Praveen and Rama (2018), divisive clustering algorithm by using the mean value of objects provided good performance in numeric clustering.

2.5.2.3 Enhanced Hierarchical Clustering

The key shortcoming in conventional hierarchical clustering is that it cannot pass inside a hierarchy of other clusters once two points of the cluster are connected to each other. Therefore, some enhanced hierarchical clustering has been proposed. COBWEB addresses the uncertainty associated with categorical attributes in the clustering by means of a probabilistic model close to Naive Bayes (Saxena et al., 2017). For this method, the dendrogram is sometimes called a classification chain, and the nodes are called concepts. In addition, the CHAMELEON method utilizes a graph-based partitioning algorithm to initially organize the data entities into large amounts of small sub-clusters such that items in each cluster are closely connected and therefore less influenced by outliers (Praveen and Rama, 2018).

2.5.3 Density Based Clustering

Density based method is used to discover clusters of arbitrary shapes. Its working principle is clustering the regions which have high point density and separate out those with low density. In other words, this method relies on distance and spatial location of data points. Therefore, it chooses the number of clusters itself based on input data instead of defining it at the beginning.

Among density-based approaches, Density-based spatial clustering of applications with noise (DBSCAN) is the most common used. DBSCAN requires two parameters which are epsilon (maximum distance from one point to another point) and minPts (minimum number of neighbour points) (Mali, Kulkarni and Bagade, 2017). Core points which have overlapping neighbourhood are determined and form the skeleton of a cluster. Objects that are not associated with cluster considered noise (Valarmathy and Krishnaveni, 2019). The cluster is expanding until all points in dataset were examined.

According to Daszykowski, Walczak and Massart (2004), DBSCAN was efficient and had high computational speed for exploration of analytical data. In similar view, Ogbuabor and F. N (2018) reported that DBSCAN had good clustering performance in healthcare dataset. In recent years, some studies have proposed the enhanced DBSCAN method can improve the its performance (Kalyani, 2012; Tran, Drab and Daszykowski, 2013). Besides that, there is a view suggested by Al-Shammari et al. (2019) which combined of Piece-wise Aggregate Approximation (PAA) and DBSCAN can provide more efficient clustering in medical data streams.

2.5.4 Grid Based Clustering

Grid based clustering creates grid structure and merge the grid cells to obtain cluster. Statistical Information Grid-Based Clustering (STING) is one example of this clustering method. Park and Lee (2004) illustrated the use of STING algorithm for data stream. STING has low computational cost, but it requires predefined density parameter which can affect the quality of clustering. Optimal grid (OPTIFRID) is another method which the dataset is partitioned in a region of low density (Oyelade et al., 2019). This approach is efficient for clustering high dimensional databases with noise. Although grid-based clustering is well known, there are still limited study show the use of this clustering algorithm in medical domain.

2.5.5 Model Based Clustering

Model based clustering approaches optimize with certain mathematical models as well as find the eligibility of the provided results. Similar to traditional clustering, model-based clustering approaches often detect characteristic information for each cluster, where each cluster reflects a category or group (Saxena et al., 2017).

A neural algorithm model-based clustering is self-organising map (SOM). It is commonly used for feature extraction, visualisation and data mining (Taşdemir and Merényi, 2009). SOM also same as other neural network approach which involved training and mapping phases. Typically, it consists of two-dimensional grid of map. Throughout the learning process, weight of neuron is randomly initialised. Data points in the input space located near each other are mapped to local map units. The training phase utilized competitive learning. The neurons of the prototype compete for the recent example. The winner is the neuron, whose weight vector is closest to the present case. The champion and his neighbours learn by changing their weights. After numerous iterations, SOM can success divide the input data into several clusters. These mechanism is discussed detail in (Azzag and Lebbah, 2008). Besides that, several studies have proposed enhancement for SOM technique (Kiang, 2001; Kumar Roy and Mohan Pandey, 2018). Its application included image processing, speech recognition and medical diagnosis (Hsu, 2006). According to Markey et al. (2003), SOM can used for cluster analysis of breast cancer database.

2.6 Clustering Validation Measures

The evaluation of validity and accuracy for generated partitions is important step in cluster analysis (Rodriguez et al., 2019). Criteria used to calculate the reliability of the partition may be classified as internal and external.

The internal validation indices are focused on compactness and separation measure. It is important to determine how similarly each instance relates to the cluster and how far the cluster is isolated from the other clusters. Similarity of points in same cluster delivers the most critical purpose in this case. Often clustering techniques are using distance calculations to determine the similarity (Saxena et al., 2017). Examples of distance measures included Euclidean distance, Manhattan distance and Jaccard distance. Apart from this, Cosine measure and Pearson correlation measure can also be used (Mali, Kulkarni and Bagade, 2017). Different measures may give different outcomes, so it is better to understand the mechanism of each methods before selection., Silhouette analysis and Davies-Bouldin criteria are the popular internal validation methods. Silhouette analysis measures on the distance of each point within the cluster to points in neighbouring cluster. Davies-Bouldin criteria measures the interclass to intraclass distance ratio.

On the other hand, external validity indices calculate the consistency between the performance of the cluster algorithm and the proper partitioning of the dataset. The Jaccard index is a well-known technique to define the equivalence of the two datasets (Mali, Kulkarni and Bagade, 2017). In addition, the Rand Index is a basic metric used to measure how close clusters are to the standard classifications.

2.7 Summary

To conclude this chapter, the main risk factors were identified throughout literature reviews. Gait and balance, muscle strength, cardiovascular disorder, cognitive impairment and demographic factors are proved to have strong correlation with fall among older people. The other risk factors such as fall history, fear of falling and visual impairment was not directly linked with fall but interact with major factors hence increase the fall risk. Besides that, various assessment tools for each fall risk factor were discussed. Choosing the right assessment tool is important for identification of actual fall risk. In general, current fall risk assessment likes HFRM II, MFS and STRATIFY can provide fast yet reliable results based on same situations. However, the accuracy may not there because the diagnosis symptoms can be different as time pass and some important features might not include in such assessment. Therefore, feature selection, feature extraction and clustering techniques are identified to develop a machine learning algorithms for fall risk assessment in older people.

CHAPTER 3

METHODOLOGY AND WORK PLAN

3.1 Introduction

In order to complete this project, the related information was collected through background study and literature review. Some appropriate dimensional reduction techniques and clustering techniques were identified and tested. After this, the clustering algorithm included stages of data pre-processing, feature selection, feature extraction, clustering and characteristic interpretation was constructed based on the selected techniques. This algorithm was evaluated and enhanced to achieve the objective of this project. In this chapter, the methodology is discussed in detail.

3.2 Equipment

This project only involved software equipment. It included:

i. Spyder (Anaconda) software (Released 2019. Version 3.3.6.)
ii. Statistical Package of Social Sciences (SPSS) software (IBM Corp. Released 2018. IBM SPSS Statistics for Windows, Version 20.0. Armonk, NY: IBM Corp)

Spyder is powerful integrated development environment (IDE) written in Python. The Python version used here was Version 3.6. Python offered advanced development tools in data exploration and visualization. The algorithm was developed in Spyder. Apart from that, a large and complex data set can quickly understand with advanced statistical procedures in SPSS software. SPSS was used for statistical analysis on the generated result.

3.3 Proposed Clustering Algorithm

The overview of clustering algorithm was proposed in Figure 3.1. It consists of several stages. At the first stage, the input dataset was imported and analysed. In the data pre-processing stage, the algorithm was handling the missing data inside the dataset and categorized data into different category. Normality testing was used to examine the data distribution of numerical variables. Feature selection was conducted in sequential order. Hypothesis testing (Independent T-test, Mann-Whitney U test or Chi-squared test) was performed on all variables. Next, correlation filter (Spearman correlation and Cramer V correlation) was applied on those significant variables. Then, the feature importance of the selected variables was computed. After feature selection, principal component analysis (PCA) and t-distributed stochastic neighbour embedding (t-SNE) were applied to transform the chosen data into lower dimensional space. In clustering stage, K-means, Hierarchical and Fuzzy C-means clustering were implemented for each transformed data. The performance of each combination was evaluated through cluster analysis. The fall risk and characteristic of each clustered group was analysed in the last stage of algorithm.

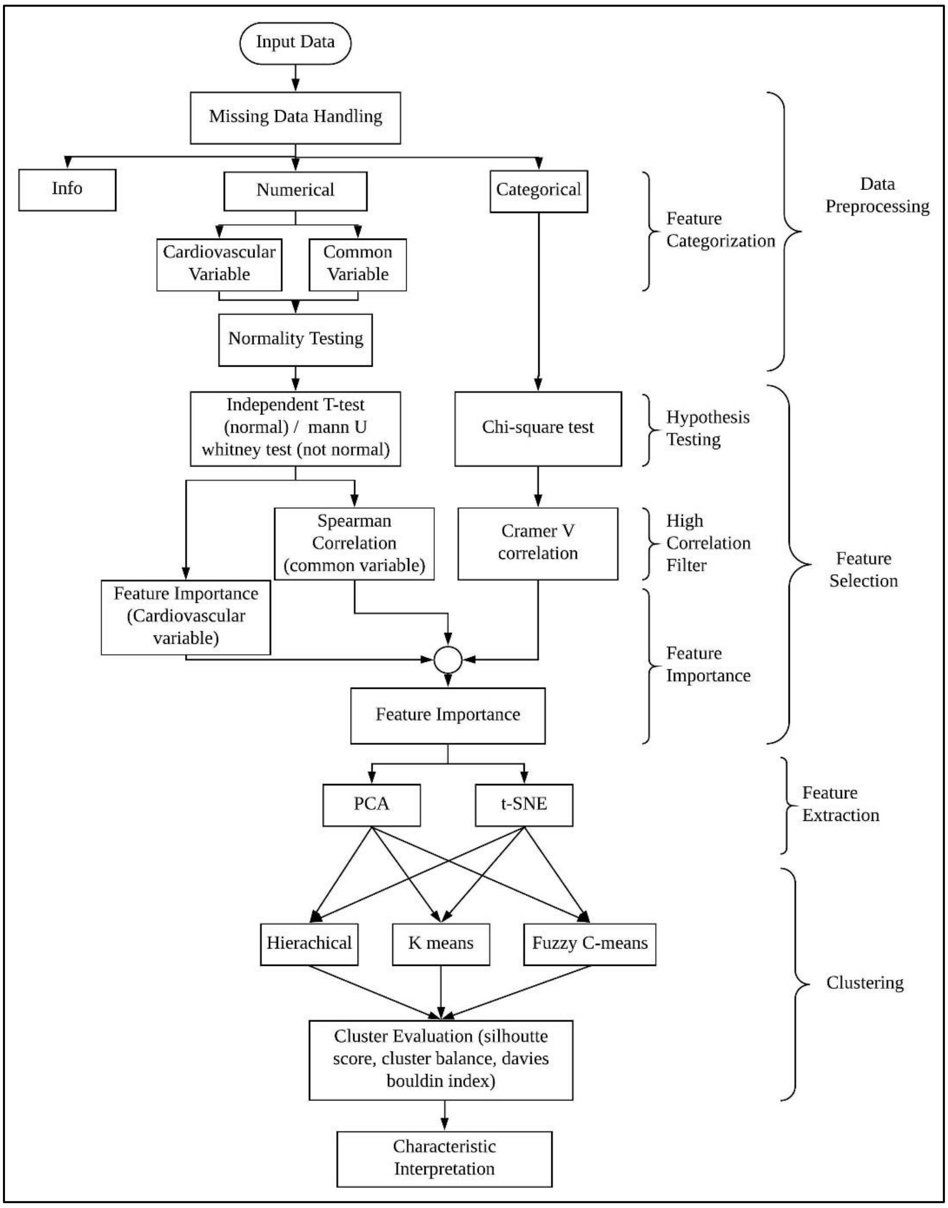


Figure 3.1: Overview for Proposed Clustering Algorithm.

3.3.1 Input Data

The study dataset is obtained from Malaysian Elders Longitudinal Research (MELoR). The aim of MELoR study is to investigate various aging effects including prevalence of falls towards older people population. This study involved participants who were community-dwelling older Malaysians and aged fifty-five and above. Individuals are selected if the inclusion criteria are met and informed consent is obtained. The data are collected from questionnaire interview in phase one then assessments in phase two.

The questionnaire is developed by a panel of experts from different areas. It is conducted as home-based interview through computer-aided platform. Participants are interviewed by interviewers. They are required to provide their basic demographic data, fall history and medication history. If the participant himself is unable to answer specific question, his relatives are asked to provide the relevant information. All the answers are double confirmed with participant’s relatives before recorded down.

In phase two, participants are required to attend at University of Malaya Medical Centre to conduct both physiological and medical assessments. Basic anthropometry included standing height, weight, waist and hip circumference are measured by appropriate measurements. The Jamar Plus + digital hand dynamometer (Sammons Preston, Illinois, USA) is used to measure hand grip strength. As for gait and balance, three tests are conducted in standard procedures. It included TUG test, frailty walk test and functional reach. Apart from that, cardiovascular autonomic reflexes are measured by Active Standing test with continuous beat-to-beat blood pressure monitored. Montreal Cognitive Assessment (MoCA) also conducted in questionnaire to screen for cognitive impairment based on the score.

All these procedures were following operating standard to ensure consistency. Fallers are identified in question “Have you fallen in the past 12 months?”. Again, the relative will requested to assist in such question if the participant found difficulty to provide reliable answer.

In overall, the study has recruited a total of one thousand four hundred eleven community-dwelling older people, who underwent a comprehensive interview and clinical assessment (one hundred and thirty-nine variables related to falls were extracted).

3.3.2 Data Pre-processing Methods

Input dataset contains huge number of rows and columns (1411 subjects x 139 variables). Data pre-processing was used to transform the raw data into understandable and accessible format. It included missing data handling, feature categorization and normality testing.

3.3.2.1 Missing Data Handling

Missing data are common in medical dataset. To illustrate this, people who did not complete blood pressure measurement cannot provide the blood pressure record. The missing data are reducing the statistical power and representativeness of the dataset. The approach chosen in this case was excluding those missing data and analysed the remaining data. The data are either excluded in rows or columns (Jason, 2017). Before the stage of feature selection, the percentage of missing data in each variable (column) was identified. If the percentage of missing ratio was exceeding the threshold (10%), such variable was excluded first. Then, the subject (row) that contained one or more missing data in selected variable was excluded after feature selection. This is to avoid large data being excluded due to those irrelevant variables.

3.3.2.2 Feature Categorization

There are different types of data included categorical and numerical data. Categorical variable contains defined set of values while numerical variable contains continuous or integer values. The univariate feature selection methods are different when deal with categorical or numerical data (Jason, 2019). Therefore, feature categorization step was used to classify the type of data so the feature selection method can be applied based on each type of data.

Besides that, some of variables may contain only basic information such as name and ID. Therefore, it was also used to filter out those irrelevant variables before subsequent stage. However, this step is based on the domain knowledge.

3.3.2.3 Normality Testing

Shapiro-Wilk test was used to examine the normality of continuous variables. The null hypothesis is stating the data is in normal distribution. The variable is indicated as non-normally distributed if this test rejects the null hypothesis (the computed p-value is less than 0.05) (Jason, 2018). The test was implemented with 95% confidence. On the other hand, passing the normality test (the computed p-value is more than 0.05) shows that no major deviation from normality has been detected. This normality testing was assisted with histogram plot. It is one of graphical method that used to evaluate whether the distribution follow familiar bell shape. The variable was classified to normal distribution if bell shape is observed.

3.3.3 Feature Selection Methods

Feature selection was conducted to select the relevant variables from original dataset. It was conducted through hypothesis testing (Independent T-test, Mann-Whitney U test or Chi-square test), high correlation filter (Spearman correlation and Cramer V correlation) and feature importance (random forest classifier) in sequential order.

3.3.3.1 Hypothesis Testing

The variable that can identify between fallers and non-fallers was considered as important variable. Therefore, hypothesis testing was performed on all variables to evaluate the difference between fallers and non-fallers. To illustrate this, it tests whether fallers have older age compare to non-fallers. If the result obtained was positive, age was an important variable. Independent T-test is commonly used to assess whether two unrelated groups are statistically different from each other, provide the data is normal distributed as shown in Figure 3.2 (Vadim Uvarov, 2018).

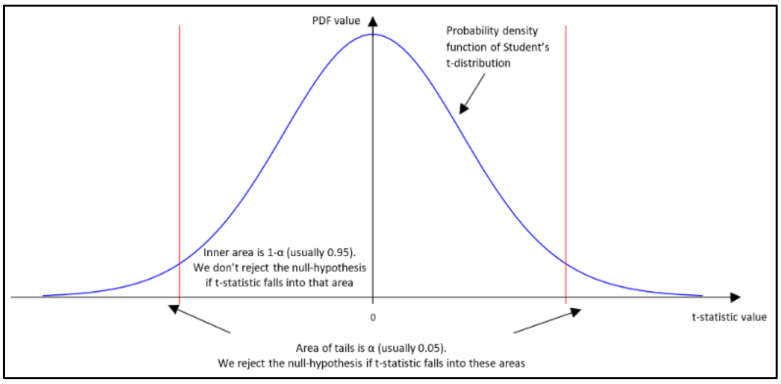

PDF value
Probability density function of Student's t-distribution
Inner area is 1-α (usually 0.95). We don't reject the null-hypothesis if t-statistic falls into that area
0
t-statistic value
Area of tails is α (usually 0.05). We reject the null-hypothesis if t-statistic falls into these areas

Figure 3.2: Probability Density Function of T-test.

A null hypothesis was stated that there is no difference between two measured variables. The probability to accept or reject hypothesis is depend on p-value. Assume that (significance level α=0.05), p-value obtained which less than α indicates the null hypothesis rejected, there are difference between two groups. In order to implement T-test, the variables were selected as test variables. After computation, the output significant value was used to compare with significance level, α. If it was less than significance level, such variable was significant variable and hence keep for subsequent stages. On the other hand, the variable was excluded to reduce the dimensionality as it cannot identify between faller and non-faller.

The Mann-Whitney U test was performed same function as independent T-test but for the not normally distributed variable (Jason, 2018). The feature selection was conducted with Mann-Whitney U test if such variable fail Shapiro-Wilk test. As for categorical variable, Chi-squared test for independence was applied (Bedre, 2019).

3.3.3.2 High Correlation Filter

The correlation method was conducted by Spearman correlation. The Spearman correlation coefficient calculates the linear relationship between variables. The value of the coefficient ranges between -1 and +1, where there is no association at 0. Correlations approach to -1 or +1 suggested a very good linear association. Besides that, coefficient of -0.5 or +0.5 represents a moderate correlation. The heatmap was constructed to visualise the correlation among the variables.

If the variable had high correlation (above coefficient of -0.8 or +0.8) with another variable, only one variable will be selected (Shetye, 2019). Cramer V correlation was conducted for the categorical variables.

3.3.3.3 Feature Importance

Feature importance is a technique that used to assign scores to input variable in a predictive model. The score indicates the relative importance of each variable when conducting prediction. Thus, the most relevant variable has the highest relative score and hence should be remained. On the other hand, the variable which has lower score is removed because it is not much important toward the model.

The random forest classifier is a meta estimator that applies a variety of decision tree classifiers to different sub-samples of the dataset and uses averaging to improve predictive precision and control over-fitting. In this case, it was used as predictive model. The input variables were fit into random forest classifier. Then, the importance score for each variable was observed. The variables with higher score were chosen for subsequent stage (Shaikh, 2018).

3.3.4 Feature Extraction Methods

Feature extraction included PCA and t-SNE transformed the input data to more manageable dimensional space for processing. Before this, all the selected variables from feature selection stage were standardized within a specific range to prevent variable from large domain dominates. Z-score standardization was used to transform the data into distribution which has mean of zero and standard deviation of one (Goyal,2020). It was computed through equation 3.1.

$$X_{new} = \frac{x - \mu}{\sigma} \tag{3.1}$$

where

x = original attribute

μ = mean before standardization

σ = standard deviation before standardization

3.3.4.1 Principal Component Analysis (PCA)

The basic idea of PCA is linear transformation from input space to another dimensional space. The coordinates of data in the new space are uncorrelated and have maximum variance. It preserved only small number of attribute (Shihab, 2004).

The covariance matrix was obtained through equation 3.2. The covariance matrix describes the association between the variables in the data set. It is important to recognize highly dependent variables as they contain bias and repetitive information. Besides that, covariance matrix consists of both eigenvector and eigenvalue were computed. The eigenvectors are used to classify and calculate the principal components. Eigenvalue describes the magnitude of respective eigenvector. After computing all the principal components, it was sorted in descending order from highest to lowest eigenvalue. Only predefined number of eigenvectors with respective eigenvalue were chosen as it already contained most of the information (Gursewak.S, 2020). The PCA steps were summarised in Figure 3.3.

$$Cov(x,y) = \frac{1}{N}\left(\sum_{j=1}^{n}\left(x_j - \bar{x}\right)\left(y_j - \bar{y}\right)^T\right) \tag{3.2}$$

where

N = number of samples in class

$\bar{x}$ = mean vector of input data

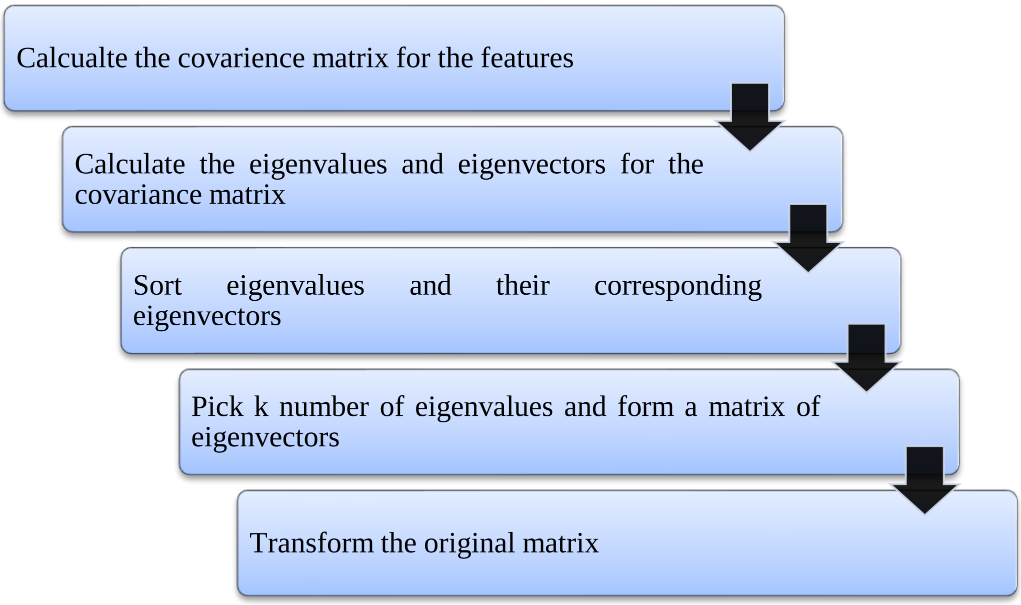

Calcualte the covarience matrix for the features
Calculate the eigenvalues and eigenvectors for the covariance matrix
Sort eigenvalues and their corresponding eigenvectors
Pick k number of eigenvalues and form a matrix of eigenvectors
Transform the original matrix

Figure 3.3: Steps Involved in PCA Feature Extraction.

3.3.4.2 t-Distributed Stochastic Neighbour Embedding (t-SNE)

t-SNE algorithm works by estimating the probability of similarity of points in high dimensional space. Then, it tries to recreate the similar probability distribution at low dimensional space (Pathak,2018).

The similarity of points was determined as the conditional probability that point A will select point B as its neighbour if neighbours were selected in proportion to their probability density under the Gaussian (normal distribution) centred at A. Then, t-SNE attempt to minimises the difference between similarities in higher dimensional and lower-dimensional space. Kullback-Leibler divergence is a calculation of how the distribution of probability varies from the predicted distribution of probability. In other words, t-SNE minimize the divergence between both distributions. After this, those data were recreated in lower dimensional space. The t-SNE steps were summarised in Figure 3.4.

The hyperparameter of perplexity was described the effective number of neighbours for any point. Alteration of this parameter will provide different results. Therefore, several testing were conducted to find the most suitable perplexity for this dataset.

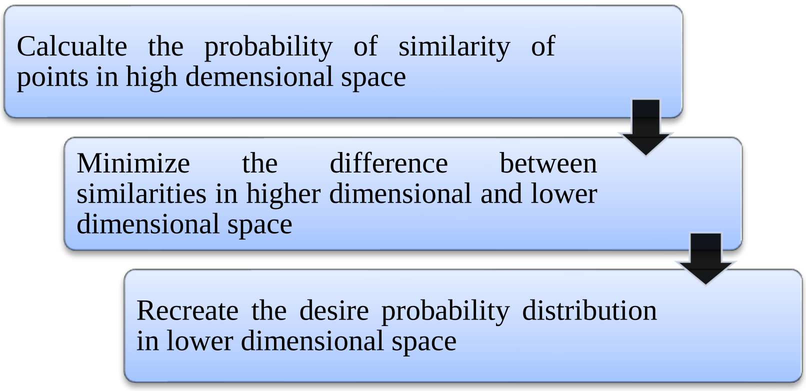

Calcualte the probability of similarity of points in high demensional space
Minimize the difference between similarities in higher dimensional and lower dimensional space
Recreate the desire probability distribution in lower dimensional space

Figure 3.4: Steps Involved in t-SNE Feature Extraction.

3.3.5 Clustering Methods

The clustering methods were included K-means clustering, hierarchical clustering, and Fuzzy C-means clustering. Different clustering methods may provide different results in different dataset.

3.3.5.1 K-means Clustering

K-means clustering is one of partition methods. It required the number of clusters before the algorithm applied (Ogbuabor and F. N, 2018). Elbow method and Silhouette coefficient were used to evaluate the suitable number of cluster. K-means clustering was chosen because it is simple to implement. First, the centroid of each cluster was set randomly. Then, each data point was allocated to closest centroids. Euclidean distance as equation 3.3 was used to calculate the distance between points and centroid. The cluster centroids were recomputing when new data points were inserted. These steps were kept iterating until convergence was observed. The steps for this clustering algorithm was summarised in Figure 3.5.

$$d_{x,y} = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2} \qquad (3.3)$$

where

x = position of data point

y = position of centroid

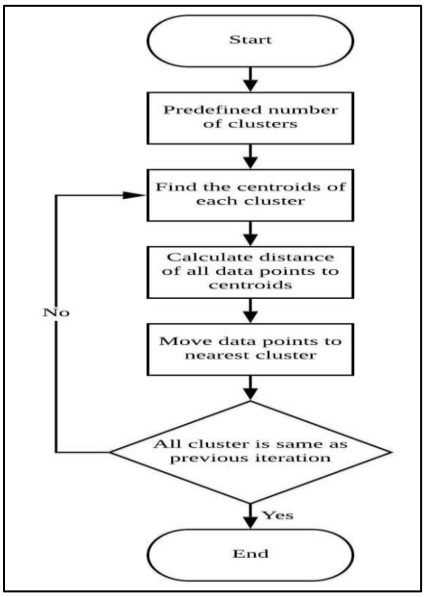

Start
Predefined number of clusters
Find the centroids of each cluster
Calculate distance of all data points to centroids
Move data points to nearest cluster
All cluster is same as previous iteration
No
Yes
End

Figure 3.5: Flow Chart for K-means Clustering Algorithm.

3.3.5.2 Hierarchical Clustering

Agglomerative clustering which using bottom-up approach was used. At beginning, each data point was considered as one single cluster. Then, the closest data clusters were start merging each other. This step was repeated until the final cluster which contain all data points was formed (Malik, 2018). The steps for agglomerative clustering algorithm was summarised in Figure 3.6.

The dendrograms were used to illustrate the number of clusters formed in different Euclidean distance. The number of clusters can be determined by defining the minimum distance required to be a separate cluster. Besides that, the criterion for choosing the pair of clusters to merge was chosen as ward method. It minimizes the total within-cluster variance.

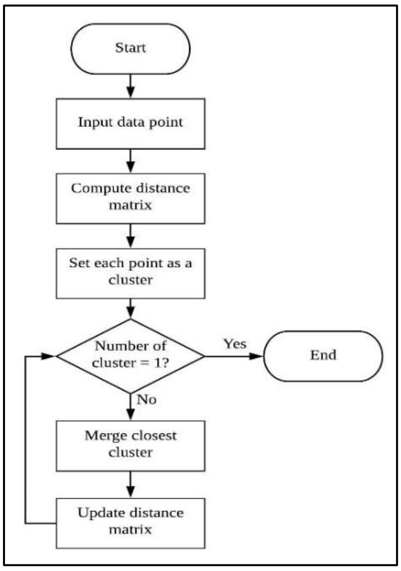


Figure 3.6: Flow Chart for Agglomerative Clustering Algorithm.

3.3.5.3 Fuzzy C-mean Clustering

The Fuzzy C-mean clustering is like K-means clustering but each data point has different membership coefficient of several clusters (Kemal, 2018). The membership coefficient is varying from zero to one. To conduct Fuzzy C-mean clustering, the number of clusters were required to define. The coefficients were assigned randomly to each data point for being in the clusters. The centroid of each cluster was determined. Then, the coefficient of each data point of being in the cluster was computed. These steps were repeated until maximum number of iterations was achieved. The steps were summarised in Figure 3.7.

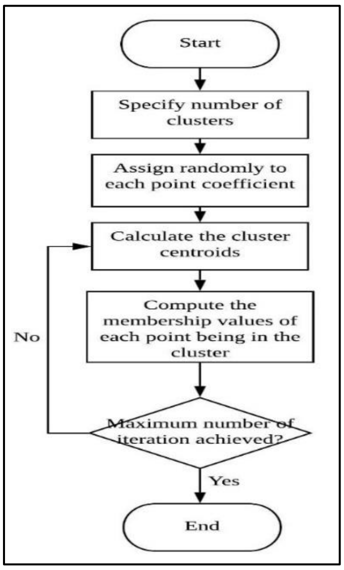

Start
Specify number of clusters
Assign randomly to each point coefficient
Calculate the cluster centroids
Compute the membership values of each point being in the cluster
No
Maximum number of iteration achieved?
Yes
End

Figure 3.7: Flow Chart for Fuzzy C-means Clustering Algorithm.

3.3.6 Cluster Evaluation Methods

The purpose of evaluation is to examine how well the clustered results obtained from different clustering methods. This evaluation checks whether the cluster is well-separated from the other clusters (Malarvizhi and Ravichandran, 2018). In this case, external clustering validation was not applied because no true class label existed in this dataset. Only internal clustering validation included range, Silhouette Coefficient and Davies Bouldin score were computed.

3.3.6.1 Range

The range is referring to maximum cluster size difference among the generated cluster. This was obtained by subtracting the number of data point between largest and smallest generated cluster. Lower value of the range is desired because it indicates all the clusters have almost similar number of data points. Therefore, the clusters are more balance with each other. The information retrieved from such cluster is more accurate and valuable.

3.3.6.2 Silhouette Coefficient

Silhouette coefficient was used to analyse and appreciate the difference between the resulting clusters. This method can determine how similar of each entity in a cluster is to entity in another cluster. The silhouette value is between - 1 and + 1. The value of + 1 implies the right clustering of objects, while the value of - 1 means that items are not correctly clustered (Ogbuabor and F. N, 2018).

In order to obtain the silhouette value, the mean intra cluster distance was calculated. After this, the nearest cluster distance from next closest cluster was obtained. The metrics then were used to compute the silhouette value by equation 3.4. The clustering performance was evaluated by the value.

$$S_i = \frac{(b_i - a_i)}{\max(a_i, b_i)} \tag{3.4}$$

where

a = average dissimilarity inside cluster

b = average dissimilarity to neighbour cluster

3.3.6.3 Davies Bouldin Score

The score is defined as each cluster's average similarity measure with its most similar cluster, where similarity is the ratio between within-cluster distances and between-cluster distances. Therefore, clusters that are more distant and less dispersed will result in a higher score. The minimum score is zero and better clustering result will indicate a lower value. The score was computed by using the equation 3.5 (Drakos, 2020).

$$D_{ij} = \frac{(d_i - d_j)}{d_{ij}} \tag{3.5}$$

where

d_i = average distance between every data point in cluster i and its centroid

d_j = average distance between every data point in cluster j and its centroid

d_{ij} = Euclidean distance between the centroids of the two clusters

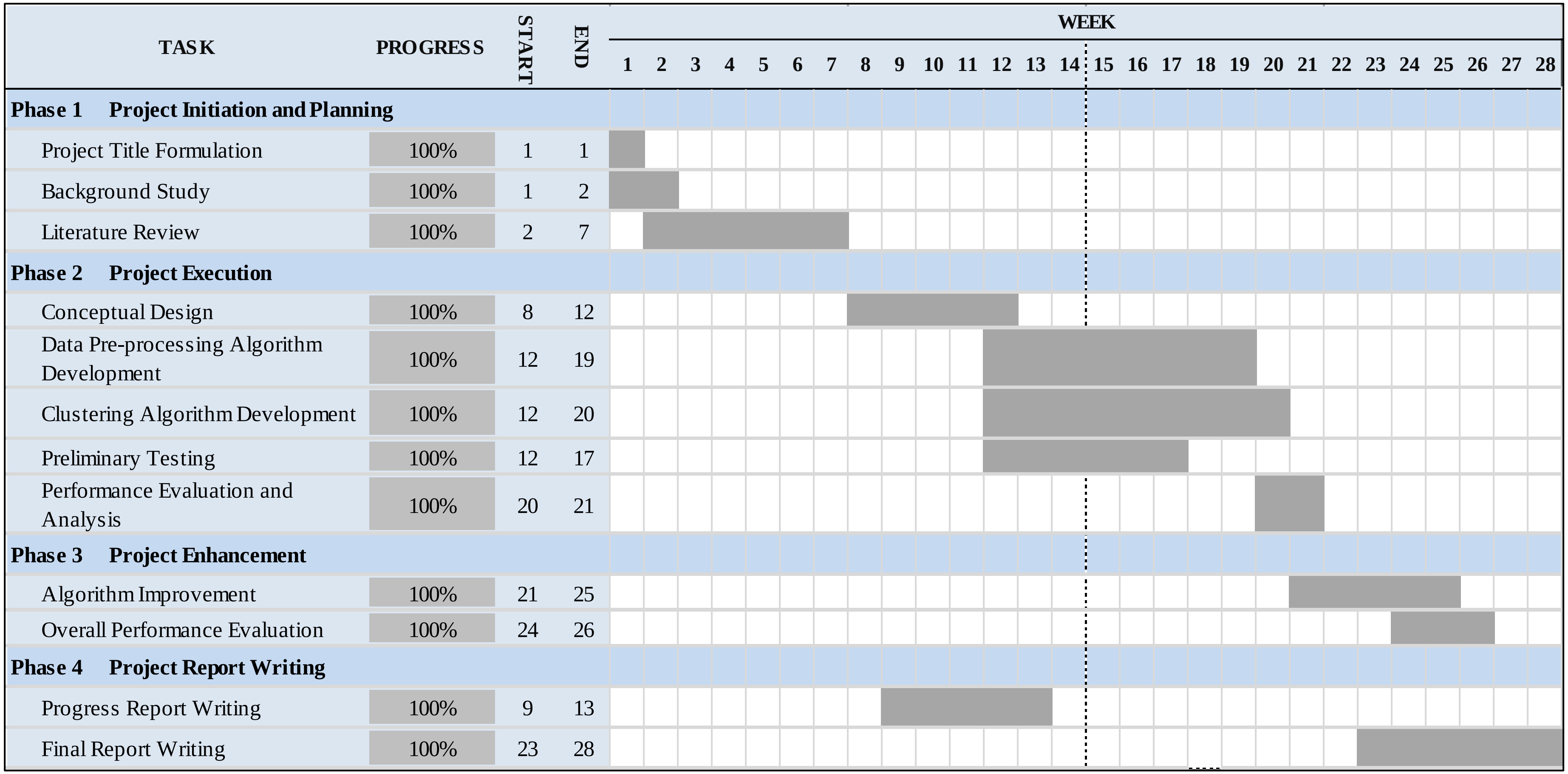
3.3.7 Characteristic Interpretation

The fall risk was calculated for each clustered group by dividing the number of fallers to total number of faller and non-faller within the group. The odd ratio (OR) was also computed by dividing the group fall risk to overall fall risk. Besides that, the characteristic of each group was indicated by computing the mean and standard deviation of each selected variable. The median and interquartile range were computed if the variable was not normally distributed.

After this, SPSS software was used to conduct Kruskal-Wallis H test and Dunn's test. Kruskal-Wallis H test is a multiple comparison test to evaluate whether there is a difference between groups. The null hypothesis states that there is no difference between group. If the obtained p-value is less than 0.05, the null hypothesis was rejected. Then, Dunn's test was used as post-hoc test to perform multiple pairwise comparison. The procedure is similar as Kruskal test.

3.4 Project Planning

A meticulous plan was developed by considering resources and time. This project had not involved any cost because only software was used. Figure 3.8 shows the Gantt chart for this project. All the tasks were conducted successfully and completed on timeline as shown in Figure 3.9.



TASK	PROGRESS	START	END
Phase 1 Project Initiation and Planning			
Project Title Formulation	100%	1	1
Background Study	100%	1	2
Literature Review	100%	2	7
Phase 2 Project Execution			
Conceptual Design	100%	8	12
Data Pre-processing Algorithm Development	100%	12	19
Clustering Algorithm Development	100%	12	20
Preliminary Testing	100%	12	17
Performance Evaluation and Analysis	100%	20	21
Phase 3 Project Enhancement			
Algorithm Improvement	100%	21	25
Overall Performance Evaluation	100%	24	26
Phase 4 Project Report Writing			
Progress Report Writing	100%	9	13
Final Report Writing	100%	23	28

Figure 3.8: FYP Gantt Chart

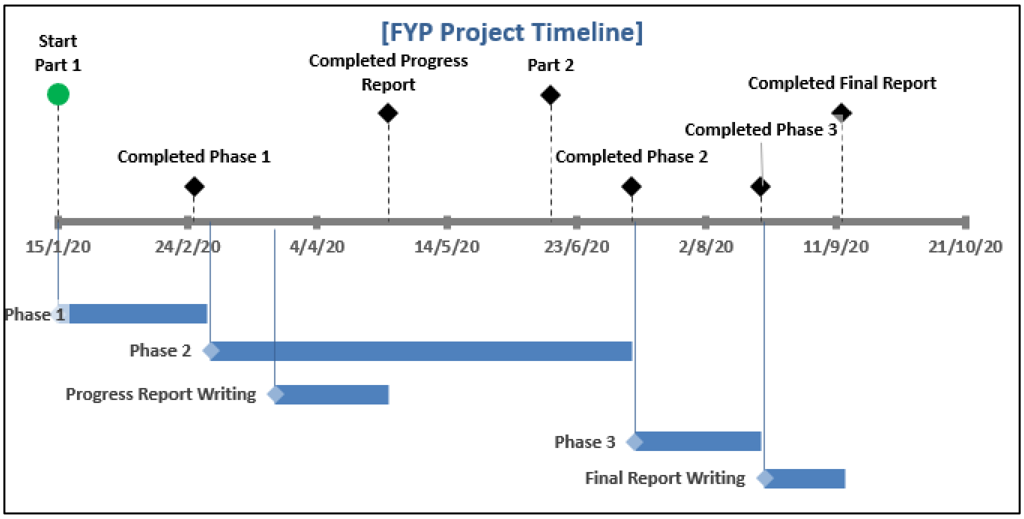


Figure 3.9: FYP Project Timeline.

3.5 Problems Encountered and Solutions

There were some problems encountered when completing this project. First, there is no best way to deal with missing data. The two common ways in this case are exclusion and imputation. In imputation method, the missing value can be replaced by mean or median. However, those imputed values are predicted from other values in dataset. This can cause the clustering result to be biased because of misleading data point. Therefore, exclusion method was chosen. although a portion of data were removed. The missing ratio in this dataset was small so it still offers a complete and true data for clustering.

Besides that, the distribution of data for some variables are not normal. Some of the statistical test such as ANOVA and independent T-test required the normality assumption. To deal with this problem, non-parametric tests (Mann-Whitney U test, Spearman correlation and Kruskal-Wallis H Test) were introduced. It is distribution-free tests and assess the group median instead of group means. In other word, it doesn’t assume data follow a specific distribution.

Moreover, the number of variables related to cardiovascular variability are huge in this dataset. Most of them are carry similar information but in different representation. This may cause redundant information. Therefore, random forest classifier model was used to evaluate the feature importance of each variable. The variable that have highest relative score was chosen for further process.

The data maybe non-linear due to the variables. It will affect the performance of methods like PCA that required linearity of data. To solve this, t-SNE which is non-linear method was conducted. The difference between the result of PCA and t-SNE were analysed.

3.6 Summary

The software used here were Spyder and SPSS software. The relevant information regarding the input dataset was discussed. A clustering algorithm included data pre-processing, feature selection, feature extraction, clustering techniques and characteristic interpretation was proposed. Missing data handling, feature categorization and normality testing were implemented in data pre-processing. Hypothesis testing, high correlation filter and feature importance were chosen as feature selection methods while PCA and t-SNE were used for feature extraction. Furthermore, K-means clustering, Hierarchical (Agglomerative) clustering and Fuzzy c-means clustering were selected as clustering methods. Relevant steps and flow charts were explained in detail. Besides that, Gantt chart and project timeline were included to illustrate the planned tasks. In last section, the encountered problem and proposed solutions were discussed.

CHAPTER 4

RESULTS AND DISCUSSION

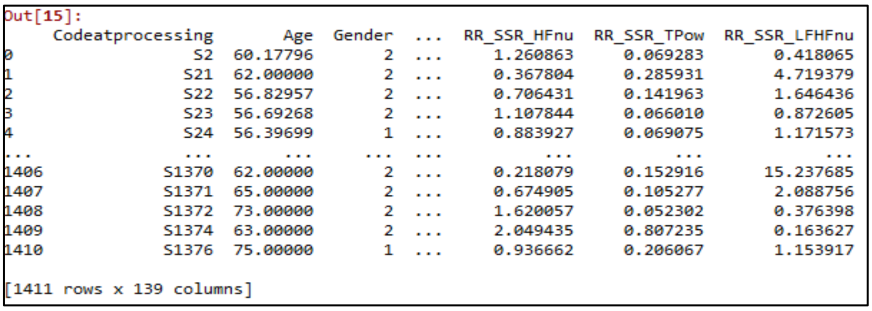
4.1 Introduction

This chapter was presenting and discussing the results obtained by proposed fall risk algorithm. It consists of several parts included data pre-processing, feature selection, feature extraction, clustering and characteristic interpretation.

4.2 Data Pre-processing

The original dataset which consists of one thousand four hundred eleven subjects and one hundred and thirty-nine variables were loaded into Pandas DataFrame as shown in Figure 4.1. The target variable 'Fall_questionnaire' which indicates the identity of faller or non-faller was removed from data frame. On the other hand, a new variable that represents the dominant hand grip strength of subject was inserted. Besides that, all the zero-value inside data frame was replace to NaN (missing data representation).

The new variable is suggested to be included because there is difference of manipulation speed between dominant hand and non-dominant hands (Cary and Dipcot, 2003). According to the research, it stated that non-dominant hand was manipulated objects slower compared to dominant hand. In addition, Petersen et al. (1989) reported that dominant hand grip strength of right-handed person was 10% stronger compared to non-dominant hand. As for zero-value, it could affect the result of statistical analysis. Therefore, replace it to NaN value can ensure the reliability of arithmetic result regardless of the operation.



```
Out[15]:
    Codeatprocessing        Age  Gender  ...  RR_SSR_HFnu  RR_SSR_TPow  RR_SSR_LFHFnu
0                 S2   60.17796       2  ...     1.260863     0.069283       0.418065
1                S21   62.00000       2  ...     0.367804     0.285931       4.719379
2                S22   56.82957       2  ...     0.706431     0.141963       1.646436
3                S23   56.69268       2  ...     1.107844     0.066010       0.872605
4                S24   56.39699       1  ...     0.883927     0.069075       1.171573
...              ...        ...     ...  ...          ...          ...            ...
1406           S1370   62.00000       2  ...     0.218079     0.152916      15.237685
1407           S1371   65.00000       2  ...     0.674905     0.105277       2.088756
1408           S1372   73.00000       2  ...     1.620057     0.052302       0.376398
1409           S1374   63.00000       2  ...     2.049435     0.807235       0.163627
1410           S1376   75.00000       1  ...     0.936662     0.206067       1.153917

[1411 rows x 139 columns]
```

Figure 4.1: Summary of Data Set (1411 Subjects, 139 Variables).

4.2.1 Missing Data Handling

The missing value percentage was computed for all variables. From the result obtained, 133 out of 139 variables had missing data with different percentages. Among of them, the variables ‘TBRSSt’ and ‘R_TBRS’ had missing data that exceed 10% so it was removed out from data frame.

Bennett (2001) reported that analysis was likely to bias when the percentage of missing data was above 10%. The statistical result determined by such variable may not be correctly estimated. Therefore, 10% was set as threshold in this case. For those variables that having lower percentage of missing data, it can still be analysed after dropping those subject with missing data.

4.2.2 Feature Categorization

The remaining variables were classified into informative, numerical and categorical group. The informative group consists of five variable which described the ID in MATLAB, dominant hand used, report of clinical falls, condition of continuous blood pressure and MoCA questionnaire language. There were only variables ‘Gender’ and ‘Ethnicity’ for the categorical group, whereas one hundred and thirty variables were categorised into numerical group. The numerical group was further classified into cardiovascular group (systolic blood pressure variability, RR variability, etc.) and common group (age, height, etc.). The numbers of cardiovascular variables were one hundred and sixteen whereas the numbers of common variables were only fourteen.

Informative variables are irrelevant with fall risk analysis. It is subjective and challenging to be evaluated especially for comparison. Therefore, it should manually be filtered out because only provides information instead of valuable metrics. Besides that, cardiovascular variables occupy almost 89% among of the numerical variables. Although the numbers of cardiovascular variables are huge, most of it are representing similar feature but in different method of computation. To illustrate this, ‘SBP_SDsp’ and ‘SBP_CVsp’ are computing systolic blood pressure variation but one with standard deviation and another one with coefficient of variation. On the other hand, the number of common and categorical variables are lower but all of it are unique in this data set. In short, this stage simplifies the dataset into understandable categories so the further steps can be implemented based on each category.

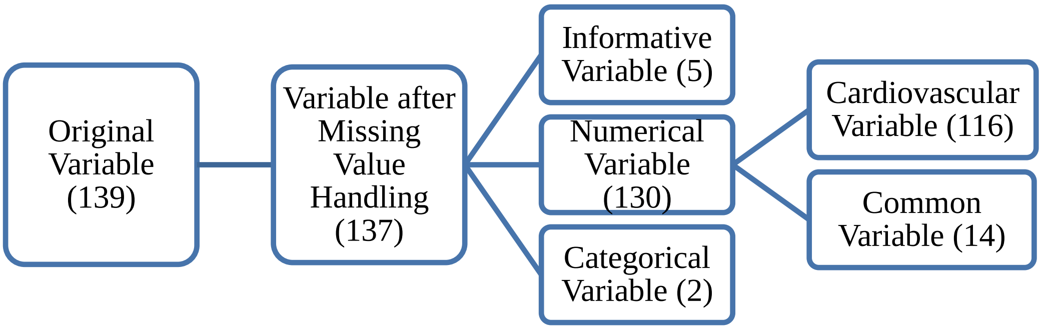


Figure 4.2: Number and Category of Variables.

4.2.3 Normality Testing

The result of Shapiro–Wilk test indicated only one variable, ‘DBP_sp’ had normal distribution. All other numeric variables had failed this normality test.

Apart from normality test, histogram also used to check the distribution of data. However, the graphical methods may not enough to provide conclusive evidence compare to normality test (Razali and Wah, 2011). Among all normality tests, Razali and Wah (2011) reported that Shapiro-Wilk test had the best performance. Park (2016) also recommended this test when sample size was less than two thousand. Therefore, Shapiro-Wilk test is the most suitable normality testing method in this data set.

The causes for non-normality are including outliers and underlying distribution. The presence of outliers will lead the data to skew. In this dataset, most of the variables such as height and hand grip strength do not have defined ranges, so the large percentage of outliers are an issue. There may also situation like multiple normal distribution combined to multimodal distribution. In this case, some variables have significant difference between faller and non-faller group so it may cause the data to give the appearance of bimodal data.

There is difference between measurements in normally distributed and non-normally distributed variables as shown in Table 4.1. Therefore, all variables except ‘DBP_sp’ were analysed by median, interquartile range and non-parametric test in further stages.

Table 4.1: Measurements Between Normally and Non-Normally Distributed Data.

Measurement	Data	
	Normally Distributed	**Non-Normally Distributed**
Measure of Central Tendency	Mean	Median
Measure of Spread of Data	Standard Deviation	Interquartile Range
Statistical Analysis	Parametric (Independent T-test, ANOVA test, Pearson Correlation)	Nonparametric (Mann-Whitney test, Kruskal-Wallis test, Spearman Correlation)

4.3 Feature Selection

After data pre-processing, the number of variables at this stage were one hundred and thirty-two. In order to eliminate the duplication in cardiovascular group, only the most representative or relevant variable among systolic blood pressure (SBP) variables, diastolic blood pressure (DBP) variables and RR interval (RR) variables will be chosen. In other words, only three out of one hundred and sixteen cardiovascular variables were selected in the end. As for common variables, it will be selected if pass all three feature selection methods.

4.3.1 Hypothesis Testing

The hypothesis test consists of independent T-test (for normal distributed numeric variables), Mann-Whitney test (for non-normal distributed numeric variables) and Chi-squared test (for categorical variables). After these tests, the number of common variables had reduced from fourteen to nine whereas the number of cardiovascular variables had reduced from one hundred and sixteen to forty-seven. Apart from that, the number of categorical variables were remained as two.

Hypothesis tests are used to determine the variable that have significance difference between faller group and non-faller group. By analysing the selected variables, the characteristic between faller and non-faller groups can be compared. This result has 95% level of confidence because the alpha value was set to 0.05.

4.3.2 High Correlation Filter

The correlation test consists of Spearman correlation (for numerical variables) and Cramer V (for categorical variables). After this test, the number of common variables had decreased from nine to only six. According to Figure 4.3, the variable 'Dominant_Hand_grip' were strongly correlated with 'RightHandAverage' and 'LeftHandAverage'. Besides that, variable 'TUGs' was strongly correlated with 'Frailty15ft'. Therefore, variables 'RightHandAverage', 'LeftHandAverage' and 'Frailty15ft' were filtered out. Number of categorical variables remained as two after Cramer V correlation test.

The threshold set is 0.8 and any value above is considering as strongly correlated. Those strongly correlated variables are redundant so keep only one of them is enough (Molala,2019). Dominant hand grip strength and TUG test are supported by literature review, so it was preferable. This test is not conducted for cardiovascular variables because those variables are very similar with each other.

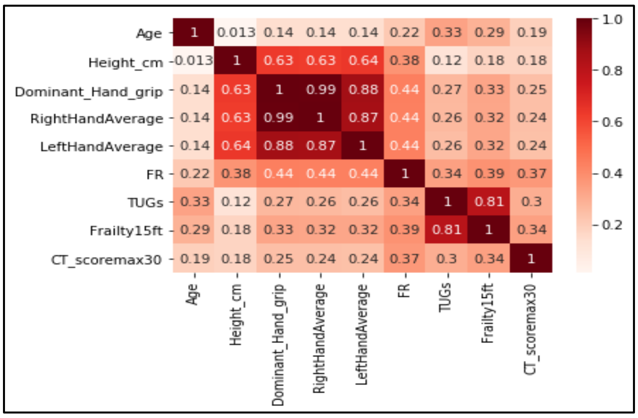


Figure 4.3: Heat Map for Correlation Test.

4.3.3 Feature Importance

The feature importance method was used to select the top representation for cardiovascular variables. Before this, the forty-seven cardiovascular variables were grouped into 'SBP', 'DBP' and 'RR' group. The feature importance of all 'RR' variables were computed. According to Figure 4.4 (left), 'RR_SSR_ARV' (standing to supine ratio of RR variability computed in average real variability) had the highest relative score. Thus, the similar representation was also chosen for ‘SBP’ and ‘DBP’ variable which were ‘SBP_SSR_ARV’ and ‘DBP_SSR_ARV’.

After combined the three selected cardiovascular variables and other eight selected variables after correlation filter, the feature importance of each variable was computed again. According to Figure 4.4 (right), the variable ‘Gender’ and ‘Ethnicity’ had the value less than threshold (0.05) so it was filtered out.

The scores of input variables in a predictive model shows its relative importance when making prediction. The random forest classifier was chosen as predictive model because it provides the method of mean decrease impurity. Gini impurity is measuring probability of incorrect classification for training data set. Random forest classifier computed the impurity decreased from each variable and rank them according to this. The relative scores can highlight the most relevant cardiovascular variable (highest score). Also, it acts as final interpretation to examine which variable should be included. After dropping the variable ‘Gender’ and ‘Ethnicity’, the remaining variables still can achieve 95.39% cumulative importance. Therefore, these two variables considered as low importance features.

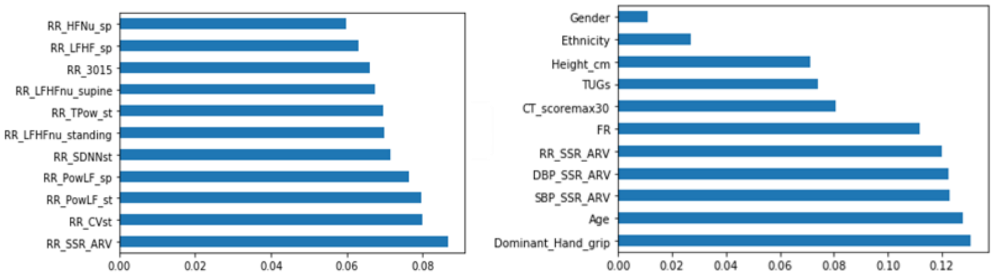


Figure 4.4: Feature Importance for Cardiovascular Variables in ‘RR’ Group (Left) and Selected Variables after Combining Cardiovascular and Common Variables (Right)

4.3.4 Discussion of Feature Selection Methods

As summarised in Table 4.2, the hypothesis testing methods are successfully reducing one hundred and thirty-two variables to only fifty-eight variables, followed by correlation method (fifty-five variables) then feature importance method (nine variables). In comparison, the filter methods (hypothesis testing and correlation method) is less computationally expensive than embedded method (feature importance method) (Shetye, 2019). Therefore, the filter methods were used to exclude those irrelevant variables first. The embedded method was then applied when the number of variables were lesser. The result obtained from this combination was satisfied.

The final selected variables were carrying most of the information from original full variables. It reduces the algorithm complexity and noise caused by misleading data. Furthermore, most of the selected variables are major fall risk factors that identified previously in Literature Review. Thus, the relationship between these variables and fall risk can be discovered.

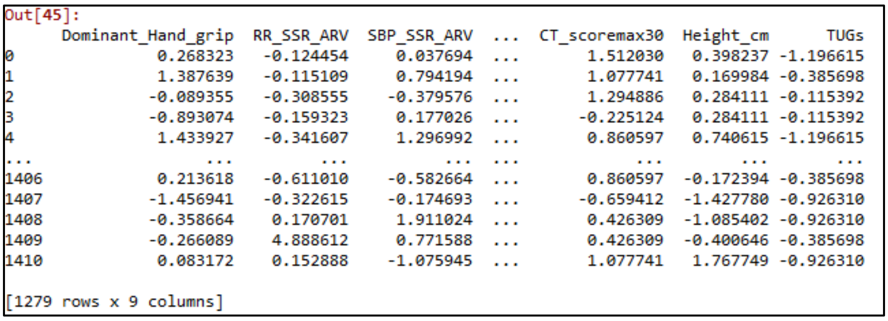
Table 4.2: Number of Variables After Each Stage in Feature Selection.

Stage	Number of Variables
After Pre-processing	132
After Hypothesis Testing	58
After Correlation Filter	55
After Feature Importance	9

4.4 Feature Extraction

The nine selected variables were further proceeded to this stage. In order to ensure the accuracy of dataset, all the subjects that contain missing data in selected variables were removed. Instead of one thousand four hundred eleven subjects, one thousand two hundred seventy-nine subjects were remained. The missing ratio was 9.36% which less than 10% so it still can be accepted.

Besides that, all the variables were standardized so it had centred around zero with standard deviation of one. This is important when comparing different variables in different units. Without this step, the variables with large unit may dominate and mislead the result of feature extraction. Figure 4.5 indicated the data after Z-score standardization. Although the data were rescaled, the distribution of data doesn’t change after this (González, 2018).



```
Out[45]:
      Dominant_Hand_grip  RR_SSR_ARV  SBP_SSR_ARV  ...  CT_scoremax30  Height_cm      TUGs
0               0.268323   -0.124454     0.037694  ...       1.512030   0.398237 -1.196615
1               1.387639   -0.115109     0.794194  ...       1.077741   0.169984 -0.385698
2              -0.089355   -0.308555    -0.379576  ...       1.294886   0.284111 -0.115392
3              -0.893074   -0.159323     0.177026  ...      -0.225124   0.284111 -0.115392
4               1.433927   -0.341607     1.296992  ...       0.860597   0.740615 -1.196615
...                  ...         ...          ...  ...            ...        ...       ...
1406            0.213618   -0.611010    -0.582664  ...       0.860597  -0.172394 -0.385698
1407           -1.456941   -0.322615    -0.174693  ...      -0.659412  -1.427780 -0.926310
1408           -0.358664    0.170701     1.911024  ...       0.426309  -1.085402 -0.926310
1409           -0.266089    4.888612     0.771588  ...       0.426309  -0.400646 -0.385698
1410            0.083172    0.152888    -1.075945  ...       1.077741   1.767749 -0.926310

[1279 rows x 9 columns]
```

Figure 4.5: Summary of Data Set after Z-score standardization (1279 Subjects, 9 Variables).

4.4.1 Principal Component Analysis (PCA)

PCA was conducted by Scikit-Learn python built in library. The parameter was set as default. All nine variables were computed as principal components. The variance caused by each component were illustrated at Figure 4.6. According to result, the first to ninth principal component was representing 28.25%, 14.33%, 12.73%, 10.52%, 9.49%, 8.24%, 6.52%, 5.87% and 4.05%. In order to obtain at least 60% variance ratio, first four components (65.83% cumulative variance) were chosen. The scatter plot of data point based on first two principal components was shown in Figure 4.7.

The criteria for choosing number of principal components are performance and cumulative variance ratio. The general rule of thumb is to take the number of key components that lead to significant variance and ignore those with declining variance returns (Malik, 2018). In this case, the variance ratio is decreasing along the components. However, the change of variance ratio is small after second principal components. It doesn’t show component with diminishing variance. As for performance, the accuracy is similar although number of chosen principal components are changing. In other words, increase number of components doesn’t improve the accuracy of classifier. According to Hair et al (2012), it reported that the appropriate variance explained for the model to be valid in factor analysis was 60%. Therefore, 60% is set as threshold to decide the number of principal components.

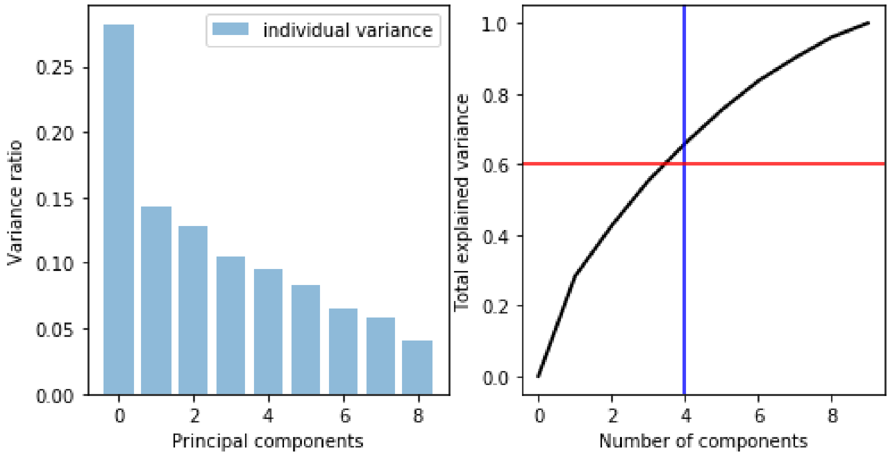


Figure 4.6: The Variance Ratio Explained by Each Principal Components (Left) and Total Explained Variance by Number of Principal Components (Right).

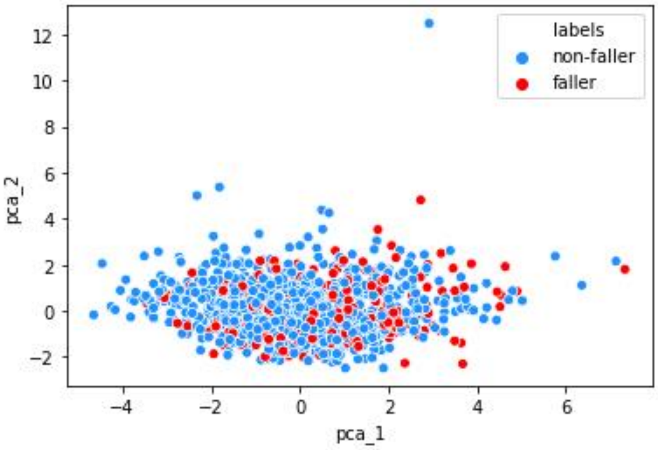


Figure 4.7: 2D Representation of 1st and 2nd Principal Components.

4.4.2 t-Distributed Stochastic Neighbour Embedding (t-SNE)

t-SNE was used to create two-dimensional representation from original nine dimensions. In t-SNE, tuneable parameters have complex effect on the generated result. After several run, the parameters were set as perplexity (180), early exaggeration factor (12), learning rate (200). The scatterplot for the result was shown in Figure 4.8 (right).

The perplexity describes the effective number of neighbours that used to compute defined structure of clusters. Larger perplexities contribute to largest nearest neighbours and less sensitive to small structures. In contrary, a lower perplexity perceives a smaller number of neighbours and hence neglects more global information in favour of the local neighbourhood. When the size of data set is larger, more points are necessary to accomplish a reasonable sampling of the local area, and thus greater perplexities may be expected (Scikit Learn, 2014). As indicated in Figure 4.8 (left), local variations are dominate with only perplexity of ten. In comparison, perplexity of one hundred and eighty loses some fine detail but retain larger and meaningful structure together.

Other than perplexity parameter, the early exaggeration factor handles how close natural structures are in the original space and how much gap is between them. The distance between existing clusters would be greater in the embedded area if larger values are provided. Apart from that, most points are clustered in a compact cloud of little outliers if the learning rate is set too small. In short, t-SNE is stochastic method that will produce different results based on hyperparameters (Wattenberg et al., 2016). In this case, the selected hyperparameters can extract the pattern inside data according to similarities of data points.

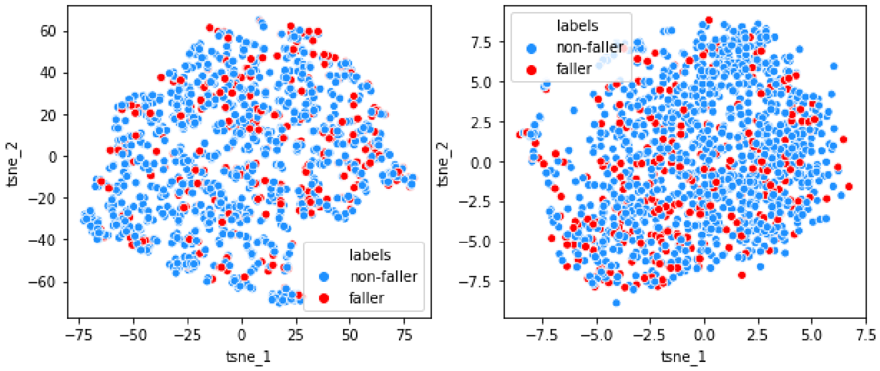


Figure 4.8: 2D Representation for t-SNE with Perplexity 10 (Left) and 180 (Right).

4.4.3 Discussion of Feature Extraction Methods

PCA is a linear feature extraction method that aims to optimize variation and maintains large pairwise distances. Data points that different from original data set will far away from each other after PCA transformation. However, data set may have the manifold structure instead of linear. In this case, PCA may not be able to interpret the data efficiently. From the result obtained, the variance ratio is not significant decrease after second principal component. All this may due to the nonlinearity of data set.

Unlike PCA, t-SNE achieves a better visualization result which the data points are spread evenly. The data points of faller are concentrated at one side. This may because t-SNE is a nonlinear technique that preserve local similarities. Linear dimensional reduction relies on putting dissimilar data points further away in a lower dimensional representation. Nevertheless, in order to illustrate high-dimensional data on low-dimensional, non-linear manifolds, it is important that similar data points be expressed close together, which is something t-SNE does not PCA.

Other than linearity, there are some key differences between PCA and t-SNE. To illustrate this, PCA is mathematical technique that look for axis that explain highest variance but t-SNE is probabilistic method that attempt to minimize the divergence between distribution. t-SNE is computationally expensive as it may take longer time than PCA, especially for large data size (Jaju, 2017).

In fact, t-SNE can work with both linear or nonlinear data sets and produces meaningful clustering (Bedre, 2020). Balamural and Melkumyan (2016) reported that t-SNE can retained local structure and revealed the global structure such as presence of clusters. This shown that t-SNE can work with clustering algorithm to produce a better result. Derksen (2016) reported that constructed PCA before t-SNE was able to improve the result when the number of variables were more than fifty. However, this method is not implemented here because the number of variables in this case have only nine.

4.5 Clustering Algorithm

After feature extraction, the data were successfully transformed to two dimensional for t-SNE and four dimensional for PCA. In this stage, the transformed data were further clustered into defined groups. The algorithm was implemented by using Sklearn library.

4.5.1 K-means Clustering

The appropriate number of clusters was evaluated by elbow method and Silhouette Coefficient before clustering. According to Figure 4.9(a), the sum square error (SSE) for both PCA and t-SNE transformed data were continued to decrease when the number of clusters increased. To illustrate this, the distance between point and closest centroids was decreased when the more centroids were exists. The elbow point for the t-SNE result was identified as four. However, it may difficult to choosing the elbow point of the PCA curve. Therefore, Silhouette Coefficient was also computed to observe the suitable number of clusters. Based on the Figure 4.9(b), the best choice of cluster number for PCA was five whereas for t-SNE was same as four. The clustering results were shown in Figure 4.9(c). As expected, the PCA data was clustered into five clusters whereas t-SNE data was clustered into four clusters.

The initial centroids were selected by using the (K-means++) to speed up the convergence. This method is recommended because it provides best initial points for k-means algorithm (Thakur, 2020). Besides that, the random state was set to zero to obtain reproducible result. According to the result, the clusters size formed from PCA data point are not balance especially for group three. Besides that, some data points seem overlap to another group. This may due to only two components are involved in visualization. Therefore, it cannot show the cluster result in full picture. On the other hand, the clustering result obtained from t-SNE is considered good. All groups are divided evenly and there are no overlapping points.

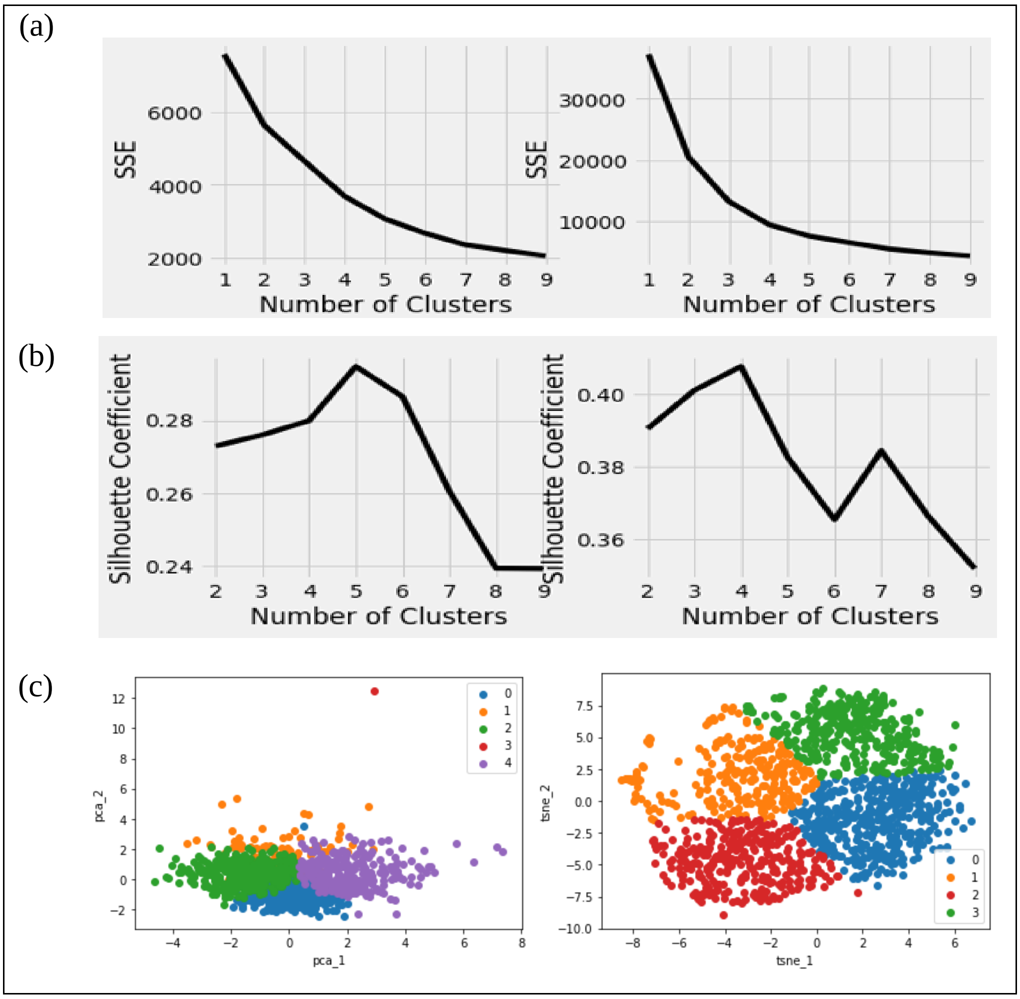


Figure 4.9: Analysis and Results in K-means Clustering Algorithm.

(Left) = PCA; (Right) = t-SNE

(a) Elbow Curve; (b) Silhouette Coefficient Curve; (c) Clustering Scatterplot

4.5.2 Hierarchical (Agglomerative) Clustering

According to Figure 4.10(a), the suitable number of clusters for both PCA and t-SNE data were four in hierarchical clustering. The hierarchical clustering was agglomerative based which points are merged as one moves up the hierarchy. Figure 4.10(b) illustrated the dendrogram that shows the hierarchical relationship between points. The linkage criterion chosen for this was ward which the variance of the clusters was minimized. As shown in Figure 4.10(c), the data point for both PCA and t-SNE were clustered into four groups.

From the dendrogram generated from PCA, it shows that there are three group when the Euclidean distance is forty. There is one point that did not merge with other as its geometric distance is far from other points. The Euclidean distance that used to separate four clusters for PCA is about thirty-seven while for t-SNE is about seventy to one hundred. It shows that the data point from PCA is more closely packed compared to t-SNE. In overall, the result obtained in this clustering algorithm is almost same with K-means clustering algorithm except the component numbers choose for PCA data.

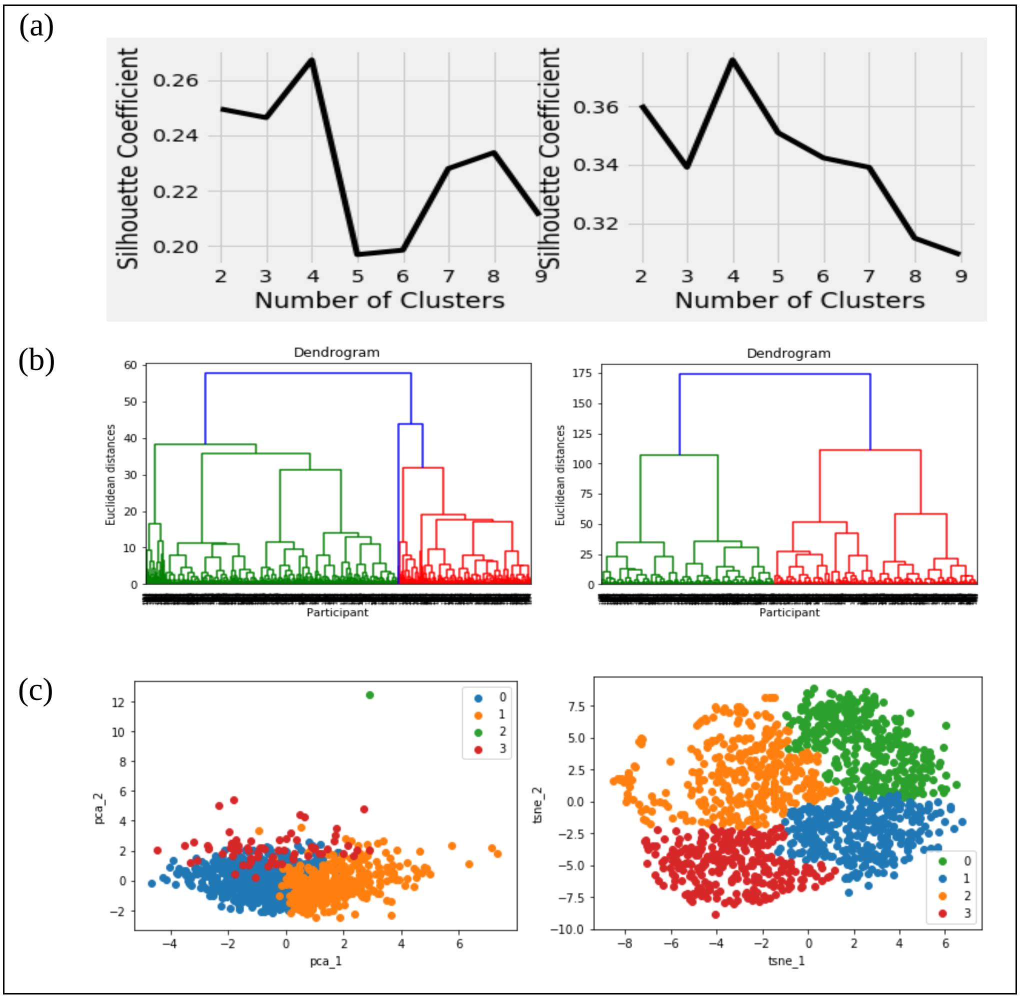


Figure 4.10: Analysis and Results in Hierarchical (Agglomerative) Clustering Algorithm.

(Left) = PCA; (Right) = t-SNE

(a) Silhouette Coefficient Curve; (b) Dendrogram; (c) Clustering Scatterplot

4.5.3 Fuzzy C-mean Clustering

According to Figure 4.11(a), the suitable number of cluster for PCA data was two whereas for t-SNE was four. After the clustering algorithm, both PCA and t-SNE data was assigned to its defined clusters as shown in Figure 4.11(b).

From the result, the PCA point was clustered into two distinguish groups. There was no overlapping point in this clustering algorithm compared to previous clustering algorithm because there were only involved two groups. As for t-SNE data, there was no big difference compared to result obtained from Hierarchical and k-mean clustering.

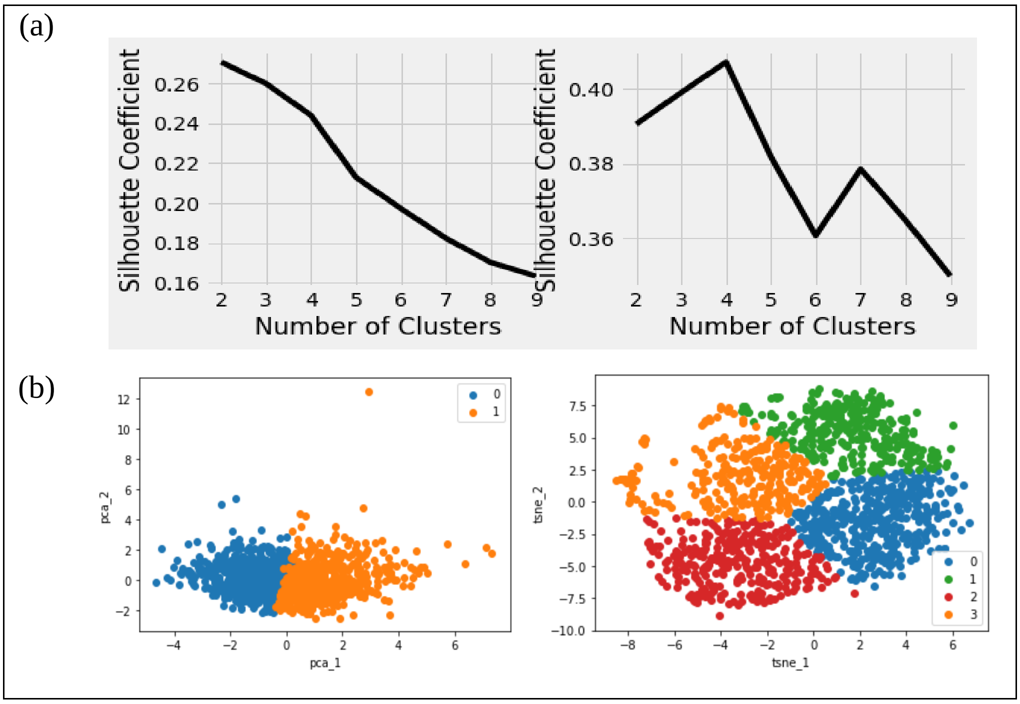


Figure 4.11: Analysis and Results in Fuzzy C-mean Clustering Algorithm. (Left) = PCA; (Right) = t-SNE

(a) Silhouette Coefficient Curve; (b) Clustering Scatterplot

4.5.4 Discussion of Clustering Methods

In short, PCA and t-SNE data input were successful clustered by each clustering algorithm. There were six different combinations of feature extraction and clustering methods. Range, Silhouette Coefficient and Davies-Bouldin score were computed as clustering validation methods to evaluate the performance of each combination. The range measured the maximum difference of clustered group size. If all the groups had a balance group size, the maximum difference would be zero. Silhouette Coefficient ranges from negative one to positive one, clusters which were dense and well separated will have higher score. As for Davies-Bouldin score, lower values indicate better clustering.

According to Table 4.3, combination of t-SNE and k-means was ranked number one among all six models. On the other hand, PCA and Hierarchical was the lowest performance combination because it had the most imbalance clusters and lowest Silhouette Coefficient. Although combination of PCA and Fuzzy C-means had the most balanced group size, its performance in Silhouette Coefficient and Davies-Bouldin score were not that good.

The number of clusters were not consistent in PCA input data. Different clustering algorithm were led to different optimised cluster number. This may because the data points are too closely packed with each other and the presence of outlier. Besides that, the size of cluster formed in PCA input data were significant different between each group except for Fuzzy C-means clustering. However, only two clusters were formed by Fuzzy C-means so it may not suitable to direct compared with other two algorithms that have more cluster. On the other hand, all clustering algorithm resulted similar cluster number in t-SNE input data. The difference in group size were almost similar for K-means and Fuzzy C-means clustering. Hierarchical clustering achieved more balance group size as it had the lower range.

In term of Silhouette Coefficient and Davies-Bouldin score, the PCA input data achieved lower performance if compared with t-SNE input data. This indicated that the intra-cluster similarity and inter-cluster differences were lower for clusters formed by PCA input data. Therefore, t-SNE resulted more validate and compact clusters thus the information retrieved was more accurate.

In overall, t-SNE input data achieved higher performance than PCA input data. Among the clustering techniques applied in t-SNE input data, the performance of Hierarchical clustering technique was slightly lower. Apart from that, the K-means and Fuzzy C-means clustering had almost similar performance. However, Fuzzy C-means produced more compact cluster whereas K-means clustering yield more distinct clusters. This result is related to finding from Panda et al. (2012). In addition, Cebeci and Yildiz (2015) reported that K-means clustering was outperforming to Fuzzy C-means clustering in term of computing time. By considering all of this, the combination of t-SNE for feature extraction and K-means clustering algorithm achieved the best clustering performance in MELoR dataset. Its clustering result was further analysed in next stage.

Table 4.3: The Comparison of Performance Among Six Different Combination.

Combination	Cluster	Clustering Validation Methods			
		Range	Silhouette Coefficient	Davies Bouldin Score	Rank
PCA and k-means	5	545 (5)	0.29 (4)	0.88 (4)	5
PCA and Hierarchical	4	771 (6)	0.26 (6)	0.96 (5)	6
PCA and Fuzzy C-means	2	11 (1)	0.27 (5)	1.37 (6)	4
t-SNE and k-means	4	149 (4)	0.41 (1)	0.78 (1)	1
t-SNE and Hierarchical	4	82 (2)	0.38 (3)	0.85 (3)	3
t-SNE and Fuzzy C-means	4	129 (3)	0.40 (2)	0.79 (2)	2

Note: The rank is evaluated by least stack ranking from all three clustering validation methods.

$^{1, 2, 3, 4, 5, 6}$ Ordinal Ranks 1 to 6

4.6 Characteristic Interpretation

After the t-SNE for feature extraction and K-means clustering algorithm, one thousand two hundred seventy-nine subjects were clustered into four groups. As shown in Figure 4.12, each cluster had occupied its specific region. According to the trend, the fall risk would likely to increase when the data point moves downward or leftward. To illustrate this, Low fall risk group was located at top right side while High fall risk group located at bottom left side. The Intermediate A and B fall risk group were located at middle.

According to Table 4.4, the clusters size for each group (from Low to High) was three hundred thirteen, four hundred six, two hundred fifty-seven and three hundred three respectively. The group size is almost balance especially for Low and High fall risk group. The percentage of faller to non-faller for each group (from Low to High) were 13%, 19%, 21% and 31%. The overall fall risk for this dataset was 21% where two hundred sixty-four faller among all one thousand two hundred seventy-nine subjects. In comparison, Low fall risk group had odd ratio of 0.62 whereas High fall risk group had odd ratio of 1.48. This reveals that older people clustered at Low fall risk group are exposed to almost 40% lower risk of falls among overall older cohort in the dataset. In contrast, almost 50% higher risk of falls are exposed for High fall risk group.

In order to evaluate the characteristic inside each group, the average value of selected features was computed. Body mass index, waist-to-hip ratio, and gender were included in interpretation because these variables were identified as relevant fall risk factors in literature review. According to Table 4.5, a total of twelve variables, which can further assign into six categories were analysed. Since all the variables were not normal distributed, the median and interquartile range were applied instead of mean and standard deviation. Apart from that, gender was a categorical variable, so the percentage of female was used to indicate the effect toward fall risk.

According to Table 4.6, all the variables were significant difference among the groups in Kruskal-Wallis H test. For the Dunn's test, variables HGS, MoCA and Age were statistically significant differences between all pairwise groups. Besides that, only one pairwise group of variables TUG, FR, DBP_SSR_ARV and Height was no difference. Lastly, variables SBP_SSR_ARV, RR_SSR_ARV, BMI and WHR had at least two pairwise group that was no significant differences.

In overall, all the selected variables from clustering algorithm were significant difference between Low and High fall risk group. Therefore, it is validated to compare the characteristic between these two groups. After analysis, characteristics of faller were analysed as older, slower or imbalanced gait, weaker muscle strength, with cardiovascular disorder and cognitive impairment. In next subsection, the result was then discussed with previous studies.

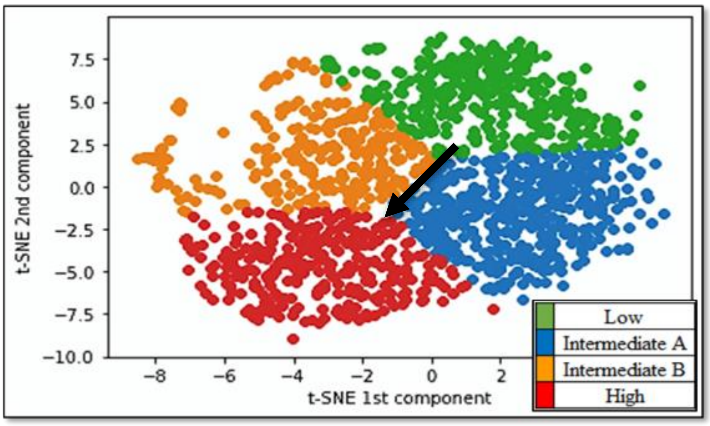


Figure 4.12: Different Fall Risk Group. Arrow = Trend of Increased Fall Risk

Table 4.4: Summary of Each Generated Cluster.

Cluster Colour	Centroid	Cluster Size	Fall Risk	Odds Ratio
Green	[1.58, 5.42]	313	Low (13%)	0.62
Blue	[2.53, -1.43]	406	Intermediate A (19%)	0.90
Orange	[-3.66, 2.06]	257	Intermediate B (21%)	1.00
Red	[-3.26, -4.73]	303	High (31%)	1.48

Table 4.5: Variables That Chosen for Characteristic Interpretation.

Fall Risk Factors	Variables
Gait and Balance	Timed Up and Go Test (TUG)
	Functional Reach (FR)
Muscle Strength	Dominant Hand Grip Strength (HGS)
Cardiovascular Disorder (Standing to Supine Ratio of Variation Computed with Average Real Variability)	Systolic Blood Pressure (SBP_SSR_ARV)
	Diastolic Blood Pressure (DBP_SSR_ARV)
	RR Interval (RR_SSR_ARV)
Cognitive Impairment	Montreal Cognitive Assessment (MoCA)
Demographic	Age
	Gender *
Other	Height
	Body Mass Index (BMI) *
	Waist to Hip Ratio (WHR) *

* Additional variables that included for characteristic interpretation, but not selected in clustering algorithm.

Table 4.6: Characteristic of Clustered Groups.

Variables	Fall Risk Group				p-value
	Low	Intermediate A	Intermediate B	High	
TUG (s)	11.0 ± 3.00 [a,b,c]	11.0 ± 2.00 [a,e]	12.0 ± 3.00 [b,f]	15.0 ± 6.00 [c,e,f]	***
FR (cm)	32.0 ± 8.00 [b,c]	26.0 ± 7.00 [d,e]	26.0 ± 8.00 [b,d,f]	19.0 ± 7.00 [c,e,f]	***
HGS (kg)	32.0 ± 9.4 [a,b,c]	19.8 ± 6.43 [a,d,e]	24.8 ± 7.83 [b,d,f]	17.3 ± 7.03 [c,e,f]	***
SBP_SSR_ARV	1.3 ± 0.75 [c]	1.2 ± 0.56 [d,e]	1.2 ± 0.68 [d,f]	1.0 ± 0.52 [c,e,f]	***
DBP_SSR_ARV	1.2 ± 0.60 [a,c]	1.1 ± 0.53 [a,d,e]	1.2 ± 0.70 [d,f]	1.0 ± 0.45 [c,e,f]	***
RR_SSR_ARV	0.8 ± 0.41 [a,b,c]	0.8 ± 0.48 [a]	1.0 ± 0.99 [b]	0.8 ± 0.51 [c]	***
MoCA	26.0 ± 4.00 [a,b,c]	26.0 ± 4.00 [a,d,e]	22.0 ± 6.00 [b,d,f]	18.0 ± 7.00 [c,e,f]	***
Age	65.7 ± 8.40 [a,b,c]	64.2 ± 9.07 [a,d,e]	71.6 ± 8.50 [b,d,f]	73.1 ± 10.02 [c,e,f]	***
Gender (Female)	63 (0.20)	378 (0.93)	69 (0.27)	221 (0.73)	***
Height (cm)	166.0 ± 10.00 [a,b,c]	153.0 ± 6.00 [a,d]	162.0 ± 8.00 [b,d,f]	153.0 ± 10.00 [c,f]	***
BMI (kg/m^2)	24.5 ± 4.51	24.4 ± 6.14 [e]	24.7 ± 5.10 [f]	25.6 ± 6.67 [e,f]	**
WHR	0.9 ± 0.08 [a]	0.9 ± 0.10 [a,d,e]	0.9 ± 0.10 [d]	0.9 ± 0.11 [e]	***

** $p<0.01$(conducted with Kruskal-Wallis H test)

*** $p<0.001$(conducted with Kruskal-Wallis H test)

Dunn's test: [a] $p<0.05$ for Low fall risk group versus Intermediate A fall risk group, [b] $p<0.05$ for Low fall risk group versus Intermediate B fall risk group, [c] $p<0.05$ for Low fall risk group versus High fall risk group, [d] $p<0.05$ for Intermediate A fall risk group versus Intermediate B fall risk group, [e] $p<0.05$ for Intermediate A fall risk group versus High fall risk group, [f] $p<0.05$ for Intermediate B fall risk group versus High fall risk group

4.6.1 Timed Up and Go Test (TUG)

Based on Table 4.6, the time to complete TUG test was increased from lower to higher fall risk group. Subject that required 15 second to complete this test would acquire higher fall risk. This was supported by Shumway-Cook et al. (2000) which reported that older people that used more than 13.5 second to complete test is at risk for falling. Apart from that, Alexandre et al. (2012) reported 12.47 second as the threshold value. Therefore, the fall risk of other three groups were lower because its TUG completion time less than threshold value. In short, subjects have greater fall risk if TUG completion time are higher, where these subjects maybe suspected for gait disorder.

4.6.2 Functional Reach (FR)

From lower to higher fall risk group, the FR scores were decreased. The lower score is representing imbalance problem so increase the risk for fall (Lin et al., 2004). According to Thomas and Lane (2005), older patient who reached less than 18.5 cm indicated higher fall risk. This is almost similar with the result obtained which 19 cm for High fall risk group. Besides that, Williams et al. (2017) stated the score within 15.24 cm to 25.40 cm acquired fall risk with two times higher. From the result, Low to Intermediate fall risk group reached average distance greater than 25.40 cm so it depicts a lower fall risk.

4.6.3 Hand Grip Strength (HGS)

As for HGS, the score was lowest in High fall risk group. This is supported by Yang et al. (2018) which stated the HGS was lower in a group that had recently fallen compared to group that had not fallen. Besides that, the HGS score is different for both genders. Based on result obtained, Low and Intermediate B fall risk group had large proportion of male whereas Intermediate A and High fall risk group had large proportion of female. Therefore, the score was lower as expected at Intermediate A and High fall risk group. Giles et al. (2003) reported that the fall risk was increased for those with HGS less than 25th percentile for both genders. This is supporting the result where fall risk increases when HGS decreased for both genders. In short, lower hand grip strength suspected for muscle weakness and increase fall risk (Moreland et al., 2004).

4.6.4 Cardiovascular Variability Ratio (Systolic Blood Pressure, Diastolic Blood Pressure and RR Interval)

High fall risk group had lower SBP_SSR_ARV and DBP_SSR_ARV value. Apart from that, no significant difference was observed in RR_SSR_ARV. Previous studies had evaluated only the absolute BP difference in postural change to identify presence of OH (Heitterachi et al., 2002). In this study, short-term blood pressure variability (BPV) provides measurement on changes in BP with posture change to assess its potential relevance to falls. Furthermore, ARV index is proven as reliable metric for prognostic significance of BPV (Mena et al., 2005). As SSR measures the change in BPV between standing to supine ratio, reduction in SBP_SSR_ARV and DBP_SSR_ARV demonstrate possible reduction in reactivity in BP control for the upright posture (Goh et al., 2017). This is considered as cardiovascular disorder and may give direct effect on susceptibility to falls. However, RR_SSR_ARV shows no difference and thus it assumed as non-relevant fall risk factor in this case.

4.6.5 Cognitive Assessment (MoCA questionnaire)

In addition, subjects with higher fall risk had lower score in MoCA test. This test aims to screen people for dementia. A score of 26 and higher is considered as normal, 22.1 with mild cognitive impairment and 16.2 with Alzheimer’s disease (Andrew, 2020). From the result obtained, Intermediate B and High fall risk group had score below 26, where these subjects maybe suspected for cognitive impairment. Cognitive impairment is known as fall risk factor in many studies (Sieri and Beretta, 2004; Rubenstein and Josephson, 2006).

4.6.6 Demographic

From Table 4.6, subjects with age 73.1 years old had higher fall risk compared to 65.7 years old. It indicates that risk increases with age (Gale, Cooper and Aihie Sayer, 2016). The gait imbalance and weaker muscle strength are associated with advanced age (Verghese et al., 2006; Keller and Engelhardt, 2013) . Across all age groups, female is prone to higher fall risk (Stevens and Sogolow, 2005). This may due to the weaker muscle strength from biological factor.

4.6.7 Height, Body Mass Index and Waist to Hip Ratio

According to result, High fall risk group had lower height compared to Low fall risk group. None of the previous study had study the relationship between height and fall risk. Therefore, it is linked with BMI where lower height associated with higher BMI in adults (Sperrin et al., 2016). The higher BMI may result in obesity and thus cause instability (Hue et al., 2007). This is found similar from the result which higher BMI in High fall risk group. However, it is not considered as major risk factor because the difference among the fall risk group was small. On the other hand, WHR was found no difference across all the groups. Therefore, it is categorized as non-relevant fall risk factor in this dataset.

4.7 Summary

The MELoR data set contains one thousand four hundred eleven subjects and one hundred and thirty-nine variables. After data pre-processing and feature selection, it had reduced to one thousand two hundred seventy-nine subjects and nine variables. Most of the selected variables were identified as major fall risk factors in literature review. Among of different combination of feature extraction and clustering algorithm, t-SNE feature extraction methods and K-means clustering algorithm had the highest performance in clustering validation. It groups the subjects into Low (13%), Intermediate A (19%), Intermediate B (21%) and High (31%) fall risk group. After conducted characteristic interpretation for each group, older people with higher fall risk have slower gait, imbalance, weaker muscle strength, with cardiovascular disorder, poor performance in cognitive test, and advancing age.

CHAPTER 5

CONCLUSIONS AND RECOMMENDATIONS

5.1 Conclusions

The aim and objectives of this project were achieved. The major risk factors for falls in older cohort were identified through literature review, which are gait and balance, muscle strength, cardiovascular disorder, cognitive impairment, age and gender. One thousand two hundred seventy-nine subjects and nine variables were chosen from MELoR dataset after feature selection. Most of the selected variables were similar to major fall risk factor that identified in literature review.

After this, t-SNE for feature extraction combined with K-means clustering demonstrated the highest performance compared to other five combinations. It achieved Silhouette Coefficient of 0.41, Davies Bouldin score of 0.78 and maximum group size difference of 149. By using this clustering algorithm, the data points were successfully clustered into four groups. It consisted of Low (13%), Intermediate A (19%), Intermediate B (21%) and High (31%) fall risk group.

The odd ratio for older people at High fall risk group to fall was 1.48 (>1). This reveals that older people clustered at this group are exposed to almost 50% higher risk of falls among overall older cohort in the dataset. After analysis, the characteristics of older people with high fall risk were interpreted as slower gait, imbalance, weaker muscle strength, with cardiovascular disorder, poorer performance in cognitive test, and advancing age. Besides that, female was prone to higher fall risk compared to male. In overall, this clustering algorithm present a potential as assessment tool in management of falls.

5.2 Recommendations for Future Work

First and foremost, a further classification algorithm can be developed as future work. In this project, the clustering algorithm can successfully group the subjects into four groups based on the characteristic of fall risk factors. The subjects in each group can be further classified into faller or non-faller by new variables that does not include in feature selection. To illustrate this, it might have more variables if the falls data is from other datasets, then these new or extra variables can be used for classification after clustering. The information obtained from clustering algorithm may improve the accuracy of classification. Therefore, the difference between faller and non-faller can be further distinguished and analysed.

Apart from that, graphical user interface (GUI) can be developed. Instead of interacting with Python console, it will be more convenient for clinician to work with GUI. It can convey the necessary information where the action is taken by user. Once the action is taken, the clustering result can direct visualise by clinician. In this case, Python provides number of GUI frameworks including Tkinter, Kivy and PyQT.

The proposed fall risk clustering algorithm only tested for MELoR dataset in this project. It may achieve different outcomes when tested with other fall datasets. Future work shall be working on evaluation of algorithm in different datasets. The chosen dataset may have higher number of subjects and variables. In addition, other risk factors such as visual impairment, uses of assistive device, medication and environment hazard can be analysed. It is also suggested that search and analyse dataset that contain balanced number between faller and non-faller.

REFERENCES

Akbar, F. and Setiati, S., 2018. Correlation between hand grip strength and nutritional status in elderly patients. *Journal of Physics: Conference Series*, 1073(4).

Al-Shammari, A., Zhou, R., Naseriparsaa, M. and Liu, C., 2019. An effective density-based clustering and dynamic maintenance framework for evolving medical data streams. *International Journal of Medical Informatics*, [online] 126(November 2018), pp.176–186. Available at: <https://doi.org/10.1016/j.ijmedinf.2019.03.016>.

Albayrak, S. and Amasyalı, F., 2003. Fuzzy C-means Clustering On Medical Diagnostic Systems. *International Twelfth Turkish Symposium on Artificial Intelligence and Neural Networks*, pp.3–5.

Allum, J.H.J., Carpenter, M.G., Honegger, F., Adkin, A.L. and Bloem, B.R., 2002. Age-dependent variations in the directional sensitivity of balance corrections and compensatory arm movements in man. *Journal of Physiology*, 542(2), pp.643–663.

Álvarez, J.D., Matias-Guiu, J.A., Cabrera-Martín, M.N., Risco-Martín, J.L. and Ayala, J.L., 2019. An application of machine learning with feature selection to improve diagnosis and classification of neurodegenerative disorders. *BMC Bioinformatics*, 20(1), pp.1–12.

Andrew, R., 2020. Montreal Cognitive Assessment (MoCA) Test for Dementia. [online] Available at: < https://www.verywellhealth.com/alzheimers-and-montreal-cognitive-assessment-moca-98617> [Accessed 20 August 2020].

Aranda-Gallardo, M., Morales-Asencio, J.M., Canca-Sanchez, J.C., Barrero-Sojo, S., Perez-Jimenez, C., Morales-Fernandez, A., De Luna-Rodriguez, M.E., Moya-Suarez, A.B. and Mora-Banderas, A.M., 2013. Instruments for assessing the risk of falls in acute hospitalized patients: A systematic review and meta-analysis. *BMC Health Services Research*, 13(1).

Arnold, C.M. and Faulkner, R.A., 2007. The history of falls and the association of the timed up and go test to falls and near-falls in older adults with hip osteoarthritis. *BMC Geriatrics*, 7, pp.1–9.

Aydin, A.E., Soysal, P. and Isik, A.T., 2017. Which is preferable for orthostatic hypotension diagnosis in older adults: Active standing test or head-up tilt table test? *Clinical Interventions in Aging*, 12, pp.207–212.

Azzag, H. and Lebbah, M., 2008. Clustering of self-organizing map. *ESANN 2008 Proceedings, 16th European Symposium on Artificial Neural Networks - Advances in Computational Intelligence and Learning*, 11(3), pp.209–214.

Balamural, M. and Melkumyan, A., 2016. t-SNE Based Visualisation and Clustering of Geological Domain. *Neural Information Processing*, pp 565-572.

Bateni, H., Zecevic, A., McIlroy, W.E. and Maki, B.E., 2004. Resolving conflicts in task demands during balance recovery: Does holding an object inhibit compensatory grasping? *Experimental Brain Research*, 157(1), pp.49–58.

Bedre, R., 2019. Chi-square test in Python. [online] Available at: < https://reneshbedre.github.io/blog/chisq.html/> [Accessed 10 August 2020].

Bennett, D.A., 2001. How can I deal with missing data in my study? *Australian and New Zealand Journal of Public Health*, 25(5), pp.464–469.

Borges, V.S., Silva, N.S., Malta, A.C., Xavier, N.C. and Bernardes, L.E.S., 2017. Falls, muscle strength, and functional abilities in community-dwelling elderly women. *Fisioterapia em Movimento*, 30(2), pp.357–366.

Cary, I. and Dipcot, J.A., 2003. A Comparison of Dominant and Non-dominant Hand Function in both Right- and Left-Handed Individuals using the Southampton Hand Assessment Procedure (SHAP). *The British Journal of Hand Therapy*, 8(1), pp.4–10.

Cebeci, Z. and Yildiz, F., 2015. Comparison of K-Means and Fuzzy C-Means Algorithms on Different Cluster Structures. *Journal of Agricultural Informatics*, 6(3), pp.13–23.

Chaiwanichsiri, D., Janchai, S. and Tantisiriwat, N., 2009. Foot disorders and falls in older persons. *Gerontology*, 55(3), pp.296–302.

Cheng, T.H., Wei, C.P. and Tseng, V.S., 2006. Feature selection for medical data mining: Comparisons of expert judgment and automatic approaches. *Proceedings - IEEE Symposium on Computer-Based Medical Systems*, 2006, pp.165–170.

Cho, B.Y., Seo, D.C., Lin, H.C., Lohrmann, D.K. and Chomistek, A.K., 2018. BMI and Central Obesity With Falls Among Community-Dwelling Older Adults. *American Journal of Preventive Medicine*, [online] 54(4), pp.e59–e66. Available at: <http://dx.doi.org/10.1016/j.amepre.2017.12.020>.

Chormunge, S. and Jena, S., 2018. Correlation based feature selection with clustering for high dimensional data. *Journal of Electrical Systems and Information Technology*, [online] 5(3), pp.542–549. Available at: <http://dx.doi.org/10.1016/j.jesit.2017.06.004>.

Chumerin, N. and Van Hulle, M.M., 2006. Comparison of two feature extraction methods based on maximization of mutual information. *Proceedings of the 2006 16th IEEE Signal Processing Society Workshop on Machine Learning for Signal Processing, MLSP 2006*, pp.343–348.

Cordeiro, R.C., Jardim, J.R., Perracini, M.R. and Ramos, L.R., 2009. Factors associated with functional balance and mobility among elderly diabetic outpatients. *Arquivos Brasileiros de Endocrinologia & Metabologia*, 53(7), pp.834–843.

D. Belka, R. and DeBeliso, M., 2019. Hand Grip Strength and Older Adults: Is Hand Grip Strength Associated with Self-Efficacy in Older Adults? *Journal of Physical Activity Research*, 4(1), pp.41–46.

Daszykowski, M., Walczak, B. and Massart, D.L., 2004. Density-based clustering for exploration of analytical data. *Analytical and Bioanalytical Chemistry*, 380(3 SPEC.ISS.), pp.370–372.

Delbaere, K., Kochan, N.A., Close, J.C.T., Menant, J.C., Sturnieks, D.L., Brodaty, H., Sachdev, P.S. and Lord, S.R., 2012. Mild cognitive impairment as a predictor of falls in community-dwelling older people. *American Journal of Geriatric Psychiatry*, [online] 20(10), pp.845–853. Available at: <http://dx.doi.org/10.1097/JGP.0b013e31824afbc4>.

Denkinger, M.D., Lukas, A., Nikolaus, T. and Hauer, K., 2015. Factors associated with fear of falling and associated activity restriction in community-dwelling older adults: A systematic review. *American Journal of Geriatric Psychiatry*, [online] 23(1), pp.72–86. Available at: <http://dx.doi.org/10.1016/j.jagp.2014.03.002>.

Dhargave, P. and Sendhilkumar, R., 2016. Prevalence of risk factors for falls among elderly people living in long-term care homes. *Journal of Clinical Gerontology and Geriatrics*, [online] 7(3), pp.99–103. Available at: <http://dx.doi.org/10.1016/j.jcgg.2016.03.004>.

Dhital, A., Pey, T. and Stanford, M.R., 2010. Visual loss and falls: A review. *Eye*, [online] 24(9), pp.1437–1446. Available at: <http://dx.doi.org/10.1038/eye.2010.60>.

Ding, C. and Li, T., 2007. Adaptive dimension reduction using discriminant analysis and K-means clustering. *ACM International Conference Proceeding Series*, 227, pp.521–528.

Ding, L. and Yang, F., 2016. Muscle weakness is related to slip-initiated falls among community-dwelling older adults. *Journal of Biomechanics*, [online] 49(2), pp.238–243. Available at: <http://dx.doi.org/10.1016/j.jbiomech.2015.12.009>.

Dong, Y., Sharma, V.K., Chan, B.P.L., Venketasubramanian, N., Teoh, H.L., Seet, R.C.S., Tanicala, S., Chan, Y.H. and Chen, C., 2010. The Montreal Cognitive Assessment (MoCA) is superior to the Mini-Mental State Examination (MMSE) for the detection of vascular cognitive impairment after acute stroke. *Journal of the Neurological Sciences*, [online] 299(1–2), pp.15–18. Available at: <http://dx.doi.org/10.1016/j.jns.2010.08.051>.

Doorn, C. Van, Gruber-baldini, A.L., Zimmerman, S., Hebel, J.R., Port, C.L., Baumgarten, M., Quinn, C.C., Taler, G., May, C. and Magaziner, J., 2003. Dementia as a Risk Factor for Falls and Fall Injuries Among Nursing Home Residents Dementia as a Risk Factor for Falls and Fall Injuries Among Nursing Home Residents. pp.1213–1218.

Drakos, G., 2020. Silhouette Analysis vs Elbow Method vs Davies-Bouldin Index: Selecting the optimal number of clusters for KMeans clustering. [online] Available at: < https://gdcoder.com/silhouette-analysis-vs-elbow-method-vs-davies-bouldin-index-selecting-the-optimal-number-of-clusters-for-kmeans-clustering/> [Accessed 10 August 2020].

Duxbury, A.S., 2000. Gait disorders and fall risk: Detection and prevention. *Comprehensive Therapy*, 26(4), pp.238–245.

Dy, J.G. and Brodley, C.E., 2004. Feature selection for unsupervised learning. *Journal of Machine Learning Research*, 5, pp.845–889.

Elliott, S., Painter, J. and Hudson, S., 2009. Living alone and fall risk factors in community-dwelling middle age and older adults. *Journal of Community Health*, 34(4), pp.301–310.

Escudero, J., Zajicek, J.P. and Ifeachor, E., 2011. Early detection and characterization of Alzheimer's disease in clinical scenarios using Bioprofile concepts and K-means. *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Conference*, 2011, pp.6470–6473.

Feldman, H.H. and Jacova, C., 2005. Mild cognitive impairment. *American Journal of Geriatric Psychiatry*, [online] 13(8), pp.645–655. Available at: <http://dx.doi.org/10.1097/00019442-200508000-00003>.

Gale, C.R., Cooper, C. and Aihie Sayer, A., 2016. Prevalence and risk factors for falls in older men and women: The English Longitudinal Study of Ageing. *Age and ageing*, 45(6), pp.789–794.

Gangavati, A., Hajjar, I., Quach, L., Jones, R.N., Kiely, D.K., Gagnon, P. and Lipsitz, L.A., 2011. Hypertension, orthostatic hypotension, and the risk of falls in a community-dwelling elderly population: The maintenance of balance, independent living, intellect, and zest in the elderly of Boston study. *Journal of the American Geriatrics Society*, 59(3), pp.383–389.

Ganz, DA, Bao Y, Shekelle PG, R.L., 2007. Clinician's corner : Will my patient fall? Patient scenario. *Journal of American Medical Association*, 297(1), pp.77–86.

Giles, L., Harrison, J., Miller, M., Andrews, G. and Crotty, M., 2003. A clinically relevant criterion for grip strength: relationship with falling in a sample of older adults. *Nutrition & Dietetics*, 60(4), pp.248-252

Gluhm, S., Goldstein, J., Loc, K., Colt, A., Liew, C. Van and Corey-Bloom, J., 2013. Cognitive performance on the mini-mental state examination and the montreal cognitive assessment across the healthy adult lifespan. *Cognitive and Behavioral Neurology*, 26(1), pp.1–5.

Goh, C.H., Ng, S.C., Kamaruzzaman, S.B., Chin, A.V., Poi, P.J.H., Chee, K.H., Imran, Z.A. and Tan, M.P., 2016. Evaluation of two new indices of blood pressure variability using postural change in older fallers. *Medicine (United States)*, 95(19), pp.1–6.

Goh, C.H., Ng, S.C., Kamaruzzaman, S.B., Chin, A.V. and Tan, M.P., 2017. Standing beat-to-beat blood pressure variability is reduced among fallers in the Malaysian Elders Longitudinal Study. *Medicine (United States)*, 96(42), pp.1–7.

González, J. 2018. Scaling/ normalisation/ standardisation: a pervasive question. QuantDare, [blog] 18 October. Available at: < https://quantdare.com/scaling-normalisation-standardisation-a-pervasive-question/ > [Accessed 10 August 2020].

Goyal, K., 2020. Data Preprocessing in Machine Learning: 7 Easy Steps To Follow. [online] Available at: < https://www.upgrad.com/blog/data-preprocessing-in-machine-learning/.> [Accessed 10 August 2020].

Guo, Q., Lu, X., Gao, Y., Zhang, J., Yan, B., Su, D., Song, A., Zhao, X. and Wang, G., 2017. Cluster analysis: A new approach for identification of underlying risk factors for coronary artery disease in essential hypertensive patients. *Scientific Reports*, [online] 7(November 2016), pp.1–7. Available at: <http://dx.doi.org/10.1038/srep43965>.

Härlein, J., Dassen, T., Halfens, R.J.G. and Heinze, C., 2009. Fall risk factors in older people with dementia or cognitive impairment: A systematic review. *Journal of Advanced Nursing*, 65(5), pp.922–933.

Heitterachi, E., Lord, S.R., Meyerkort, P., McCloskey, I. and Fitzpatrick, R., 2002. Blood pressure changes on upright tilting predict falls in older people. *Age and Ageing*, 31(3), pp.181–186.

Van Helden, S., Wyers, C.E., Dagnelie, P.C., Van Dongen, M.C., Willems, G., Brink, P.R.G. and Geusens, P.P., 2007. Risk of falling in patients with a recent fracture. *BMC Musculoskeletal Disorders*, 8, pp.1–7.

Hendrich, A.L., Bender, P.S. and Nyhuis, A., 2003. Validation of the Hendrich II Fall Risk Model: A large concurrent case/control study of hospitalized patients. *Applied Nursing Research*, 16(1), pp.9–21.

Herman, T., Mirelman, A., Giladi, N., Schweiger, A. and Hausdorff, J.M., 2010. Executive control deficits as a prodrome to falls in healthy older adults: A prospective study linking thinking, walking, and falling. *Journals of Gerontology - Series A Biological Sciences and Medical Sciences*, 65 A(10), pp.1086–1092.

Himes, C.L. and Reynolds, S.L., 2012. Effect of obesity on falls, injury, and disability. *Journal of the American Geriatrics Society*, 60(1), pp.124–129.

Hira, Z.M. and Gillies, D.F., 2015. A review of feature selection and feature extraction methods applied on microarray data. *Advances in Bioinformatics*, 2015(1).

Hohler, A.D., Zuzuárregui, J.R.P., Katz, D.I., DePiero, T.J., Hehl, C.L., Leonard, A., Allen, V., Dentino, J., Gardner, M., Phenix, H., Saint-Hilaire, M. and Ellis, T., 2012. Differences in motor and cognitive function in patients with Parkinson's disease with and without orthostatic hypotension. *International Journal of Neuroscience*, 122(5), pp.233–236.

Horlings, C.G.C., van Engelen, B.G.M., Allum, J.H.J. and Bloem, B.R., 2008. A weak balance: The contribution of muscle weakness to postural instability and falls. *Nature Clinical Practice Neurology*, 4(9), pp.504–515.

Horton, K., 2007. Gender and the risk of falling: A sociological approach. *Journal of Advanced Nursing*, 57(1), pp.69–76.

Hsu, C.C., 2006. Generalizing self-organizing map for categorical data. *IEEE Transactions on Neural Networks*, 17(2), pp.294–304.

Hue, O., Simoneau, M., Marcotte, J., Berrigan, F., Doré, J., Marceau, P., Marceau, S., Tremblay, A. and Teasdale, N., 2007. Body weight is a strong predictor of postural stability. *Gait and Posture*, 26(1), pp.32–38.

Jade, A.M., Srikanth, B., Jayaraman, V.K., Kulkarni, B.D., Jog, J.P. and Priya, L., 2003. Feature extraction and denoising using kernel PCA. *Chemical Engineering Science*, 58(19), pp.4441–4448.

Jason, B., 2017. How to Handle Missing Data with Python. [online] Available at: < https://machinelearningmastery.com/handle-missing-data-python/> [Accessed 10 August 2020].

Jason, B., 2018. A Gentle Introduction to Normality Tests in Python. [online] Available at: < https://machinelearningmastery.com/a-gentle-introduction-to-normality-tests-in-python/#:~:text=It%20can%20be%20used%20to,than%20a%20single%20p%2Dvalue.> [Accessed 20 August 2020].

Jason, B., 2019. How to Choose a Feature Selection Method For Machine Learning. [online] Available at: < https://machinelearningmastery.com/feature-selection-with-real-and-categorical-data/> [Accessed 12 August 2020].

Johnsson, E., Henriksson, M. and Hirschfeld, H., 2003. Does the functional reach test reflect stability limits in elderly people? *Journal of Rehabilitation Medicine*, 35(1), pp.26–30.

Jung, D., 2008. Fear of Falling in Older Adults: Comprehensive Review. *Asian Nursing Research*, [online] 2(4), pp.214–222. Available at: <http://dx.doi.org/10.1016/S1976-1317(09)60003-7>.

Kalyani, P., 2012. Approaches to Partition Medical Data using Clustering Algorithms. *International Journal of Computer Applications*, 49(23), pp.7–10.

Keller, K. and Engelhardt, M., 2013. Strength and muscle mass loss with aging process. Age and strength loss. *Muscles, Ligaments and Tendons Journal*, 3(4), pp.346–350.

Kemal, E.,2018. Fuzzy clustering. [online] Available at: < http://eneskemalergin.github.io/blog//blog/Fuzzy_Clustering/> [Accessed 10 August 2020].

Khalid, S., Khalil, T. and Nasreen, S., 2014. A survey of feature selection and feature extraction techniques in machine learning. *Proceedings of 2014 Science and Information Conference, SAI 2014*, pp.372–378.

Khalid, S. and Prieto-Alhambra, D., 2019. Machine Learning for Feature Selection and Cluster Analysis in Drug Utilisation Research. *Current Epidemiology Reports*, 6(3), pp.364–372.

Krakovska, O., Christie, G., Sixsmith, A., Ester, M. and Moreno, S., 2019. Performance comparison of linear and nonlinear feature selection methods for the analysis of large survey datasets. *PLoS ONE*, 14(3), pp.1–17.

Kumar Roy, D. and Mohan Pandey, H., 2018. A New Clustering Method Using an Augmentation to the Self Organizing Maps. *Proceedings of the 8th International Conference Confluence 2018 on Cloud Computing, Data Science and Engineering, Confluence 2018*, pp.739–742.

De La Torre, F. and Kanade, T., 2006. Discriminative cluster analysis. *ACM International Conference Proceeding Series*, 148, pp.241–248.

Lakany, H., 2008. Extracting a diagnostic gait signature. *Pattern Recognition*, 41(5), pp.1627–1637.

Large, J., Gan, N., Basic, D. and Jennings, N., 2006. Using the timed up and go test to stratify elderly inpatients at risk of falls. *Clinical Rehabilitation*, 20(5), pp.421–428.

Larner, A.J., 2012. Screening utility of the montreal cognitive assessment (MoCA): In place of - Or as well as - The MMSE? *International Psychogeriatrics*, 24(3), pp.391–396.

Letts, L., Moreland, J., Richardson, J., Coman, L., Edwards, M., Ginis, K.M., Wilkins, S. and Wishart, L., 2010. The physical environment as a fall risk factor in older adults: Systematic review and meta-analysis of cross-sectional and cohort studies. *Australian Occupational Therapy Journal*, 57(1), pp.51–64.

Li, D.C., Liu, C.W. and Hu, S.C., 2011. A fuzzy-based data transformation for feature extraction to increase classification performance with small medical data sets. *Artificial Intelligence in Medicine*, [online] 52(1), pp.45–52. Available at: <http://dx.doi.org/10.1016/j.artmed.2011.02.001>.

Lin, M., Hwang, H., Hu, M., Wu, H.I., Wang, Y. and Huang, F., 2004. Psychometric comparisons of the timed up and go. *Journal of the American Geriatrics Society*, 52(8), pp.1343–1348.

Lindsay, R., James, E.L. and Kippen, S., 2004. The timed up and go test: Unable to predict falls on the acute medical ward. *Australian Journal of Physiotherapy*, [online] 50(4), pp.249–251. Available at: <http://dx.doi.org/10.1016/S0004-9514(14)60115-X>.

Lord, S.R., Murray, S.M., Chapman, K., Munro, B. and Tiedemann, A., 2002. Sit-to-stand performance depends on sensation, speed, balance, and psychological status in addition to strength in older people. *Journals of Gerontology - Series A Biological Sciences and Medical Sciences*, 57(8), pp.539–543.

Low, P.A., 2008. Prevalence of orthostatic hypotension. *Clinical Autonomic Research*, 18(SUPPL. 1), pp.8–13.

Malarvizhi, A. and Ravichandran, S., 2018. Data mining's role in mining medical datasets for disease assessments – A case study. *International Journal of Pure and Applied Mathematics*, [online] 119(12), pp.16255–16259. Available at: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85048827315&partnerID=40&md5=d58a7a1278f41f25b4673d6bfa2e94ff>.

Mali, M., Kulkarni, P. and Bagade, V., 2017. Medical Records Clustering: A Survey. *International Journal of Innovative Research in Computer and Communication Engineering*, [online] 5(4), pp.8198–8205. Available at: <www.ijircce.com>.

Malik, U. 2018. Implementing PCA in Python with Scikit-Learn. Stack Abuse, [blog] 18 October. Available at: < https://stackabuse.com/implementing-pca-in-python-with-scikit-learn/> [Accessed 10 August 2020].

Maranhao Neto, G.A., Oliveira, A.J., Pedreiro, R.C. de M., Pereira-Junior, P.P., Machado, S., Marques Neto, S. and Farinatti, P.T.V., 2017. Normalizing handgrip strength in older adults: An allometric approach. *Archives of Gerontology and Geriatrics*, [online] 70, pp.230–234. Available at: <http://dx.doi.org/10.1016/j.archger.2017.02.007>.

Markey, M.K., Lo, J.Y., Tourassi, G.D. and Floyd, C.E., 2003. Self-organizing map for cluster analysis of a breast cancer database. *Artificial Intelligence in Medicine*, 27(2), pp.113–127.

Morcelli, M.H., Crozara, L.F., Rossi, D.M., Laroche, D.P., Ribeiro Marques, N., Hallal, C.Z., Castro, A., Cardozo, A.C., Gonçalves, M. and Navega, M.T., 2014. Hip muscles strength and activation in older fallers and non-fallers. *Isokinetics and Exercise Science*, 22(3), pp.191–196.

Molala, R.,2019. MLmuse: Correlation and Collinearity — How they can make or break a model. [online] Available at: < https://blog.clairvoyantsoft.com/correlation-and-collinearity-how-they-can-make-or-break-a-model-9135fbe6936a> [Accessed 10 August 2020].

Moreland, J.D., Richardson, J.A., Goldsmith, C.H. and Clase, C.M., 2004. Muscle weakness and falls in older adults: A systematic review and meta-analysis. *Journal of the American Geriatrics Society*, 52(7), pp.1121–1129.

Morse, J.M., Morse, R.M. and Tylko, S.J., 1989. Development of a Scale to Identify the Fall-Prone Patient. *Canadian Journal on Aging / La Revue canadienne du vieillissement*, 8(4), pp.366–377.

Motoda, H. and Liu, H., 2002. Feature selection, extraction and construction. *Communication of IICM*, 5, pp.67–72.

Muir, S.W., Gopaul, K. and Montero Odasso, M.M., 2012. The role of cognitive impairment in fall risk among older adults: A systematic review and meta-analysis. *Age and Ageing*, 41(3), pp.299–308.

Murman, D.L., 2015. The Impact of Age on Cognition Cognition and the Aging Auditory System. *Seminars in Hearing*, [online] 36(1), pp.111–121. Available at: <http://dx.doi.org/10.1055/s-0035-1555115.>.

Murtagh, F. and Contreras, P., 2012. Algorithms for hierarchical clustering: An overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2(1), pp.86–97.

Nasreddine, Z.S., Phillips, N.A., Bedirian, V., Charbonneau, S., Whitehead, V., Collin, I., Cummings, J.L. and Chertkow, H., 2005. The Montreal Cognitive Assessment, MoCA: A Brief Screening Tool For Mild Cognitive Impairment. *Journal of the American Geriatrics Society*, pp.695–699.

Neumann, D.A., 2010. Kinesiology of the hip: A focus on muscular actions. *Journal of Orthopaedic and Sports Physical Therapy*, 40(2), pp.82–94.

Nithya, N.S., Duraiswamy, K. and Gomathy, P., 2013. A Survey on Clustering Techniques in Medical Diagnosis. *International Journal of Computer Science Trends and Technology*, [online] 1(2), pp.17–22. Available at: <www.ijcstjournal.org>.

Ogbuabor, G. and F. N, U., 2018. Clustering Algorithm for a Healthcare Dataset Using Silhouette Score Value. *International Journal of Computer Science and Information Technology*, 10(2), pp.27–37.

Oliver, D., Britton, M., Seed, P., Martin, F.C. and Hopper, A.H., 1997. Development and evaluation of evidence based risk assessment tool (STRATIFY) to predict which elderly inpatients will fall: Case-control and cohort studies. *British Medical Journal*, 315(7115), pp.1049–1053.

Oyelade, J., Isewon, I., Oladipupo, O., Emebo, O., Omogbadegun, Z., Aromolaran, O., Uwoghiren, E., Olaniyan, D. and Olawole, O., 2019. Data Clustering: Algorithms and Its Applications. *Proceedings - 2019 19th International Conference on Computational Science and Its Applications, ICCSA 2019*, (ii), pp.71–81.

Panda, S., Sahu, S., Jena, P. and Chattopadhyay, S., 2012. Comparing fuzzy-C means and K-means clustering techniques: A comprehensive study. *Advances in Intelligent and Soft Computing*, 166 AISC(VOL. 1), pp.451–460.

Park, H.M., 2016. Univariate Analysis and Normality Test Using SAS, STATA, and SPSS. *Measurement*, (January), pp.1–67.

Park, N.H. and Lee, W.S., 2004. Statistical grid-based clustering over data streams. *SIGMOD Record*, 33(1), pp.32–37.

Pawan, J.,2019. Hierarchical clustering Clearly Explained. [online] Available at: < https://towardsdatascience.com/https-towardsdatascience-com-hierarchical-clustering-6f3c98c9d0ca> [Accessed 10 August 2020].

Pathak, M.,2018. Introduction to t-SNE. [online] Available at: < https://www.datacamp.com/community/tutorials/introduction-t-sne?utm_source=adwords_ppc&utm_campaignid=1455363063&utm_adgroupid=65083631748&utm_device=c&utm_keyword=&utm_matchtype=b&utm_network=g&utm_adpostion=&utm_creative=332602034358&utm_targetid=dsa429603003980&utm_loc_interest_ms=&utm_loc_physical_ms=1010274&gclid=EAIaIQobChMI9Z2L5fPW6wIVFQ4rCh3F1wL8EAAYASAAEgIwVvD_BwE> [Accessed 10 August 2020].

Peretz, C., Herman, T., Hausdorff, J.M. and Giladi, N., 2006. Assesing fear of failing: Can a short version of the activities-specific balance confidence scale be useful? *Movement Disorders*, 21(12), pp.2101–2105.

Petersen, P., Petrick, M., Connor, H. and Conklin, D., 1989. Grip Strength and Hand Dominance: Challenging the 10% Rule. *American Journal of Occupational Therapy*, 43(7), pp.444–447.

Pfortmueller, C.A., Kunz, M., Lindner, G., Zisakis, A., Puig, S. and Exadaktylos, A.K., 2014. Reducing fall risk in the elderly: risk factors and fall prevention, a systematic review. *The Scientific World Journal*, 2014(October).

Pijnappels, M., van der Burg, J.C.E., Reeves, N.D. and van Dieën, J.H., 2008. Identification of elderly fallers by muscle strength measures. *European Journal of Applied Physiology*, 102(5), pp.585–592.

Polat, K. and Güneş, S., 2007. An expert system approach based on principal component analysis and adaptive neuro-fuzzy inference system to diagnosis of diabetes disease. *Digital Signal Processing: A Review Journal*, 17(4), pp.702–710.

Praveen, P. and Rama, B., 2018. A Novel Approach to Improve the Performance of Divisive Clustering-BST. *Data Engineering and Intelligent Computing*, [online] 542, pp.283–291. Available at: <http://link.springer.com/10.1007/978-981-10-3223-3>.

Ramya, T.B., 2018. Disease Prediction System Using Fuzzy C-Means Algorithm. *International Journal of Engineering Research & Technology (IJERT)*, 6(3), pp.1–5.

Razali, N.M. and Wah, Y.B., 2011. Power comparisons of Shapiro-Wilk , Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. Journal of Statistical Modeling and Analytics, [online] 2(1), pp.21–33. Available at: <http://instatmy.org.my/downloads/e-jurnal 2/3.pdf%0Afiles/1576/Razali and Wah - 2011 - Power comparisons of Shapiro-Wilk, Kolmogorov-Smir.pdf>.

Rodriguez, M.Z., Comin, C.H., Casanova, D., Bruno, O.M., Amancio, D.R., Costa, L. da F. and Rodrigues, F.A., 2019. Clustering algorithms: A comparative approach. *PLoS ONE*.

Rogers, M.W. and Mille, M.L., 2003. Lateral stability and falls in older people. *Exercise and Sport Sciences Reviews*, 31(4), pp.182–187.

Rubenstein, L.Z. and Josephson, K.R., 2006. Falls and Their Prevention in Elderly People: What Does the Evidence Show? *Medical Clinics of North America*, 90(5), pp.807–824.

Rustempasic, I. and Can, M., 2013. Diagnosis of Parkinson's Disease using Fuzzy C-Means Clustering and Pattern Recognition. *Southeast Europe Journal of Soft Computing*, 2(1), pp.42–49.

Sahu, S., Jena, P. and Chattopadhyay, S., 2012. Comparing fuzzy-C means and K-means clustering techniques: A comprehensive study. *Advances in Intelligent and Soft Computing*, 166 AISC(VOL. 1), pp.451–460.

Salzman, B., 2011. Gait and balance disorders in older adults. *American Family Physician*, 82(1), pp.61–68.

Samant, R. and Rao, S., 2013. A study on Feature Selection Methods in Medical Decision Support Systems. *International Journal of Engineering Research & Technology (IJERT)*, 2(11), pp.615–620.

Santos, V., Datia, N. and Pato, M.P.M., 2014. Ensemble Feature Ranking Applied to Medical Data. *Procedia Technology*, 17, pp.223–230.

Saxena, A., Prasad, M., Gupta, A., Bharill, N., Patel, O.P., Tiwari, A., Er, M.J., Ding, W. and Lin, C.T., 2017. A review of clustering techniques and developments. *Neurocomputing*, [online] 267, pp.664–681. Available at: <http://dx.doi.org/10.1016/j.neucom.2017.06.053>.

Schatz, I.J., Bannister, R., Freeman, R.L., Jankovic, J., Koller, W.C., Low, P.A., Mathias, C.J., Polinsky, R.J., Quinn, N.P., Robertson, D. and Streeten, D.H.P., 1996. Consensus statement on the definition of orthostatic hypotension, pure autonomic failure and multiple system atrophy. *Clinical Autonomic Research*, 6(2), pp.125–126.

Schoene, D., Wu, S.M.S., Mikolaizak, A.S., Menant, J.C., Smith, S.T., Delbaere, K. and Lord, S.R., 2013. Discriminative ability and predictive validity of the timed up and go test in identifying older people who fall: Systematic review and meta-analysis. *Journal of the American Geriatrics Society*, 61(2), pp.202–208.

Schrager, M.A., Kelly, V.E., Price, R., Ferrucci, L. and Shumway-Cook, A., 2008. The effects of age on medio-lateral stability during normal and narrow base walking. *Gait and Posture*, 28(3), pp.466–471.

Scikit Learn, 2014. Manifold learning: Optimizing t-SNE. [online] Available at: < https://scikit-learn.org/stable/modules/manifold.html#t-sne> [Accessed 20 August 2020].

Sevene, T.G., Berning, J., Harris, C., Climstein, M., Adams, K.J. and DeBeliso, M., 2017. Hand Grip Strength and Gender: Allometric Normalization in Older Adults and Implications for the NIOSH Lifting Equation. *Journal of Lifestyle Medicine*, 7(2), pp.63–68.

Shaw, B.H. and Claydon, V.E., 2014. The relationship between orthostatic hypotension and falling in older adults. *Clinical Autonomic Research*, 24(1), pp.3–13.

Shaikh, R., 2018. Feature Selection Techniques in Machine Learning with Python. [online] Available at: < https://towardsdatascience.com/feature-selection-techniques-in-machine-learning-with-python-f24e7da3f36e#:~:text=using%20SelectKBest%20class-,2.,feature%20towards%20your%20output%20variable.> [Accessed 10 August 2020].

Shetye, A., 2019. Feature Selection with sklearn and Pandas. Towards Data Science, [blog] 18 February. Available at: < https://towardsdatascience.com/feature-selection-with-pandas-e3690ad8504b > [Accessed 10 August 2020].

Shibao, C., Grijalva, C.G., Raj, S.R., Biaggioni, I. and Griffin, M.R., 2007. Orthostatic Hypotension-Related Hospitalizations in the United States. *American Journal of Medicine*, 120(11), pp.975–980.

Shihab, K., 2004. Improving clustering performance by using feature selection and extraction techniques. *Journal of Intelligent Systems*, 13(3), pp.249–273.

Shumway-Cook, S. Brauer and Woollacott, M. , 2000. Predicting the Probability for Falls in Community-Dwelling Older Adults Using the Timed Up & Go Test, Physical Therapy, vol. 80, no. 9, pp. 896-903. Available: 10.1093/ptj/80.9.896.

Sidal-Gidan, F., 2013. Cognitive screening tools. *Clinician Reviews*, 23(1), pp.12–18.

Sieri, T. and Beretta, G., 2004. Fall risk assessment in very old males and females living in nursing homes. *Disability and Rehabilitation*, 26(12), pp.718–723.

Sinaki, M., Brey, R.H., Hughes, C.A., Larson, D.R. and Kaufman, K.R., 2005. Balance disorder and increased risk of falls in osteoporosis and kyphosis: Significance of kyphotic posture and muscle strength. *Osteoporosis International*, 16(8), pp.1004–1010.

Sperrin, M., Marshall, A.D., Higgins, V., Renehan, A.G. and Buchan, I.E., 2016. Body mass index relates weight to height differently in women and older adults: Serial cross-sectional surveys in England (1992-2011). *Journal of Public Health (United Kingdom)*, 38(3), pp.607–613.

Stasny, B.M., Newton, R.A., LoCascio, L.V., Bedio, N., Lauke, C., Conroy, M., Thompson, A., Vakhnenko, L. and Polidoro, C., 2011. The ABC scale and fall risk: A systematic review. *Physical and Occupational Therapy in Geriatrics*, 29(3), pp.233–242.

Stevens, J.A., Ballesteros, M.F., Mack, K.A., Rudd, R.A., DeCaro, E. and Adler, G., 2012. Gender differences in seeking care for falls in the aged medicare population. *American Journal of Preventive Medicine*, [online] 43(1), pp.59–62. Available at: <http://dx.doi.org/10.1016/j.amepre.2012.03.008>.

Stevens, J.A. and Sogolow, E.D., 2005. Gender differences for non-fatal unintentional fall related injuries among older adults. *Injury Prevention*, 11(2), pp.115–119.

T. SenthilSelvi and R. Parimala, 2018. Improving Clustering Accuracy using Feature Extraction Method. *International Journal of Scientific Research in Computer Science and Engineering*, 6(2), pp.15–19.

Tan, M.P. and Kenny, R.A., 2006. Cardiovascular assessment of falls in older people. *Clinical interventions in Aging*, 1(1), pp.57–66.

Taşdemir, K. and Merényi, E., 2009. Exploiting data topology in visualization and clustering of self-organizing maps. *IEEE Transactions on Neural Networks*, 20(4), pp.549–562.

Thakur, N. 2020. k-Means Clustering: Comparison of Initialization strategies. Analytics Vidhya, [blog] 11 April. Available at: < https://medium.com/analytics-vidhya/comparison-of-initialization-strategies-for-k-means-d5ddd8b0350e> [Accessed 10 August 2020].

Thomas, J.I. and Lane, J. V., 2005. A pilot study to explore the predictive validity of 4 measures of falls risk in frail elderly patients. *Archives of Physical Medicine and Rehabilitation*, 86(8), pp.1636–1640.

Todd, C. and Skelton, D., 2004. What are the main risk factors for falls amongst older people and what are the most effective interventions to prevent these falls ? *World Health*, [online] (March), p.28. Available at: <http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:What+are+the+main+risk+factors+for+falls+amongst+older+people+and+what+are+the+most+effective+interventions+to+prevent+these+falls+?#0>.

Tran, T., Luo, W., Phung, D., Gupta, S., Rana, S., Kennedy, R.L., Larkins, A. and Venkatesh, S., 2014. A framework for feature extraction from hospital medical data with applications in risk prediction. *BMC Bioinformatics*, 15(1), pp.1–9.

Tran, T.N., Drab, K. and Daszykowski, M., 2013. Revised DBSCAN algorithm to cluster data with dense adjacent clusters. *Chemometrics and Intelligent Laboratory Systems*, [online] 120, pp.92–96. Available at: <http://dx.doi.org/10.1016/j.chemolab.2012.11.006>.

Tromp, A.., Pluijm, S.M.., Smit, J.., Deeg, D.J.., Bouter, L.. and Lips, P., 2001. Fall-risk screening test. *Journal of Clinical Epidemiology*, 54(8), pp.837–844.

Uvarov, V., 2018. *Feature Selection Using Statistical Testing*. [online] Medium. Available at: <https://medium.com/@vadim_uvarov/feature-selection-using-statistical-testing-2d8e7b5e27b8> [Accessed 19 Apr. 2020].

Valarmathy, N. and Krishnaveni, S., 2019. Performance evaluation and comparison of clustering algorithms used in educational data mining. *International Journal of Recent Technology and Engineering*, 7(6), pp.103–112.

Vassallo, M., Mallela, S.K., Williams, A., Kwan, J., Allen, S. and Sharma, J.C., 2009. Fall risk factors in elderly patients with cognitive impairment on rehabilitation wards. *Geriatrics and Gerontology International*, 9(1), pp.41–46.

Van Der Velde, N., Van Den Meiracker, A.H., Pols, H.A.P., Stricker, B.H.C. and Van Der Cammen, T.J.M., 2007a. Withdrawal of fall-risk-increasing drugs in older persons: Effect on tilt-table test outcomes. *Journal of the American Geriatrics Society*, 55(5), pp.734–739.

Van Der Velde, N., Van Den Meiracker, A.H., Stricker, B.H.C. and Van Der Cammen, T.J.M., 2007b. Measuring orthostatic hypotension with the Finometer device: Is a blood pressure drop of one heartbeat clinically relevant? *Blood Pressure Monitoring*, 12(3), pp.167–171.

Verghese, J., LeValley, A., Hall, C.B., Katz, M.J., Ambrose, A.F. and Lipton, R.B., 2006. Epidemiology of gait disorders in community-residing older adults. *Journal of the American Geriatrics Society*, 54(2), pp.255–261.

Verghese, J., Robbins, M., Holtzer, R., Zimmerman, M., Wang, C., Xue, X. and Lipton, R.B., 2008. Gait dysfunction in mild cognitive impairment syndromes. *Journal of the American Geriatrics Society*, 56(7), pp.1244–1251.

Van Vost Moncada, L. and Mire, L.G., 2017. Preventing Falls in Older Patients. *Amerian Family Physician*, 96(4), pp.240–247.

Vries, M., Seppala, L.J., Daams, J.G., van de Glind, E.M.M., Masud, T., van der Velde, N., Blain, H., Bousquet, J., Bucht, G., Caballero-Mora, M.A., van der Cammen, T., Eklund, P., Emmelot-Vonk, M., Gustafson, Y., Hartikainen, S., Kenny, R.A., Laflamme, L., Landi, F., Masud, T., O'Byrne-Maguire, I., Petrovic, M., Rodriguez, L., Seppälä, L., Svensson, O., Szczerbińska, K., Thaler, H. and van der Velde, N., 2018. Fall-Risk-Increasing Drugs: A Systematic Review and Meta-Analysis: I. Cardiovascular Drugs. *Journal of the American Medical Directors Association*, [online] 19(4), pp.371.e1-371.e9. Available at: <https://doi.org/10.1016/j.jamda.2017.12.013>.

Wattenberg et al., 2016. How to Use t-SNE Effectively. Distill, [blog] 13 October. Available at: < https://distill.pub/2016/misread-tsne/> [Accessed 10 August 2020].

Williams, B., Allen, B., Hu, Z., True, H., Cho, J., Harris, A., Fell, N. and Sartipi, M., 2017. Real-time fall risk assessment using functional reach test. *International Journal of Telemedicine and Applications*, 2017.

Wrisley, D.M., Walker, M.L., Echternach, J.L. and Strasnick, B., 2003. Reliability of the Dynamic Gait Index in people with vestibular disorders. *Archives of Physical Medicine and Rehabilitation*, 84(10), pp.1528–1533.

Yang, N.P., Hsu, N.W., Lin, C.H., Chen, H.C., Tsao, H.M., Lo, S.S. and Chou, P., 2018. Relationship between muscle strength and fall episodes among the elderly: The Yilan study, Taiwan. BMC Geriatrics, 18(1), pp.1–7.

Ye, J., Janardan, R. and Li, Q., 2004. Two-dimensional linear discriminant analysis. *17th International Conference on Neural Information Processing Systems*, 2004, pp.1569-1576.

APPENDICES

APPENDIX A: Clustering Algorithm (Python Codes)

```
import pandas as pd
import numpy as np
import scipy.stats as stats
import matplotlib.pyplot as plt
import seaborn as sn
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from math import sqrt
from sklearn.decomposition import PCA
from sklearn.manifold import TSNE
from sklearn.preprocessing import StandardScaler
from sklearn.cluster import KMeans
from sklearn.cluster import AgglomerativeClustering
import skfuzzy as fuzz
from sklearn.metrics import silhouette_score,davies_bouldin_score

def data_preprocessing(df):
    df.Weight_kg = df.Weight_kg.astype('float64')
    df.Hip_cm = df.Hip_cm.astype('float64')
    Hand_grip=[]
    for x,y in enumerate(df['GS_DominantHand']): # add dominant hand grip strength
        if y =='Right':
            Hand_grip.append(df['RightHandAverage'][x])
        elif y=='Left':
            Hand_grip.append(df['LeftHandAverage'][x])
        else:
            Hand_grip.append(np.nan)
    df.insert(9, 'Dominant_Hand_grip', Hand_grip, True)
    y=df['Fall_questionnaire'] # target variable
    df=df.drop(['Fall_questionnaire','BPcondition'],axis=1)
    df=df.replace(0,np.nan)
    return(df,y)

def group_variable(df_list):
    categorical_list,numerical_list, info_list=[],[],[]
    for i in df_list:
        if df[i].dtype=='object':
            info_list.append(i)
        elif df[i].dtype=='int64':
            categorical_list.append(i)
        else:
            numerical_list.append(i)
    return(categorical_list,numerical_list, info_list)

def shapiro_wilk_test(_list):
   normal=[]
   not_normal=[]
   for x in _list:
       n = df[x]
       n=n.dropna()
       stat, p = stats.shapiro(n)
       alpha = 0.05
       if p > alpha:
           normal.append(x)
       else:
           not_normal.append(x)
   return(normal,not_normal)

def independent_t_test(_list):
    for x in _list:
        n=df[x].dropna()
        data1=n[target==1]
        data2=n[target==2]
        se1, se2 = ((np.std(data1, ddof=1))/sqrt(len(data1)),
                    (np.std(data2, ddof=1))/sqrt(len(data2))) # calculate standard errors
        sed = sqrt(se1**2.0 + se2**2.0) # standard error on the difference between the samples
        t_stat = (np.mean(data1) - np.mean(data2)) / sed # calculate the t statistic
        degree = len(data1) + len(data2) - 2 # degrees of freedom
        p = (1.0 - stats.t.cdf(abs(t_stat), degree)) * 2.0 # calculate the p-value
        t_score.append(t_stat)
        p_value1.append(p)
    return(t_score,p_value1)

def mannwhitneyu_test(_list):
    for x in not_normal: #for numerical feature that not normal distributed
        n = df[x].dropna()
        data1=n[target==1]
        data2=n[target==2]
        stat, p = stats.mannwhitneyu(data1, data2, alternative='two-sided')
        statistic.append(stat)
        p_value2.append(p)
    return(statistic,p_value2)

def chi2_test(_list):
    for x in _list:
        crosstab = pd.crosstab(target, df[x])
        chi2, value,dof,doo=stats.chi2_contingency(crosstab)
        chi2_score.append(chi2)
        p_value3.append(value)
    return(chi2_score,p_value3)

def cramer_v(_list):
    for var1 in _list:
        col = []
        for var2 in indp_categorical_list :
            crosstab =np.array(pd.crosstab(df[var1],df[var2], rownames=None, colnames=None)) # Cross table building
            stat = stats.chi2_contingency(crosstab)[0] # Keeping of the test statistic of the Chi2 test
            obs = np.sum(crosstab) # Number of observations
            mini = min(crosstab.shape)-1 # Take the minimum value between the columns and the rows of the cross table
            cramers =stat/(obs*mini)
            col.append(round(cramers,2)) # Keeping of the rounded value of the Cramer's V
        rows.append(col)
    return(rows)

def group_BP_feature(_list):
    for x in _list:
        if 'RR' in x:
            RR_feature.append(x)
        elif 'SBP' in x:
            SBP_feature.append(x)
        else:
            DBP_feature.append(x)
    return(RR_feature,SBP_feature,DBP_feature)
```

```
def feature_imp(_list):
    X=pd.concat([df[_list],target], axis=1)
    X=X.dropna()
    y=X['Fall_questionnaire']
    X=X.drop('Fall_questionnaire', axis=1)
    X_train, X_test, y_train, y_test = train_test_split(X, y,test_size=0.30,train_size=0.70,
stratify=y,random_state=0) # test split set
    model = RandomForestClassifier(max_depth=30,random_state=0)
    model.fit(X_train, y_train)
    col_sorted_by_importance=model.feature_importances_.argsort()[::-1]
    feat_imp=pd.DataFrame({ 'cols':X.columns[col_sorted_by_importance],
'imps':model.feature_importances_[col_sorted_by_importance]})
    return(feat_imp)

#################
### Main Code ###
#################
df = pd.read_excel(r'C:\Users\User\Desktop\Fyp Python Code\Fall Data.xlsx') # add directory for dataset
df_copy = df.copy()
df, target = data_preprocessing(df)
feature_description=df.describe().T
feature_list=df.columns[df.isnull().mean() < 0.1].tolist() # remove those feature with high missing ratio
categorical_list, numerical_list, info_list = group_variable(feature_list) # categorization
common_feature=numerical_list[0:14]
BP_feature=numerical_list[14:]
for x in categorical_list:
    df[x]=df[x].astype('category')
normal,not_normal=shapiro_wilk_test(numerical_list) # normality test
```

```
###Feature Selection###
t_score, p_value1,statistic,p_value2,chi2_score,
p_value3,rows,RR_feature,SBP_feature,DBP_feature=[],[],[],[],[],[],[],[],[],[]
alpha=0.05
t_score,p_value1=independent_t_test(normal) # independent T-test
t_test = pd.DataFrame({'feature':normal,'t-value': t_score,'p': p_value1})
t_test=t_test[t_test['p']<alpha]

statistic,p_value2=mannwhitneyu_test(normal) # mann-Whitney U test
mann_test = pd.DataFrame({'feature':not_normal,'statistic': statistic,'p': p_value2})
mann_test=mann_test[mann_test['p']<alpha]
indp_common_feature=([v for v in mann_test['feature'].tolist() if v in common_feature]+
                     [v for v in t_test['feature'].tolist() if v in common_feature])
indp_BP_feature=([v for v in t_test['feature'].tolist() if v in BP_feature]+
                 [v for v in mann_test['feature'].tolist() if v in BP_feature])

chi2_score,p_value3=chi2_test(categorical_list) # chi-square test
chi_square = pd.DataFrame({'feature':categorical_list,'chi2': chi2_score,'p': p_value3})
indp_categorical_list=[v for v in chi_square['feature'].tolist() if v in categorical_list]

spearman = df[indp_common_feature].corr(method='spearman').abs() # spearman correlation
upper = spearman.where(np.triu(np.ones(spearman.shape), k=1).astype(np.bool))
to_drop = [column for column in upper.columns if any(upper[column] > 0.8)]
final_feature=[v for v in indp_common_feature if v not in to_drop]

rows=cramer_v(indp_categorical_list) # cramer's V correlation
cramers_results = np.array(rows)
cramer = pd.DataFrame(cramers_results, columns = indp_categorical_list, index =indp_categorical_list)
final_feature=final_feature+categorical_list
```

```
RR_feature,SBP_feature,DBP_feature=group_BP_feature(indp_BP_feature)
feat_imp=feature_imp(RR_feature) # feature importance
for i in range(len(RR_feature)):
    temp=feat_imp.cols.tolist()[i]
    temp=temp[3:]
    if (('SBP_'+temp) in SBP_feature) and (('DBP_'+temp) in DBP_feature):
        break
[final_feature.append(x) for x in indp_BP_feature if temp in x]
feat_imp=feature_imp(final_feature)

final_feature=feat_imp[feat_imp['imps']>0.05].cols.tolist() # final selected feature

###Feature Extraction###
df_refer=pd.concat([df[final_feature],target], axis=1)
missing_count=df_refer.isnull().sum(axis=0)
df_refer=df_refer.dropna() # drop the subjects with missing data
Y=df_refer['Fall_questionnaire']
scaler =StandardScaler()
X=scaler.fit_transform(df_refer[final_feature]) # standardization
df_scale=pd.DataFrame(data=X,columns=final_feature,index=df_refer.index)
label=Y.replace(to_replace=[1,2], value=['non-faller','faller'])

tsne = TSNE(n_components=2,perplexity=180, learning_rate=200,random_state=0) # t-SNE
X_TSNE= tsne.fit_transform(X)
e = {'tsne_1': X_TSNE[:,0], 'tsne_2': X_TSNE[:,1],'labels':label}
tsne_df = pd.DataFrame(data=e)

pca = PCA(n_components=4) # PCA
X_pca = pca.fit_transform(X)
explained_variance = pca.explained_variance_ratio_
```

```
d = {'pca_1': X_pca[:,0], 'pca_2': X_pca[:,1],'labels':label}
pca_df = pd.DataFrame(data=d)

plt.figure()
sn.scatterplot(x="pca_1", y="pca_2",hue="labels",data=pca_df,palette=['dodgerblue','red'],legend="full")#
visualisation of pca
plt.figure()
sn.scatterplot(x="tsne_1",
y="tsne_2",hue="labels",data=tsne_df,palette=['dodgerblue','red'],legend="full")# visualisation of t-SNE

###Clustering###
ks = range(2, 10)
inertia_pca,inertia_tsne,silhouette_coefficients_pca,silhouette_coefficients_tsne=[],[],[],[]
for k in ks: # evaluate the suitable number of components
    model = KMeans(n_clusters=k,random_state=0)
    model.fit(X_pca)
    inertia_pca.append(model.inertia_)
    scores = silhouette_score(X_pca, model.labels_)
    silhouette_coefficients_pca.append(scores)

    models=KMeans(n_clusters=k,random_state=0)
    models.fit(X_TSNE)
    inertia_tsne.append(models.inertia_)
    score = silhouette_score(X_TSNE, models.labels_)
    silhouette_coefficients_tsne.append(score)

k_class=KMeans(n_clusters=4,random_state=0).fit_predict(X_TSNE)
h_class=AgglomerativeClustering(n_clusters = 4).fit_predict(X_TSNE)
cntr, u, u0, d, jm, p, fpc = fuzz.cluster.cmeans(X_TSNE.T, 4, 2, error=0.005, maxiter=1000,seed=20)
f_class = np.argmax(u, axis=0)
```

```
###Characteristic Interpretation###
Overview={}
difference,score1,score2,name=[],[],[],['group1','group2','group3','group4']
for y, x in enumerate([k_class,h_class,f_class]):
    group,fall_rate,feature_median,gender,feature_IQR,fall_risk,count,rows=[],[],[],[],[],[],[],[]
    clusters = np.unique(x) # number of clusters
    df_refer['class']=x
    for i in clusters:
        group.append(df_refer[df_refer['class'] == i])
    for i,j in enumerate(group):
        fall_rate.append(group[i]['Fall_questionnaire'].value_counts(1))
        feature_median.append(group[i].median())
        feature_IQR.append(group[i].quantile(.75)-group[i].quantile(.25))
        count.append(len(j))
    for i in range(len(fall_rate)):
        fall_risk.append(1-fall_rate[i][1])
    Overview["overview%s" %y] = pd.DataFrame(feature_median)
    Overview["overview%s" %y].insert(0, "Fall risk", pd.Series(fall_risk))
    Overview["overview%s" %y].insert(0, "Count", pd.Series(count))
    Overview["overview%s" %y]=Overview["overview%s" %y].drop(['Fall_questionnaire','class'],axis=1)
    Overview["overview%s" %y]=Overview["overview%s" %y].T
    std=['std%s' %s for s in range(len(feature_IQR))]
    order=[a for a in range(len(feature_IQR)*2) if a % 2 != 0]
    for a,b,c in zip(feature_IQR,std,order):
        Overview["overview%s" %y].insert(c, b, a)
    difference.append(max(count)-min(count)) # maximum difference
    score1.append(silhouette_score(X_TSNE, x))
    score2.append(davies_bouldin_score(X_TSNE, x))
overview_new = {'overview0':'k_mean', 'overview1':'hierachical', 'overview2':'fuzzy_c_mean'}
```

```
Overview=dict((overview_new[key], value) for (key, value) in Overview.items())
Cluster_validation=pd.DataFrame([difference,score1,score2],index=['maximum_difference',

'silhouette_score','davies_bouldin_score'],columns=['k_mean','hierachical','fuzzy_c_mean'])

###Clustering Result###
plt.figure()
colour=['#1f77b4', '#ff7f0e','#d62728','#2ca02c']
for cluster,colour in zip(clusters,colour):
    row_ix = np.where(k_class == cluster)
    plt.scatter(X_TSNE[row_ix, 0], X_TSNE[row_ix, 1],label=cluster,c=colour)
plt.xlabel('tsne_1')
plt.ylabel('tsne_2')
plt.legend()
plt.show()
```