An Intelligent Recommender System Based on Short-term Disease Risk Prediction for Patients with Chronic Diseases in a Telehealth Environment

A Thesis submitted by

Raid Luaibi Lafta

For the award of

Doctor of Philosophy

2018

Abstract

Clinical decisions are usually made based on the practitioners' experiences with limited support from data-centric analytic processes from medical databases. This often leads to undesirable biases, human errors and high medical costs affecting the quality of services provided to patients. Recently, the use of intelligent technologies in clinical decision making in the telehealth environment has begun to play a vital role in improving the quality of patients' lives and reducing the costs and workload involved in their daily healthcare. In the telehealth environment, patients suffering from chronic diseases such as heart disease or diabetes have to take various medical tests such as measuring blood pressure, blood sugar and blood oxygen, etc. This practice adversely affects the overall convenience and quality of their everyday living.

In this PhD thesis, an effective recommender system is proposed utilizing a set of innovative disease risk prediction algorithms and models for short-term disease risk prediction to provide chronic disease patients with appropriate recommendations regarding the need to take a medical test on the coming day.

The input sequence of sliding windows based on the patient's time series data, is analyzed in both the time domain and the frequency domain. The time series medical data obtained for each chronicle disease patient is partitioned into consecutive sliding windows for analysis in both the time and the frequency domains. The available time series data are readily available in time domains which can be used for analysis without any further conversion. For data analysis in the frequency domain, Fast Fourier Transformation (FFT) and Dual-Tree Complex Wavelet Transformation (DTCWT) are applied to convert the data into the frequency domain and extract the frequency information.

In the time domain, four innovative predictive algorithms, Basic Heuristic Algorithm (BHA), Regression-Based Algorithm (RBA) and Hybrid Algorithm (HA) as well as a structural graph-based method (SG), are proposed to study the time series data for producing recommendations. While, in the frequency domain, three predictive classifiers, Artificial Neural Network, Least Squares-Support Vector Machine, and Naive Bayes, are used to produce the recommendations. An ensemble machine learning model is utilized to combine all the used predictive models and algorithms in both the time and frequency domains to produce the final recommendation.

Two real-life telehealth datasets collected from chronic disease patients (i.e., heart disease and diabetes patients) are utilized for a comprehensive experimental evaluation in this study. The results show that the proposed system is effective in analyzing time series medical data and providing accurate and reliable (very low risk) recommendations to patients suffering from chronic diseases such as heart disease and diabetes.

This research work will help provide high-quality evidence-based intelligent decision support to clinical disease patients that significantly reduces workload associated with medical checkups would otherwise have to be conducted every day in a telehealth environment.

Certification of Thesis

This Thesis is entirely the work of **Raid Luaibi Lafta** except where otherwise acknowledged. The work is original and has not previously been submitted for any other award, except where acknowledged.

Student and supervisors signatures of endorsement are held at the University.

Principal Supervisor: Ji Zhang

Associate Supervisor: Xiaohui Tao

Associate Supervisor: Yan Li

Acknowledgments

I would like to express my deepest appreciation to my Principle Supervisor, Associate Professor Ji Zhang, for his useful guidance, valuable support, insightful comments, and considerable encouragement in completing this thesis. I would also like to express my deepest heartfelt appreciation to my co-supervisors Dr. Xiaohui Tao and Professor Yan Li for their valuable support, advice and effective guidance.

Many thanks to the Ministry of Higher Education and Scientific Research in Iraq for their financial support by offering the PhD scholarship. Their support is gratefully acknowledged.

I would also like to extend my appreciation to my parents, sisters and brothers who have given me everlasting encouragement and support during the completion of the thesis. I am specially thankful to my wife for her patience, love and understanding.

Also, I wish to thank Sandra Cochrane for proofreading.

Finally, I would like to acknowledge the support by Australian Commonwealth Government through the Research Training Program (RTP) Fees Offset scheme during my research.

Raid Luaibi Lafta

Associated Publications

The following publications were produced during the period of candidature:

1. Zhang, J., Lafta, R., Tao, X., Li, Y., Zhu, X., Luo, Y. and Chen, F. (2017), “Coupling a fast fourier transformation with a machine learning ensemble model to support recommendations for heart disease patients in a telehealth environment”, *IEEE Access*, Vol.5, pp. 10674 - 10685.
2. Lafta, R., Zhang, J., Tao, X., Li, Y., Abbas, W., Luo, Y., Chen, F. and Tseng, V. S. (2017), “A fast fourier transform-coupled machine learning-based ensemble model for disease risk prediction using a real-life dataset”, *Proceeding of the Pacic-Asia Conference on Knowledge Discovery and Data Mining, South Korea.*
3. Lafta, R., Zhang, J., Tao, X., Li, Y. and Tseng, V. S. (2016), “IRS-HD: an intelligent personalized recommender system for heart disease patients in a tele-health environment”, *Proceeding of the 12th International Conference on Advanced Data Mining and Applications (ADMA), Gold Coast, QLD, Australia.*
4. Lafta, R., Zhang, J., Tao, X., Li, Y., Tseng, V. S., Luo, Y. and Chen, F. (2016), “An intelligent recommender system based on predictive analysis in telehealthcare environment”, *Web Intelligence*, Vol. 14, *IOS Press*, pp. 325-336.

5. Lafta, R., Zhang, J., Tao, X., Li, Y. and Tseng, V. S. (2015), "An intelligent recommender system based on short-term risk prediction for heart disease patients", *Proceeding of the Web IEEE/WIC/ACM International Conference on Intelligence and Intelligent Agent Technology (WI-IAT), Singapore.*

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Most healthcare organizations, such as hospitals and medical centers, generate huge volumes of data in the form of text, numbers, charts and images. Clinical decisions are usually made based on practitioners' experiences, with limited support from medical databases. Quite often this leads to undesirable biases, human errors and high medical costs affecting the quality of services provided to patients. There is strong evidence that clinical decisions based on risk assessment can improve chronic disease management (Ali et al., 2016). In telehealth environments, the continuous remote monitoring of patients' key measurements generates a lot of time series data. Furthermore, short term prediction is more difficult than long-term projection due to the high levels of uncertainty when reading these measurements.

Given our increasingly aging society, telehealth care is becoming more popular, delivering health-related services via telecommunication technologies such as the telephone or the Internet (O'Neil et al., 2014; Hawkes et al., 2013). Evidence shows that tele-health care could save money and time by reducing emergency department visits, hospital admissions, associated travel and physical limitations (Lu, Chi and Chen, 2014). Intelligent technologies can be developed to greatly facilitate the development and deployment of telehealth systems for patients, especially those who suffer from chronic disease and require continuous monitoring and chronic-related medical measurements.

These intelligent technologies will improve the quality of clinical evidence-based decisions made by medical practitioners and help reduce the financial and time costs experienced by patients, as well as help reduce workload from unnecessary medical checkups.

1.1 Chronic Diseases

Chronic disease, such as heart failure or diabetes, is a major health problem in the world and the rates of incidence are rising significantly. Over 57 million people around the world died in 2008, and 58% (33 million) of these deaths were due to chronic diseases (Alwan et al., 2010). Chronic diseases also impose huge costs on the health care systems responsible for their prevention and treatment. In the US, the direct health care system costs for patients with chronic diseases, especially patients with diabetes, asthma, heart disease and hypertension, were around $52.1 billion in 1996 (Druss et al., 2001). Therefore, chronic diseases require more and more medical attention and resources in today's increasingly aging societies.

1.2 Time Series Data in a Telehealth Environment

Time series data can be defined as a collection of observations obtained through repeated measurements over time (Chatfield, 2016). In some cases, time series data is a set of recorded values about one variable taken at evenly spaced time points. In the health care domain, the sequence of medical body tests is a time series. A time series can be generated in the fields of science, health care, engineering, economics and sociology. Telehealth services can provide real-life time series data for patients' day-to-day medical readings of different measurements.

Time series analysis is a statistical process which deals with techniques developed to draw important inferences from time series. The initial step of time series analysis is the choice of an appropriate model for the data. Based on the analysis of the previous time series, a satisfactory model will forecast future values of the time series.

The huge volumes of generated time series data also grow quickly. However, the analysis and management of time series data still pose a great challenge to medical practitioners and patients who have little knowledge and experience of time series analysis models. Although, many analytic tools exist for time series analysis, a lot of them are designed for data experts and analysts.

1.3 Recommendations for Patients with Chronic Diseases in a Telehealth Environment

Recommendation systems are computer-based information systems designed to support and assist medical practitioners in the implementation evidence-based practices and improved decision-making (Njie et al., 2015). These recommendation systems can help in minimizing medical errors and providing more detailed data analysis in shorter periods of time (Thong and Le, 2015).

In the era of telehealth, many applications have been designed to help patients and medical practitioners deliver their medical information easily. These digital tools allow patients and practitioners to exchange personalized reminders and medical recommendations, upload information from devices such as blood pressure cuffs or blood glucose monitors, store health records, and share information with health care providers (Van Ma et al., 2016). Health care providers can interact face-to-face and in real-time with their patients using the telehealth services (such as the Internet or video equipment) to provide appropriate recommendations for overall health conditions. Devices such as heart rate monitors and blood pressure monitors can be connected to web-based applications in a telehealth environment to exchange medical information between patients and health care providers.

Telehealth systems can be particularly useful for people with chronic diseases, as well as those who live in rural or remote areas where fewer visits to health care organizations such as hospitals and medical centers and are possible access to medical care and medical recommendations can be limited.

Telehealth devices can assess images, sounds, body motion, and vital signs such as blood pressure, heart rate and pulse, body temperature, body weight, and blood oxygenation. Thus, medical practitioners can take the advantage of technologies to provide appropriate recommendations to their patients.

Telehealth systems are becoming increasingly popular and are an effective means of coping with the challenges posed by the care of patients with chronic diseases and, therefore have enjoyed fast development in many countries due to their low-cost and fast service delivery. Most telehealth services are delivered through web-based applications which use web browsers and the Internet, together with sensors, mobiles and wearable devices.

Given the importance of disease risk prediction in the medical domain (Zhang et al., 2017) as well as the urgency of obtaining more effective analytic techniques for disease risk prediction, great efforts are needed to enhance the quality of evidence-based decisions and recommendations in the telehealth environment. Patients with chronic disease need to undertake various daily medical tests to monitor their overall chronic health conditions through the telehealth system. Carrying out various medical tests every day brings lots of inconvenience, even burden, to patients and adversely affects their life quality. Producing accurate intelligent recommendations to guide their daily medical tests can effectively reduce the workload associated with taking those tests while keeping the associated health risk at an acceptably low level.

1.4 Research Objectives

This research aims to develop an intelligent recommendation system to provide patients in a telehealth environment with appropriate recommendations for the need to take a medical body test on the coming day. The proposed system aims to improve the quality of clinical evidence-based decisions and reduce the financial and time costs incurred by patients. This project develops an intelligent recommendations system which uses an innovative time series prediction models to provide useful recommendations to chronic disease patients in the telehealth environment. The system can be used by both patients and medical practitioners to improve their decision making processes and to reduce the workload associated with unnecessary tests for patients. It also offers an effective way to reduce the risk of incorrect recommendations. Specifically, this research is designed to achieve the following three major objectives:

1. Develop an intelligent recommendation system supported by several time and frequency domain analysis techniques for the efficient and effective prediction of short-term risk associated with chronic diseases based on time series medical data analysis.
2. Implement recommendation techniques and develop a system prototype that interacts with patients and provides recommendations based on the assessment

of the disease risk.

3. Evaluate the performance, including both efficiency and effectiveness, of the proposed system using real-life datasets and conduct a comparative study with existing recommendation models.

1.5 Research Questions

This study attempts to answer the following questions:

1. How to effectively assess and predict the short-term risk of chronic disease?
2. How to effectively provide accurate and reliable recommendations to patients with chronic diseases?

1.6 Outcomes and Results

In this study, the time series data of patients are employed for short-term disease risk prediction using the time and frequency analysis techniques. Several techniques are applied to assess short-term disease risk of patients in a telehealth environment based on the analysis of a patient's historical medical data. Three innovative predictive algorithms with a developed structural graph-based method are used for the time domain analysis in time domain. Two existing decomposition methods, fast Fourier transformation and dual tree complex wavelet transformation, are developed for time series analysis in the time frequency.

Figure 1.1 illustrates the overall architecture of the recommendation system. The proposed system produces accurate recommendations and the assessment of short-term risk for patients. The study leads to many different advantages for medical practitioners and patients including:

- Improving the quality of clinical evidenced-based decisions made by medical practitioners and reduction of the financial and time cost taken by patients
- Offering a promising tool for analyzing time series medical data and providing accurate and reliable recommendations to patients
- Contributing the better monitoring, assessment and management the disease risk for patients
- Reducing the workload in medical checkups and the potential risk of incorrect recommendations for patients who must conduct body tests every day

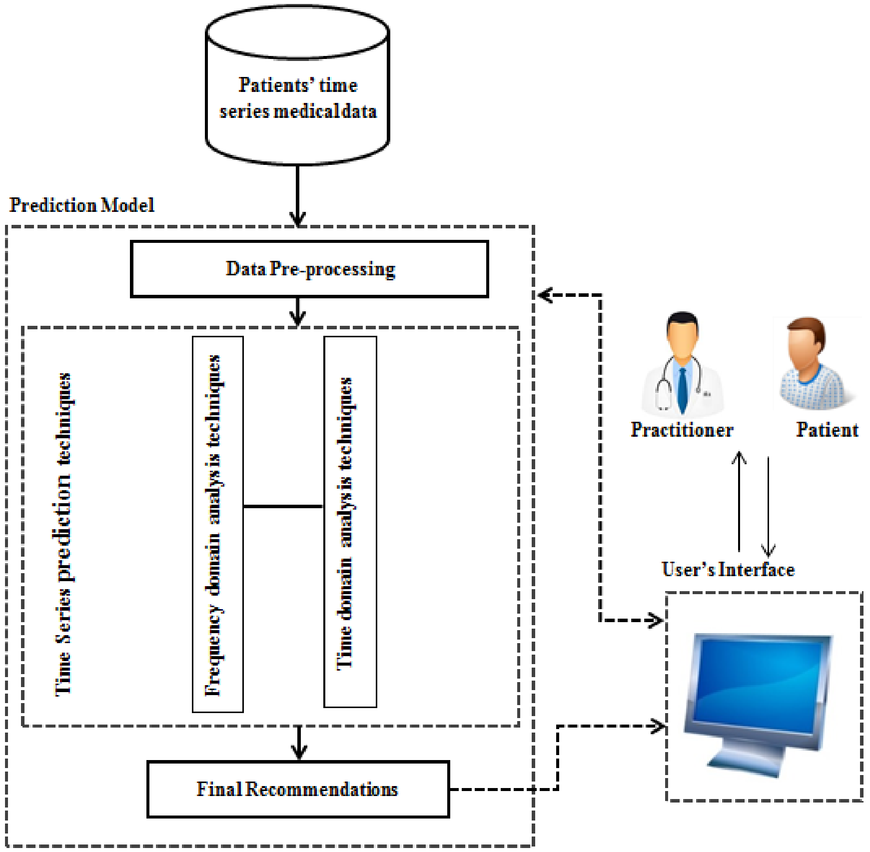


Figure 1.1: Architecture of the recommendation system

1.7 Statement of the Problems for Disease Risk Prediction

The goal of this thesis is to develop an intelligent system to analyze patients' time series medical data to assess the short-term risk of disease and provide quality decision-making support, thus reducing the associated cost without sacrificing the quality of healthcare services.

The problem can be formulated mathematically. Let D_{med} be the full set of observations including measurements and symptoms, and $T = t_1, t_2, ..., t_m$ represent the full set of timestamps in the collection. D_{med} is a m × (n+1) table where $D_{med}(i, j)$ means the jth observations at the ith timestamp. Conceptually, the prediction model can also be understood as the following mapping function:

$$f(D_{med}, k) \longrightarrow R \tag{1.1}$$

$$f(R) \longrightarrow Recommendations \tag{1.2}$$

The first function is the risk prediction function in which observations over the last k timestamps are analyzed to predict the risk of disease; R represents the disease risk level. The second function is the recommendation function that provides recommendations to patients based on the outcome of the risk prediction function.

1.8 Structure of the Thesis

Figure 1.2 illustrates the logical structure of the dissertation. The dissertation is organised as follows:

Chapter 1 includes four major topics: the time series data in a telehealth environment, telehealth environment recommendations for patients with chronic diseases, objectives of the research, research questions, statement of the problems, and outcomes and results.

Chapter 2 provides a comprehensive literature review of disease risk assessment in a health domain including the chronic disease risk assessment, basic methods and techniques, limitations of the existing analytic tools and potential improvements.

Chapter 3 presents the new methods proposed to improve the predictive performance of the system using three innovative predictive algorithms, a structural graph-based method, the fast Fourier transformation (FFT), and dual-tree complex wavelet transformation (DTCWT).

Chapter 4 describes the datasets used and the experimental evaluation of the proposed system and elaborates on the results of the experiments conducted to evaluate the performance of the system for short-term disease risk prediction using the time and frequency analysis techniques.

Chapter 5 provides a summary of the main outcomes of this study and highlights future research directions.

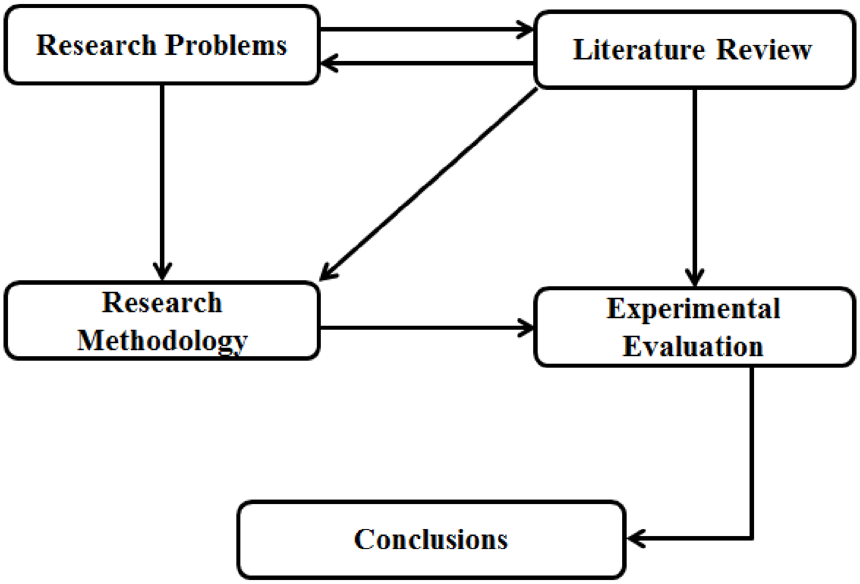


Figure 1.2: Logical structure of thesis

Chapter 2

A Literature Review of Disease Risk Prediction

This chapter reviews related state of the art work in predictive data analytics and mining, as well as their applications in the telehealth environment.

2.1 Chronic Disease Risk Assessment

Chronic diseases have been one of the main public health concerns worldwide, accounting for over 50% of global mortality (Kuh and Shlomo, 2004) and thus require more and more medical attention and resources in today's increasingly aging societies. As patients with chronic diseases require frequent medical attention, the lack of automatically generated medical recommendation systems is currently limiting options for improvements in treatment and care, thus affecting patients' quality of life. In the last decades, much research effort has been invested in the assessment of different disease risk in the healthcare domain. Many of these research studies have been carried out in data mining and analytics on medical data (Reddy et al., 2018; Nilashi et al., 2017; Braamse et al., 2016; Sánchez et al., 2016; Beaney et al., 2015; Mohktar et al., 2015; Yang et al., 2014; Lu, Huang, Su and Lin, 2014; Krishnaiah et al., 2013; Huang et al., 2012; Chang et al., 2011; Yeh et al., 2011; Polat and Güneş, 2007). The following subsections review the use of various predictive modelling techniques in medical and health domains.

2.1.1 Disease risk assessment

Extensive research work has been carried out in the prediction of the long-term or lifetime risk for the patients with different diseases. Statistical analyses were used by Aoki et al. (2017) to predict the development of hepatocellular carcinoma within five years in patients with chronic liver disease. They proposed a new score system based on combining several risk factors with high predictive accuracy. Statistical analysis such as Kaplan-Meier's, t-test and Fisher's exact test were applied by Nihtyanova et al. (2014) on patients who suffered from long-term (15 year) systemic sclerosis and assessed the survival and incidence of organ-based complications for these patients. Prediction models estimating survival and lung-based SSc complications and prediction of death, pulmonary fibrosis, and pulmonary hypertension were developed. This work basically relied on some indicators and measurements to assess survival and pulmonary complications for systemic sclerosis patients. Nobel et al. (2014) developed a personalized risk assessment tool for stroke by using statistical analysis such as the Cox proportional-hazards regression which enables a person to count their personal risk of stroke for a ten-year period based on a personal risk profile. They also estimated the risk of stroke for specific combinations of health and lifestyle-related factors. A risk prediction model to detect pregnant women at increased risk of death or major hypertensive-related complications in developing countries was developed by Payne et al. (2014). They developed a clinical prediction model for adverse maternal outcomes among women who suffer from hypertensive disorders in pregnancy for use in primary health care facilities in low and middle income countries.

However, all of these focus on long-term medical prediction. Short-term prediction, which is studied in this research, is more challenging than long-term prediction as patients' conditions may experience more dramatic and abrupt changes during a short timeframe. In addition, short-term recommendations are equally useful for patients as they provide guidance as to what the patients need to do for the next few days.

2.1.2 Prediction of the risk level of diseases

Data mining technologies and statistical analysis have been used to predict the level of disease in patients. The long-term prognostic model of primary biliary cholangitis was used by Carbone et al. (2016) to estimate the exact risk level of developing end-stage liver disease within specific time points in the future. The researchers developed their model based on corporate measures of treatment response with other prognostic variables. A Genetic Algorithm was used by Sabibullah et al. (2013) to predict the risk level for diabetes, heart attack, and stroke. This work employed laboratory measurements to predict the level of risk associated with either heart attack or stroke. A k-means clustering algorithm was used by Ahmed et al. (2013) to predict the risk level of skin cancer (Ahmed et al., 2013). The researchers relied on some risk factors and measurements to detect skin cancer risk level. A prediction model for the detection of patients at high risk of hepatocellular carcinoma was developed using decision tree analysis by Kurosaki et al. (2012). This model was constructed on the basis of simple and generally available factors to identify patients at a high risk of HCC development. Decision tree analysis was applied to make the proposed model easy for patients and practitioners to use in the clinic. Models for predicting long-term cirrhosis and hepatocellular carcinoma risk in chronic Hepatitis B Virus patients were developed by Lee et al. (2013). The models categorized participants accurately with low, medium, and high cirrhosis and HCC risk. Applying these models was totally based on measurements and relevant risk factors. Statistical analysis was applied in this study to evaluate the performance of the risk models. Boateng et al. (2018) developed cardiovascular disease (CVD) risk algorithms for sub-Saharan African populations. The Framingham laboratory, non-laboratory and Pooled Cohort Equations (PCE) algorithms were used in this study to classify individuals as high, moderate or high risk.

2.1.3 Disease risk prediction models

Samuel et al. (2017) proposed a hybrid approach based on both artificial neural networks (ANN) and the Fuzzy analytic hierarchy process (Fuzzy-AHP) technique to predict heart failure in patients. They used 13 heart failure attributes in their study and the authors' contributions were determined based on an experienced cardiac clinician. The newly proposed method achieved a high prediction accuracy compared with the conventional ANN approach using a real dataset of heart failure patients. A decision tree-based predictive model was proposed by Tayefi et al. (2017) for coronary heart disease. They used different associated risk factors including gender, age, systolic blood pressure, diastolic blood pressure, body mass index, triglyceride level, highly sensitive C-reactive protein, total cholesterol, high density lipoprotein and left anterior descending to differentiate healthy and coronary heart disease patients. Ye et al. (2018) developed a risk prediction model for essential hypertension within a year. The predictive analytic model was developed using a machine learning algorithm to predict the incidence of hypertension in the following year by analysing the electronic health records of patients. Kim et al. (2014) proposed a prediction support model to predict coronary heart disease risk in patients. The proposed model used a mining technique which was validated by medical experts to give content recommendation to coronary heart disease using knowledge-based rules.

The improvement of the overall accuracy of prediction models was the main purpose of these studies and the high accuracies of the predictive models show that these techniques have a high probability of predicting patient status quickly in order to treat patients effectively. However, a challenge in securing an effective analytical tool with high accuracy to help support personalized evidence-based decisions remains. In addition, the existing work does not deal with chronic disease using the medical measurements of patients as this study does.

2.2 Basic Concepts and Methods for Disease Risk Prediction

This section addresses the details concerning the different methods and techniques for disease risk prevention used in this research project.

2.2.1 Fast Fourier transformation of the time series data

The Fast Fourier Transformation (FFT) is an efficient technique for computing the Discrete Fourier Transform (DFT) and its inverse. It is similar to the wavelet transformation as a windowing technique (Deo et al., 2016; Kiplangat et al., 2016).

Due to its predictive effectiveness, the fast Fourier transformation has been used in many different research areas such as analyzing and forecasting electricity consumption in buildings (Tsai et al., 2016; Ji et al., 2015), detecting epileptic seizures in electroencephalography (EEG) (Samiee et al., 2015; Kovacs et al., 2014) and forecasting water demand (Brentan et al., 2017; Odan and Reis, 2012).

The DFT decomposes the input data sequence (i.e., the data in the sliding windows of the given time series data of the patients) to extract frequency information for the purpose of predicting the patient's condition one day in advance.

Let $x(t)$ be a time series bounded by a sliding window. The Discrete Fourier Transformation of $x(t)$ can be defined as:

$$X(c^{jw}) = \sum_{t=-\infty}^{\infty} x(t) c^{-jwt} \tag{2.1}$$

where t is a discrete time index and w refers to the frequency. There are T input time series $x(t)$, so the transform pair of the DFT can be defined as:

$$X(P) = \sum_{t=0}^{T-1} x(t)\, W_T^{tp} \Leftrightarrow x(t) = \frac{1}{T} \sum_{p=0}^{T-1} X(P)\, W_T^{-tp}, \qquad where \quad W = c^{-j2\Pi/T} \tag{2.2}$$

Furthermore, the DFT can be presented as a discrete-time Fourier transformation of a cyclic signal with period T as:

$$x = \begin{bmatrix} x(0) \\ x(1) \\ \vdots \\ x(T-1) \end{bmatrix}, \quad X = \begin{bmatrix} X(0) \\ X(1) \\ \vdots \\ X(T-1) \end{bmatrix} \tag{2.3}$$

$$W = [W_T^{pt}] = \begin{bmatrix} 1 & 1 & \dots & 1 \\ 1 & W_T & \dots & W_T^{T-1} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & W_T^{T-1} & \dots & W_T^{(T-1)(T-1)} \end{bmatrix} \tag{2.4}$$

The following equation presents the relationship between x and X:

$$X = Wx \Leftrightarrow x = \frac{1}{T} W^H X \tag{2.5}$$

Based on the above equations, the DFT matrix W requires T^2 complex multiplications for a given time series input signal $x(t)$ with a length of T. Therefore, the required implementation cost for factorizing the fast Fourier transformation W into a matrix is lower than that of the direct Fourier transformation since each stage of the fast Fourier transform only requires $T/2$ multiplications and T additions (Cooley et al., 1969; Mertins, 1999).

2.2.2 Dual tree complex wavelet transformation

The Discrete Wavelet Transformation (DWT) is a powerful tool for analyzing signals in the time-frequency domain. Despite its ability to provide an efficient time-frequency analysis of non-stationary signals, the DWT suffers from several problems such as aliasing, shift variance and lack of directionality (Kingsbury, 1998).

Over previous decades, there has been a growing interest in using transformation techniques such as a wavelet, Fourier and hybrid transformations to analyze different time series data, such as biomedical. One of the developed transformation techniques is Dual-Tree Complex Wavelet Transformation (DTCWT). It is an improved version of wavelet transformation designed to tackle some limitations in the wavelet transformation. Due to its predictive effectiveness, the dual-tree complex wavelet has been successfully utilized in many different research areas such as the processing and de-noising of images (Yu et al., 2016; Jung et al., 2017), and detecting epileptic seizures in electroencephalography (EEG) (Swami et al., 2017; Li et al., 2017).

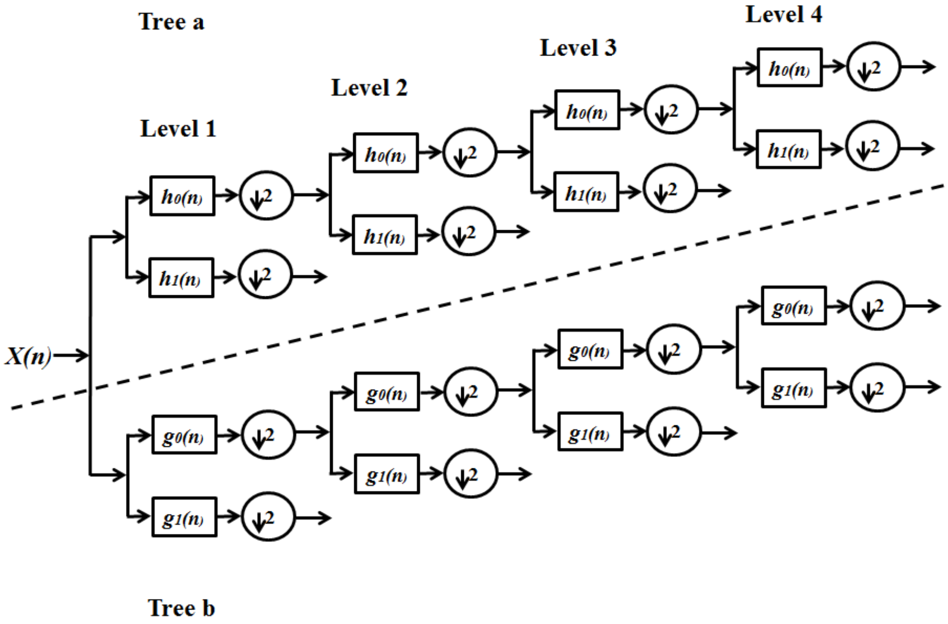


Figure 2.1: 1-D dual-tree complex wavelet transformation

The drawbacks of DWT are ameliorated by using the dual tree complex wavelet transformation (DTCWT) which offers a better time-frequency representation of signals (Das et al., 2016). Basically, the dual tree complex wavelet transformation employs two real DWT trees (Yu et al., 2016), where the top tree represents the real part of the complex wavelet coefficient whilst the tree at the bottom represents the imaginary part, as illustrated in Figure 2.1.

Therefore, the components obtained after four levels of decomposition include five sub-bands y1, y2, y3 ,y4 and z4 for each part (tree). In this study's system, DTCWT is adopted to decompose the input time series data into sub-bands of delta, theta, alpha, beta and gamma. Figure 2.2 shows four-level decomposition of DTCWT. Each DTCWT coefficient has two parts: real and imaginary. As a result, ten sub-bands in total are obtained after four-level decomposition (five sub-bands for each part). In the present work, the DTCWT sub-bands for both real and imaginary parts are presented by (y1,1),(y1,2),(y2,1),(y2,2),(y3,1),(y3,2),(y4,1),(y4,2),(z4,1) and (z4,2).

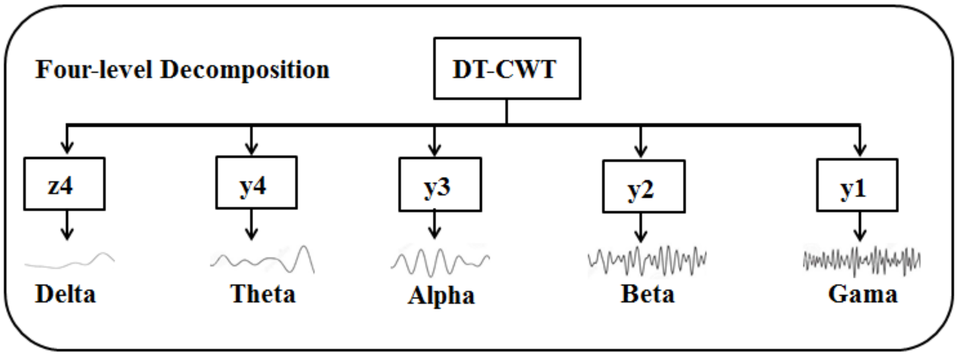


Figure 2.2: Four-level decomposition of WTCWT

Each coefficient sub-band has two parts, for example (y1,1),(y1,2) represent the y1 sub-band for the real and imaginary parts respectively.

2.2.3 Graph theory

Graphs can be mathematically defined as abstract representations of networks that consist of a set of nodes linked by edges (Micheloyannis et al., 2006). In previous years, graph theory has been increasingly used in the analysis and classification of complex network relationships such as social networks, biological and brain networks, signal and image processing. It is used in neuroscience research to analyze and study brain diseases (Panzica et al., 2013; Vural and Yildiz, 2010; Wang et al., 2014). Some studies (He and Evans, 2010; Li, Hu, Jin, Han, Liu, Guo, Hao and Li, 2013; Stam and Reijneveld, 2007) have shown that graph theory can be considered as a robust tool to characterize the functional topological properties of brain networks for both normal and abnormal brain functioning (Lang, 2017; Vecchio et al., 2017). It is also used in image processing as a powerful tool to analyze and classify digital images (Sarsoh et al., 2012). The time series of EEG signals are converted into graphs by Diykh and Li (2016) and Diykh et al. (2016) for EEG sleep stages classification.

2.2.4 Machine learning ensembles

The intelligent, accurate medical recommendations in this work rely on the use of classification approaches to produce reliable predictions of short-term medical risks. By nature, this is a classification problem which involves using classification methods (called classifiers) to predict the necessity of taking a body test of a given medical measurement. In this work, machine learning techniques are combined to build an ensemble classifier.

There are many advantages for using the ensemble classier in the medical field. First, it provides an efficient solution for building a single model for applications of very large volumes of data (Valentini and Masulli, 2002). Second, it has been proven to be an effective tool thanks to its ability to improve the overall accuracy of the prediction model. Empirical results show that machine learning ensembles are often more accurate than the individual classifiers that make them up (Valentini and Masulli, 2002). Bagging aggregation is a machine learning ensemble algorithm designed to enhance the accuracy and stability of machine learning algorithms (Voyant et al., 2017; Bashir et al., 2015), and was proposed by Breiman in the mid-1990's (Breiman, 1996). It has been proven to be a very popular, efficient and effective method for building ensemble models.

2.3 Limitations of Existing Time Series Prediction Models

Although many time series prediction models have been effectively applied in the medical domain to assess the risk of chronic diseases, a challenge remains in securing an effective analytical tool with high accuracy to help make high quality, efficient evidence-based decisions. A review of the literature shows the main limitations of time series prediction models to be:

- Most existing work focuses on long-term medical prediction. Nevertheless, short-term prediction, which is studied in this work, has turned out to be more challenging than long-term prediction as patients' conditions may experience more dramatic and abrupt changes during a short-term timeframe
- None of the existing work approached or solved the problem dealt with in this research: to provide accurate recommendations to chronic disease patients about whether they should to take or skip the medical tests on the coming day
- Most of the existing work analyzes time series data using only time domain analysis techniques while this study employs both the time and frequency analysis techniques for analysing the time series medical data of patients
- None of the existing work use the medical measurements (or medical tests) of patients in the prediction of the short-term disease risk in a telehealth environment.

Chapter 3

Disease Risk Prediction Algorithms and Models

This chapter introduces a set of disease risk prediction algorithms and the models proposed in the system for short-term disease risk prediction. The time and frequency domains analysis methods are applied to provide patients with appropriate recommendations about the need to take a medical test or not on the coming day based on the analysis of their medical data.

The project investigates the effectiveness of the recommendation system in analysing the time series data of patients in both the time domain and the time-frequency domain, thus providing patients in a telehealth environment with appropriate recommendations for taking a medical body test on the coming day. The time domain refers to variation of amplitude of signal with time, while the frequency domain refers to how much signals lie in the frequency range (Machac et al., 1988). Three innovative predictive algorithms, the Basic Heuristic Algorithm (BHA), Regression-Based Algorithm (RBA) and Hybrid Algorithm (HA), are proposed as well as a structural graph-based method (SG) to study the behavior of time series data in the time domain for the short-term risk assessment of patients in a telehealth environment based on the analysis of a patient's historical medical data. In the frequency domain, the fast Fourier transformation and dual-tree complex wavelet transformation are applied to study the behaviour of the time series data of patients. A bagging-based ensemble model is utilized to predict the patient's condition one day in advance for producing the final recommendation. A combination of three classifiers – Artificial Neural Network, Least Squares-Support Vector Machine, and Naive Bayes – is used to construct an ensemble framework. The conceptual framework of the study is presented in Figure 3.1.

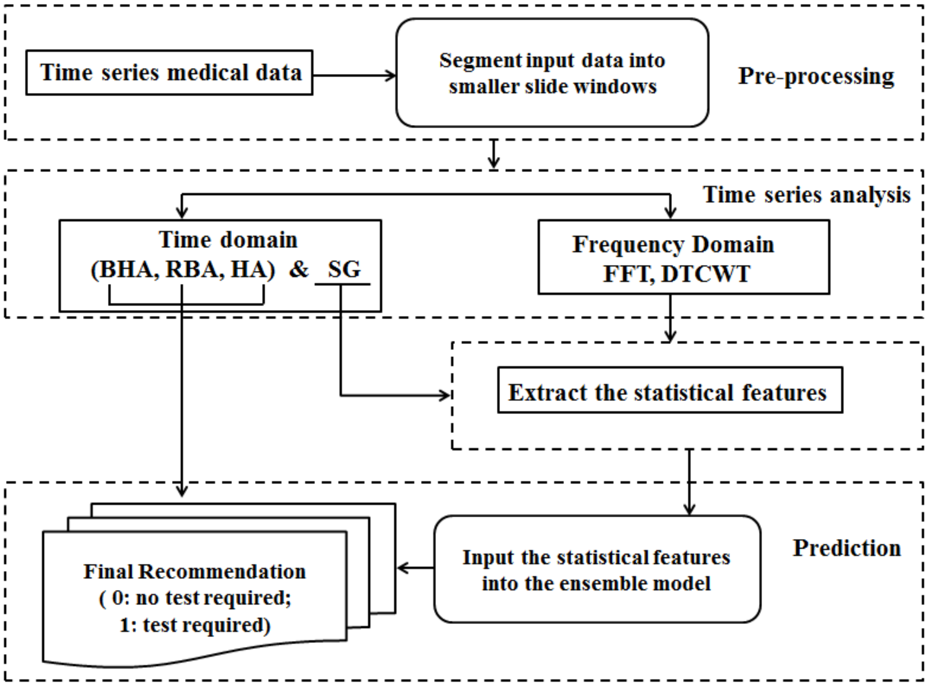


Figure 3.1: Diagram for methodology

3.1 Time Domains Analysis Algorithms for Short-Time Disease Risk Prediction

This section begins by presenting an overview of the three innovative predictive algorithms used for the short-term risk assessment of patients (in a telehealth environment), and then provides a detailed structural graph-based method; the major technical components of the system developed to analyze the time series data of patients in the time domain.

The key component of the proposed system is the recommendation algorithms based on time series data analysis. The system is based on time series prediction algorithms which predict short-term disease risk for patients, and facilitates the generation of recommendations as to whether a particular patient needs to take the physical body test on the current date. The basic heuristic algorithm is used to analyze the medical data of patients, assess their risk and provide them with appropriate recommendations for the necessity of taking a medical test on the following day based on the outcomes of prediction.

3.1.1 Basic Heuristic Algorithm

A basic heuristic algorithm is developed to decide whether a given patient needs to take a medical measurement, such as the heart rate test, today based on a study of his/her measurement readings for the past k days. If the patient satisfies both of the following two heuristic rules for a measurement, a recommendation of "no test required" will be generated, and the patient does not need to take the test for that measurement on the following day:

- *She (he) has taken the test for no less than $p\%$ of the past k days for this* *measurement (*$0 <= p <= 100$*), and*

- *All the readings of this measurement during the past k days are normal.*

The "no test required" recommendation will be provided to the patient and stored in the backend database as a part of the patient's historical records. If any of the conditions is not satisfied, a recommendation of "test required" will be generated and the patient will be advised to take the medical test on the following day. Again, the recommendation will be stored in the system as a historical record.

Overall, there are four parameters in the recommendation algorithm, i.e., the minimum (min) and maximum (max) of normal values for each measurement, setting up the boundary of healthy range; the length of the sliding time window k, and the minimum percentage (p) of days when the medical test is conducted for the measurement in the past k days. The recommendation algorithm is presented in Algorithm

1.

Algorithm 1: Basic Time Series Prediction Algorithm

```
Input   : Patient's time series medical testing data (e.g., heart rates).
Output: Rick = [0 | 1] (0: low risk; 1: high risk); Recomm = [0 | 1] (0: no
          test required; 1: test required.)
1  let $k$ be a limited number from the past days (the length of time window $k$);
2  let $p\%$ be a value between the range ($0 <= p <= 100$);
3  let max and min be the boundary of healthy values in the test;
4  foreach days for the patient do
5      if the patient has taken the test for no less than $p\%$ of the past $k$ days
         ($<= k - (p \times k)$) AND all the measured values during the past $k$ days are
         normal then
6          Risk = 0;
7      else
8          Risk = 1;
9      end
10     if Risk == 0 then
11         Recomm = 0;
12     else
13         Recomm = 1;
14     end
15 end
16 return Recomm.
```

The algorithm of the recommender system is presented in Algorithm 1. It evaluates the time series data collected for a patient on a continuous basis using a sliding window with a length of k, which is the number of days in the past that the algorithm will look at in support of an assessment evaluation for the following day. Two conditions are evaluated in the *IF* predicate from Lines 5 to 14. The first evaluates the percentage of the actual medical tests that have been carried out in the past k days. If a test is skipped for a day, the actual reading will be missing and, as a result, the certainty and accuracy for risk assessment will drop for future days. Therefore, in this condition, an upper bound is imposed on the total number of days when the medical testing is skipped in each sliding window. In addition to this bound, we also require that the readings of all the medical checkups conducted during the past k days are in the normal range for the measurement, as dictated by its corresponding minimum and maximum threshold values. Intuitively speaking, normal readings improve the confidence that the short-term risk is low, whereby a skip of the test on this measurement can be recommended. If both two conditions are satisfied, then the risk for skipping the physical test for this measurement is deemed low and a recommendation for skipping a test for the measurement can be made for the following day. Otherwise, the system will provide a recommendation urging patients to take a medical test for the measurement on the following day and the reading of the measurement will be received and stored in the database. It is worth mentioning that the above algorithm will be applied to a single measurement once at the time for risk assessment and recommendation. If the recommendations are to be performed for multiple measurements, then the system simply runs each separately to produce recommendations for different measurements.

3.1.2 Regression-Based Algorithm

Regression analysis is a simple method for investigating functional relationships among variables (Chatterjee and Hadi, 2015). Let Y be the response variable and $X_1, X_2, X_3, ..., X_p$ be the set of independent variables, where p is the number of independent variables. The relationship between the response variable and independent variables can be approximated by the regression model:

$$Y = f(X_1, X_2, X_3, ..., X_p) + \epsilon \tag{3.1}$$

where ϵ is a random error representing the discrepancy in the approximation.

The following equation illustrates the linear regression model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + ... + \beta_p X_p + \epsilon \tag{3.2}$$

where $\beta_0, \beta_1, \beta_2, ..., \beta_p$ are unknown constants to be estimated from the data.

The second version of the predictive algorithm is developed based on the linear regression model. This involves modifying the second rule of the basic heuristic algorithm by leveraging the prediction generated by the linear regression model, as follows:

- *She (he) has taken the test for no less than $p\%$ of the past k days for this measurement ($0 <= p <= 100$), and*
- *The predicted value for the current day generated by the linear regression based on the past k days is normal.*

Compared with the basic heuristic algorithm, only the second rule has been modified in the regression-based algorithm. Using the data from the past k days, a linear regression model is generated for the prediction of the value of the measurement in question for the next day. The predicted measurement value must be normal to produce the "no test required" recommendation which will be provided to the patient and stored in the backend database as a part of the patient's historical records. If either of the two conditions is not satisfied, a recommendation of "test required" will be generated and the patient is advised to take the medical test on the following day. Again, the recommendation will be stored in the system as a historical record. The regression-based algorithm is presented in Algorithm 2.

Algorithm 2: Regression-Based Time Series Prediction Algorithm

```
Input  : Patient's time series medical testing data (e.g., heart rates).
Output: Rick = [0 | 1] (0: low risk; 1: high risk); Recomm = [0 | 1] (0: no
        test required; 1: test required.)
let $k$ be a limited number from the past days (the length of time window $k$);
let $p\%$ be a value between the range ($0 <= p <= 100$);
let max and min be the boundary of healthy values in the test;
foreach days for the patient do
    if the patient has taken the test for no less than $p\%$ of the past $k$ days
      ($<= k - (p \times k)$) AND the predicted measurement value for the current day
      based on the linear regression model is normal $<$ max then
        Risk = 0;
    else
        Risk = 1;
    end
    if Risk == 0 then
        Recomm = 0;
    else
        Recomm = 1;
    end
end
return Recomm.
```

3.1.3 Hybrid Algorithm

By combining both the basic heuristic and the regression-based algorithms, a new hybrid algorithm is obtained. The second condition in the basic heuristic and the regression-based algorithms are combined to generate a new condition as follows:

- *She (he) has taken the test for no less than p% of the past k days for this measurement (*$0 <= p <= 100$*), and*
- *All the readings of this measurement during the past k days are normal and the predicted value for the current day generated by the linear regression based on the past k days is normal.*

The hybrid recommendation algorithm is demonstrated in Algorithm 3.

Algorithm 3: Hybrid Time Series Prediction Algorithm

```
Input  : Patient's time series medical testing data (e.g., heart rates).
Output: Rick = [0 | 1] (0: low risk; 1: high risk); Recomm = [0 | 1] (0: no
        test required; 1: test required.)
let $k$ be a limited number from the past days (the length of time window $k$);
let $p$% be a value between the range ($0 <= p <= 100$);
let max and min be the boundary of healthy values in the test;
foreach days for the patient do
    if the patient has taken the test for no less than $p$% of the past $k$ days
      ($<= k - (p \times k)$) AND the predicted measurement value for the current day
      based on the linear regression model is normal AND all the measured value
      $k$ days are normal then
        Risk = 0;
    else
        Risk = 1;
    end
    if Risk == 0 then
        Recomm = 0;
    else
        Recomm = 1;
    end
end
return Recomm.
```

3.1.4 A Structural Graph-Based Method

Another time domain method is the structural graph-based method. This method is used to assess the short-term disease risk for patients suffering chronic diseases for producing recommendations in a telehealth environment. The following subsections illustrate the main concepts and equations used in the proposed method.

3.1.4.1 Graph construction and structural graph similarity

Each slide window was mapped as an undirected graph. A graph is a pair of sets $G = (V, E)$, where V is a set of nodes (vertices, or points) so that each node represents the value of a test measurement for that day and E is a set of connections between the nodes of graphs. Therefore, each pair of nodes in a graph is connected by a link if there is a relationship between them (Blondel et al., 2004; Bernhardt et al., 2015; Miraglia et al., 2017). The Euclidean distance has been widely used as a similarity measuring method (Boccaletti et al., 2006; Huang and Lai, 2006; Jain et al., 1999). Let $D_{ij} = 1, 2, 3, ..., M$ be the set of time series of M test measurements in each slide window. Each test measurement in a slide window is assigned to be a node in an undirected graph. Lets n_1 and n_2 be nodes in an undirected graph. They are connected if the distance (d) between them is less than or equal to a determined threshold (Diykh et al., 2017):

$$(n_1, n_2) \in E, if \quad d(n_1, n_2) \leq \theta \tag{3.3}$$

where θ is a predefined threshold. An example of an undirect graph is shown in Figure 3.2. A graph G can be described by giving a square matrix $N \times N$ called an adjacency or connectivity matrix A to describe the connections between the nodes of the graph. The adjacency matrix contains zeros in its diagonal and thus it is a symmetric matrix. The adjacency matrix is equal to one if there is a connection between two nodes, and zero otherwise (Boccaletti et al., 2006).

$$A(n_i, n_j) = \begin{cases} 1 & \text{if}(n_i, n_j) \in E, \\ 0 & \text{otherwise.} \end{cases} \tag{3.4}$$

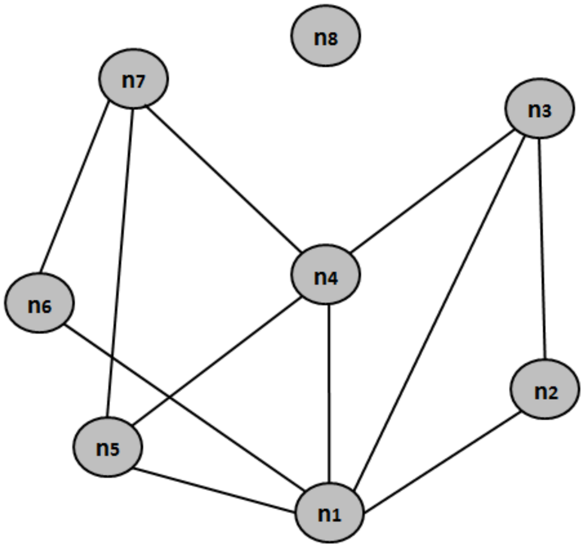


Figure 3.2: An example of an undirect graph

For example, Table 3.1 shows the adjacency matrix of a graph G that consists of 7 nodes. We can see that each element a_{ij} in an adjacency matrix A is equal to 1 when the connection exists, and zero otherwise. The diagonal of matrix A is zero for all its elements.

Table 3.1: The adjacency matrix of a graph G

	n_1	n_2	n_3	n_4	n_5	n_6	n_7
n_1	0	1	1	1	1	1	0
n_2	1	0	1	0	0	0	0
n_3	1	1	0	1	0	0	0
n_4	1	0	1	0	1	0	1
n_5	1	0	0	1	0	0	1
n_6	1	0	0	0	0	0	1
n_7	0	0	0	1	1	1	0

3.1.4.2 Graph features

The adjacency matrix of a graph G can be used to extract the statistical features of a graph G (Li et al., 2013; Diykh et al., 2016; Fang and Wang, 2014). In this study, the statistical features of a graph can be used for prediction. The following sections illustrate the most common extracted features from a graph G:

1. **Degree distributions of the graph**

   The degree distribution, denoted by $P(q)$, refers to the proportion of nodes with degree q divided by the total number of nodes in the graph (Diykh et al., 2016). It can be mathematically defined as follows:

   $$P(q) = \frac{\mid \{z \mid d(z) = q\} \mid}{Z} \tag{3.5}$$

   where $d(z)$ refers to the degree of node z, Z is the total number of nodes in the graph.

2. **The clustering coefficient of the graph**

The Clustering Coefficient(CC) is one of the most important measures used to characterize the local and global structures of a graph (Micheloyannis et al., 2006; Li et al., 2013; Diykh et al., 2016). Let n_i be a node in a graph G. Thus the local clustering coefficient of a given node n_i is computed as the proportion of links between n_i's neighbours which are actually compared with the total number of possible connections. For example, the clustering coefficient of a node n_3 in Figure 3.2 is 1 because the node n_3 has three neighbours, which can have a maximum of 3 connections among them and all of them are realised. The overall level of clustering in a graph is measured as the average of the local clustering coefficients of all the nodes:

$$C' = \frac{1}{N}\sum_{i=1}^{N} C_{ni} \tag{3.6}$$

where, N is the number of nodes in a graph G and C_{ni} is the local clustering coefficient of the node n_i.

3. **Jaccard coefficient of the graph**

Jaccard Coefficient (also called the Jaccard Index) is a statistical tool used to measure the similarity and diversity between two nodes of a graph (Jain et al., 1999). Let n_i and n_j be two nodes in a graph G. Thus, the Jaccard coefficient $\Gamma(n_i, n_j)$ is defined as the ratio of the set of the neighboring intersections between these two nodes to the set of the neighboring unions for the two nodes. It can be mathematically defined as follows:

$$\Gamma(n_i, n_j) = \frac{\mid N(n_i) \cap N(n_j) \mid}{\mid N(n_i) \cup N(n_j) \mid} \tag{3.7}$$

where $N(n_i)$ is the set of neighbors of the node n_i that have an edge from n_i to them, and $N(n_j)$ is the set of neighbors of the node n_j that have an edge from n_j to them.

4. **Average degree**

The average degree (AD) points to the average number of links connecting in a node n_i to the other nodes in the graph (Artameeyanant et al., 2015). The average degree of a graph can be defined as the total number of links for each node divided by the number of nodes in a graph (Diykh and Li, 2016):

$$AD = \frac{1}{N}\sum_{i=1}^{m} K_i \tag{3.8}$$

where K_i is the degree of node n_i and N is the total number of nodes in a graph.

For example, we can easily calculate the degree of each node for a graph G shown in Figure 3.2 and then calculate the average degree (AD) as follows:

$K(n_1)$=5, $K(n_2)$=2, $K(n_3)$=3, $K(n_4)$=4, $K(n_5)$=3, $K(n_6)$=2, $K(n_7)$=3, $K(n_8)$=0, and AD=2.

3.2 Frequency Domains Analysis Techniques for Short-Time Disease Risk Prediction

In this section, two decomposition methods (fast Fourier transformation and Dual-tree complex wavelet transformation) are used to analyze patients' time series data in the frequency domain.

3.2.1 Fast Fourier Transformation

The Fast Fourier Transformation (FFT) was utilized to process patients' time series medical data to facilitate the subsequent data analysis to produce accurate prediction and recommendations. Fast Fourier transformation is an efficient technique for computing the Discrete Fourier Transformation (DFT) and its inverse. It is a robust tool for classification and is faster than the standard DFT.

The input time series data, represented as $X = \{y_1, y_2, y_3, \ldots, y_n\}$ which contains n data, is segmented into a set of overlapped sub-segments based on a predefined value of parameter k that specifies the size of the sliding window, corresponding to each sub-segment. The input time series are analyzed using the fast Fourier transformation to extract frequency information in order to predict the patient's condition. Five frequency bands (i.e., $\alpha, \beta, \gamma, \delta$, and θ) are acquired using the fast Fourier transformation for each sliding window. Since the high frequency band captures most of the information of the sliding window, it is further divided into eight sub-bands. The original sliding window is also added, as a reference, to the extracted feature set. As a result, a total of 14 ($5 + 8 + 1 = 14$) frequency bands are generated for each sliding window. In addition, the power of the FFT coefficients is calculated for all the 14 frequency bands. This allows the computation of the square of the absolute value of the Fourier coefficients. As a result, there can be a total of 28 ($14 + 14 = 28$) frequency bands extracted for a sliding window. Figure 3.3 shows the 28 frequency bands that can be extracted from a sliding window.

From each frequency band, eight different statistical features can be extracted. The extracted features are denoted by X_{Min}, X_{Max}, X_{SD}, X_{Med}, X_{Mean}, X_{RG}, X_{FQ} and X_{SQ}, respectively. Short explanations of the statistical features are provided in Table 3.2. The best performing features are dataset dependent. Some time series data are symmetrically distributed while others may have a more skewed distribution. Generally speaking, the min and max are considered appropriate measures for a time series with a symmetric distribution but, for a skewed distribution, the mean and standard deviation are better used to measure the centre and spread of a dataset (Şen et al., 2014; Li et al., 2011; Diykh and Li, 2016). In this work, the features extracted

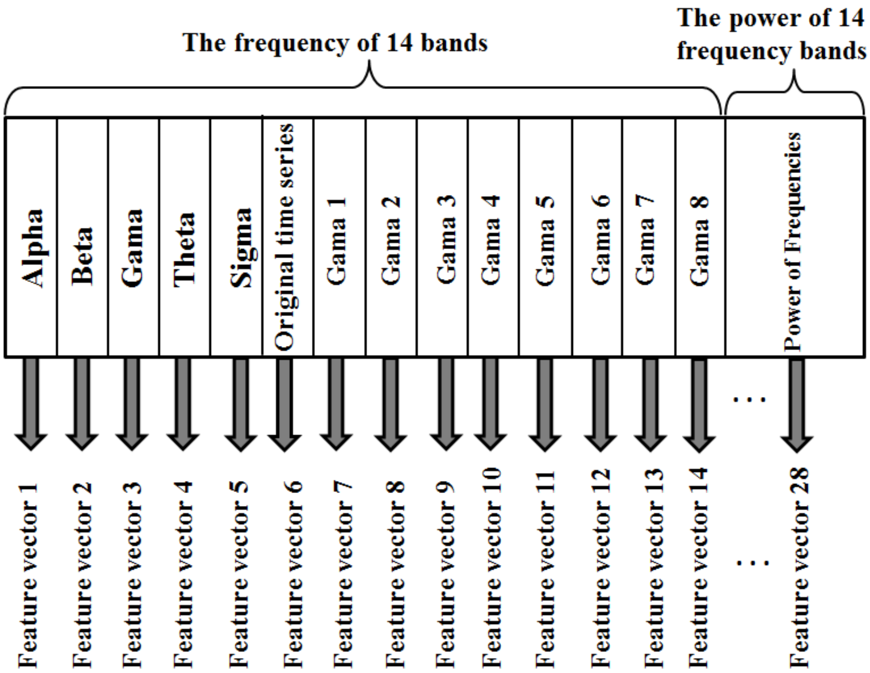


Figure 3.3: The decomposition of a sliding window into 28 bands

Table 3.2: Short explanations of the statistical features

Feature name	Formula	Description
Maximum value	$X_{Max} = Max[x_n]$	Where $x_n = 1, 2, 3, ..., n$ is a time series, n is the size of the sliding window and AM is the mean of the sliding window.
Minimum value	$X_{Min} = Min[x_n]$	
Mean	$X_{Mean} = \frac{1}{n}\sum_{1}^{n} x_i$	
Median	$X_{Med} = (\frac{N+1}{2})^{th}$	
Standard Deviation	$X_{SD} = \sqrt{\sum_{n=1}^{N}(x_n - AM)}\frac{2}{n-1}$	
Range	$X_{RG} = X_{Max} - X_{Min}$	
First Quartile	$X_{FQ} = \frac{1}{4(N+1)}$	
Second Quartile	$X_{SQ} = \frac{4}{4(N+1)}$	

from each band are grouped into one vector and used as the input to the ensemble learning model predicting the patient's condition. Experimental evaluations are conducted to evaluate the impact of different combinations of the extracted features on the final recommendation generated. The detailed results are discussed in the experimental results section.

3.2.2 Dual-Tree Complex Wavelet Transformation

The Discrete Wavelet Transformation (DWT) is a powerful tool for analyzing signals in time-frequency domain. The drawbacks of DWT are ameliorated by using the Dual Tree Complex Wavelet Transformation (DTCWT) which offers a better time-frequency representation of signals.

Here, the input time series data is segmented into a set of overlapping sub-segments based on a pre-defined value of parameter k that specifies the size of the sliding window corresponding to each sub-segment. Then, each of the sub-segments is subjected to a four-level DTCWT decomposition, giving 10 wavelet coefficients sub-bands. The input time series are decomposed using the dual tree complex wavelet transformation to extract frequency information to predict the patient's condition. DTCWT decomposition techniques represent the input time series as a few sets of wavelet coefficients (sub-bands).

The ability to convert the set of obtained wavelet coefficients into a reduced set of features is a very important stage in the proposed system. Since these features characterize the behavior of the time series data of patients, the process of feature selection is of crucial importance. The following six statistical features were used to predict the patient's condition:

1. Mean of coefficients absolute values in every sub-band, $\eta = \frac{1}{N}\sum_{j=1}^{N} \mid y_j \mid$
2. Average power of the coefficients in every sub-band, $\mu = \sqrt{\frac{1}{N}\sum_{j=1}^{N} y_j^2}$
3. Standard deviation of the coefficients in every sub-band, $\nu = \sqrt{\frac{1}{N}\sum_{j=1}^{N}(y_j - \eta)^2}$
4. Ratio of the absolute mean values of coefficients of adjacent sub-bands, $\xi = \frac{\sum_{j=1}^{N} \mid y_j \mid}{\sum_{j=1}^{N} \mid z_j \mid}$
5. Kurtosis of the coefficients in every sub-band, $\chi = \sqrt{\frac{1}{N}\sum_{j=1}^{N} \frac{(y_j - \eta)^4}{\nu^4}}$
6. Skewness of the coefficients in every sub-band, $\phi = \sqrt{\frac{1}{N}\sum_{j=1}^{N} \frac{(y_j - \eta)^3}{\nu^3}}$

where y and z are two adjacent sub-bands and N is the length of a sub-band. In the literature, Features 5 and 6 are considered as a higher order than other statistical features.

The obtained statistical features will be entered to the machine learning ensemble to predict the patient's condition. The detailed results are discussed in the experimental results section. The following section illustrates the machine learning ensemble in more detail.

3.3 Machine Learning Ensemble Model

As the theoretical background, a detailed description of the constituent classifiers used in the ensemble model is provided in this section. Specifically, the ensemble model consists of three popular and capable machine learning classifiers: Artificial Neural Networks (ANN), Least Square-Support Vector Machine (LS-SVM) and Naive Bayes (NB). Based on the literature, these classifiers were found to be the most popular in data mining and classification (Sharawardi et al., 2014; Diykh and Li, 2016; Geng et al., 2015).

The statistical features of the structural graph-based method and frequency domains analysis techniques (FFT and DTCWT) are individually entered to the ensemble model as shown in Figure 3.1. The final recommendation is considered based on the aggregation of the three classifiers. The detailed information for each method is presented in the next chapter (Experimental results).

3.3.1 Artificial Neural Networks

Artificial Neural Network (ANN) is a supervised learning machine algorithm that can be used to provide effective solutions for many complex modeling problems. This method uses a set of processing neurons or nodes which are interconnected. The network can be considered as a directed graph in which each neuron i executes the transfer function f_i as in the following equation (Yao, 1999):

$$y_i = f_i\Big(\sum_{j=1}^{n} w_{ij} x_j - \theta i\Big) \tag{3.9}$$

where y_i is the output of neuron i, x_j is the jth input to the neuron and w_{ij} is the connection weight between neurons i and j. θ is the threshold bias of the neuron. f_i is usually nonlinear taking, for example, a sigmoid, Heaviside, or Gaussian function.

3.3.2 Least Square-Support Vector Machine

The Least Square-Support Vector Machine (LS-SVM) (Suykens and Vandewalle, 1999) is a supervised machine learning technique that is based on statistical learning theory. It has been used in some applications for medical prediction, such as heart disease prediction (Bai et al., 2015), muscle fatigue prediction in electromyogram (sEMG) signals (Suykens and Vandewalle, 1999) and breast cancer prediction (Sharawardi et al., 2014).

A linear LS-SVM, which is the most popular form of LS-SVM, is designed to classify a dataset that contains two separable classes represented as $\{1, -1\}$ (Polat and Güneş, 2007). This method attempts to map the given data into a high-dimensional space and then uses a hyperplane separating the two classes involved by maximizing the distance between the plane and the support vectors. Suppose the training data consists of n data $(x_1, y_1), (x_2, y_2), ..., (x_m, y_m) \in R^m$. LS-SVM finds the optimal separating hyperplane (with the maximum margin) to separate these classes. LS-SVM is based on the following rules for solving a given problem:

$$y_i[(wx_i) + w_0] = 1 - \xi_i, i = 1, ..., m, \tag{3.10}$$

$$1/2\|w\|^2 + \frac{c}{2}\sum_{i=1}^{m}\xi_i^2 \tag{3.11}$$

According to the above formulas, the problem to be solved by LS-SVM can be formulated as follows:

$$(w, b, \alpha, \xi) = 1/2\|w\|^2 + \frac{c}{2}\sum_{i=1}^{m}\xi_i^2 - \sum_{i=1}^{m}\alpha_i\{y_i[(wx_i) + w_0] - 1 + \xi_i\} \tag{3.12}$$

3.3.3 Naive Bayes

The Naive Bayes classifier (NB) is a machine learning classifier that is based on probability theory with the supposition that the features are independent of each other (Geng et al., 2015). It is often used for classification problems in machine learning as well as for medical predictions. According to the Bayes theorem, the conditional probabilities involved in the problem can be obtained using the following formula:

$$P(C_x \mid Y) = P(C_x) \times \frac{P(Y \mid C_x)}{P(Y)} \tag{3.13}$$

where Y is a given example that needs to be classified. C_x is a class label and $P(C_x \mid Y)$ is the probability that the vector Y belongs to C_x.

3.3.4 Bootstrap Aggregation (Bagging)

An ensemble approach is a very effective method that combines the decisions of multiple base classifiers to overcome the limited generalization performance of each base classifier, and generate more accurate predictions than individual base classifiers. Bootstrap aggregation is a machine learning ensemble algorithm designed to enhance the accuracy and stability of machine learning algorithms (Li et al., 2016; Gao et al., 2016). In the bootstrap method, the classifiers are trained independently and then aggregated by an appropriate combination strategy. Specifically, this ensemble model can be divided into two phases. In the first phase, the model uses bootstrap sampling to generate a number of training sets. In the second phase, the training of the three base classifiers, Neural Network, Least Square-Support Vector Machine and Naive Bayes, is performed using the bootstrap training sets generated during the first phase. Figure 3.4 shows an example of the bagging algorithm which involves the three classifiers to build our ensemble model.

In this study, the training set was divided into multiple datasets using the bootstrap aggregation approach, and then the classifiers were individually applied to these datasets to generate the final prediction. It is noted that different individual classifiers in the bagging approach may perform differently. Therefore, this study assigns a weight to each classifier's vote based on how well the classifier performs. The classifier's weight is calculated based on its error rate. The classifier that has the lower error rate is considered more accurate and is, therefore, assigned a higher weight. The weight of classifier C_i's vote is calculated as follows:

$$w(C_i) = \log \frac{1 - error(C_i)}{error(C_i)}, 1 \leq i \leq 3 \tag{3.14}$$

The following example is presented to facilitate the understanding of the proposed weighted bagging ensemble model:

1. Neural Network, Least Square-Support Vector Machine and Naive Bayes are used as individual base classifiers in the ensemble model. Suppose that the classifier training is performed on the training data and the error rate is calculated for each base classifier as 0.25 for NN, 0.14 for LS-SVM, and 0.30 for NB;

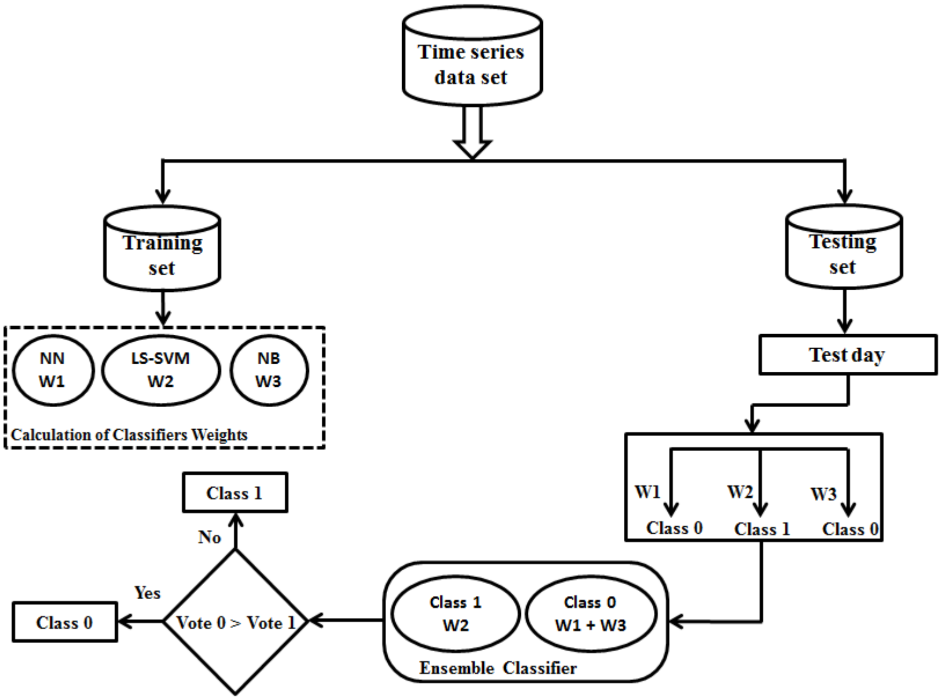


Figure 3.4: An example of a bagging algorithm

2. As per Equation (3.14), the weight 0.47 is assigned to NN, 0.78 to LS-SVM, and 0.36 to NB;

3. Suppose that the three base classifiers generate the following predictions for a coming testing day: NN predicts 0, LS-SVM predicts 1 and NB predicts 0 (Here, 0 means no test is required on the testing day for a medical measurement; 1 means a test is required otherwise);

4. The ensemble classifier will use the weighted vote to generate the following prediction results:

   Class 0: NN + NB $\longrightarrow 0.47 + 0.36 \longrightarrow 0.83$,

   Class 1: LS-SVM $\longrightarrow 0.78$.

5. Finally, according to the weighted vote, the Class 0 has a higher value than Class 1. Therefore, the ensemble classifier will classify this testing day as being

in Class 0, suggesting that the patient in question can skip the test on that day for a medical measurement.

3.4 Summary

This chapter introduced a set of disease risk prediction algorithms and models proposed for the system for short-term disease risk prediction. First, three innovative predictive algorithms (basic heuristic algorithm, regression-based algorithm and hybrid algorithm) as well as a structural graph-based method were described. The proposed algorithms were applied to study the behavior of time series data in the time domain for short-term risk assessment of patients in a telehealth environment based on the analysis of a patient's historical medical data. Then, the fast Fourier transformation and dual-tree complex wavelet transformation were presented to study the behavior of the time series data of chronic disease patients in the frequency domain.

The experimental results of the innovative predictive algorithms and existing techniques will be presented in the next chapter.

Chapter 4

Experimental Evaluation

This chapter discusses the experimental evaluation design including performance metrics and the experimental platform of the algorithms and models for short-term disease risk prediction using the time and frequency analysis techniques.

The main purpose of this research is to investigate the effectiveness of the proposed recommendation system in analysing the time series data of patients in both the time domain and the time-frequency domain to provide patients in a telehealth environment with appropriate recommendations regarding the need to undertake a medical body test on the coming day. The conceptual framework of the study is presented in Figure 3.1.

Three innovative predictive algorithms, Basic Heuristic Algorithm (BHA), Regression-Based Algorithm (RBA) and Hybrid Algorithm (HA), and a structural graph-based method (SG) are proposed to study the behaviour of time series data in the time domain for short-term risk assessment of patients in a telehealth environment based on the analysis of patients' historical medical data. In the frequency domain, the fast Fourier transformation and dual-tree complex wavelet transformation are applied to study the behavior of the time series data of patients in frequency domain.

4.1 Time Series Datasets

Two time series telehealth datasets collected from patients are used for experimentation and evaluation. Extensive experiments were designed and conducted to evaluate the performance of the proposed recommendation system using the two real-life datasets. The first dataset, called the Tunstall dataset, was obtained from a group of chronic heart disease patients while the second dataset was obtained from a set of chronic diabetes patients.

4.1.1 Data Acquisition

4.1.1.1 Tunstall datset

The study used a real-life dataset obtained from our industry collaborator, Tunstall Healthcare, to test the practical applicability of the proposed system. This Tunstall dataset was obtained from a pilot study conducted on a group of chronic heart disease patients, and the data collected contained the patients' day-to-day medical readings of different medical measurements in a telehealth care environment. The dataset is a time series and comprises data from six patients with a total of 7,147 different time series records acquired between May and October 2012. Each dataset's record consists of a number of patient-related meta-data attributes, such as patient-id, visit-id, measurement type, measurement unit, measurement value, measurement question, date and date-received.

The characteristics of the meta-data attributes of the dataset are shown in Table 4.1. The dataset also contains the numerical readings of several critical medical measurements for each patient on every day during the time period of the pilot study, including Heart Rate, Diastolic Blood Pressure (DBP), Mean Arterial Pressure (MAP), Oxygen Saturation (SO2), Blood Glucose and Weight, of which the data of Heart

Table 4.1: Meta-data attributes of Tunstall dataset

Attribute name	Attribute type
id	Numeric
patient-id	Numeric
hcn	Numeric
visit-id	Numeric
measurement type	Nominal
measurement unit	Nominal
measurement value	Numeric
measurement question	Nominal
date	Numeric
date-received	Numeric

Rate, DBP, MAP and SO2 are used in the evaluation.

4.1.1.2 Diabetes datset

The Diabetes dataset used in this research was obtained from the Repository of Machine Learning Databases which is available online from the URL link of `https://archive.ics.uci.edu/ml/datasets/diabetes` (Kahn, 1994). The data contains measurements taken multiple times per day from 70 patients. Blood glucose measurements, symptoms and insulin treatments were recorded with timestamps for each patient over the course of several weeks to months. Two different sources, an automatic electronic recording device and paper records, were used to acquire the data records. The automatic device had an internal clock to timestamp events, whereas the paper records only provided “logical time” slots (breakfast, lunch, dinner, bedtime) (Moghaddass et al., 2016).

The dataset is a time series and comprises data from 70 Diabetes patients. Each

Table 4.2: Meta-data attributes of Diabetes dataset

Attribute name	Attribute type
Date	Numeric
Time	Numeric
Code	Numeric
Value	Numeric

record in the Diabetes dataset consists of four fields for the date of measurement, time of measurement, code of measurement and value of measurement. The characteristics of the meta-data attributes of the dataset are shown in Table 4.2. Each code represents one of twenty numerical readings about the patients such as regular insulin dose, NPH insulin dose, ultralente insulin dose, unspecified blood glucose measurement, pre-breakfast blood glucose measurement, post-breakfast blood glucose measurement, pre-lunch blood glucose measurement, post-lunch blood glucose measurement, pre-supper blood glucose measurement, post-supper blood glucose measurement, pre-snack blood glucose measurement, hypoglycemic symptoms, typical meal ingestion, more-than-usual meal ingestion, less-than-usual meal ingestion, typical exercise activity, more-than-usual exercise activity, less-than-usual exercise activity and unspecified special event.

4.1.2 Ethics Clearance

Prior to the use of the Tunstall dataset, ethics clearance was obtained from the Human Research Ethics Committee of the University of Southern Queensland (No: H15REA167).

4.1.3 Data Pre-processing

The real time series datasets were transferred to a desktop computer configured with a 3.40 GHz Intel core i7 CPU processor with 8.00 GB RAM for off-line analysis. Data pre-processing is necessary to ensure that raw data is ready for the analysis tasks (Famili et al., 1997). Analytic tools can be misled and give inaccurate results if data have impurities such as missing or duplicate data. As a result, it is necessary to prepare the data before starting the data analysis process. In this phase, the data missing problem, caused during data collection or transmission, is resolved by filling the missing data with a global constant. Also, duplicated records are removed from the dataset to make the data consistent. Another important task of data pre-processing, in this work, is to extract the information for each individual patient from the original dataset for personalized data analysis and recommendations.

In this study, the input time series data, represented as $X = \{y_1, y_2, y_3, \ldots, y_n\}$ which contains n data, is segmented into a set of overlapping sub-segments based on a predefined value of parameter k that specifies the size of the sliding window, corresponding to each sub-segment.

Before the proposed algorithms and techniques were performed, the input variables were scaled using a normalization technique. The normalized variables take on values within the interval of [0,1] (Singh and Gupta, 2007). Normalization is performed based on the following equation:

$$D_{nor} = \frac{D_{ori} - D_{min}}{D_{max} - D_{min}} \tag{4.1}$$

Here, D_{nor} is the normalized data value, D_{ori} is the original raw data, D_{min} and D_{max} are the minimum and maximum data values respectively for the entire dataset.

4.1.4 Summary

This chapter presents the time series data of patients in both the Tunstall and Diabetes datasets. The raw data of Tunstall and diabetes were collected and preprocessed using C/C++ in Microsoft Visual Studio 2008 on a desktop computer with the configurations of a 3.40 GHz Intel Core i7 CPU processor with 8.00 GB RAM. The medical checkups for each patient were selected from the datasets and saved in a separate file for later processing. In addition, the medical checkups were normalised to reduce the exaction time of the proposed algorithms and techniques.

4.2 The Experimental Setup

The performance of the proposed recommendation system has been evaluated by three performance metrics proposed for this work, namely *accuracy*, *workload saving* and *risk*. Accuracy refers to the percentage of correctly recommended days against the total number of days for which recommendations are provided. Workload saving refers to the percentage of the total number of days when recommendations are provided for skipping the medical test against the total number of days in the training set. Risk refers to the percentage of incorrectly recommended days against the total number of days in the training set. Mathematically, they are defined as follows:

$$Accuracy = \frac{NN}{NN + NA} \times 100\% \tag{4.2}$$

$$Saving = \frac{NN + NA}{\mid \mathcal{D} \mid} \times 100\% \tag{4.3}$$

$$Risk = \frac{NR}{\mid \mathcal{D} \mid} \times 100\% \tag{4.4}$$

Here, NN denotes the number of days with correct recommendations, NA denotes the number of days with incorrect recommendations, NR denotes the number of days with a risky recommendation that refers to a recommendation that suggests skipping a medical test for a given medical measurement but the actual reading of the measurement in the testing set is abnormal), and $\mid \mathcal{D} \mid$ refers to the total number of days in the training set. Here, a correct recommendation means that the system produces the recommendation of "no test required" for the following day and the actual reading for that day in the dataset is normal. Any recommendations other than this case are considered incorrect.

Two metrics, *Difference* and *Rotating frequency*, have also been devised in this study. These have been developed to quantitatively measure data characteristics for each medical measurement. Their impact on the performance of the recommender system in terms of accuracy, workload saving and risk will be evaluated. Difference refers to the difference in terms of the number of normal and abnormal values for a given medical measurement in the dataset, as dictated by their corresponding minimum and maximum threshold values; whilst rotating frequency refers to the total number of instances where two consecutive values for the medical measurement have different states (one is normal and the other one is abnormal) in the dataset. Both metrics will be normalized against the total number of days in the dataset. Mathematically, difference and rotating frequency for a medical measurement m are defined as follows:

$$Difference(m) = \frac{\mid NV_m - AV_m \mid}{\mid \mathcal{D} \mid} \times 100\% \tag{4.5}$$

Table 4.3: Examples of rotating frequency

Example1	Exampe2
normal	normal
normal	abnormal
normal	normal
normal	normal
abnormal	abnormal
abnormal	abnormal
abnormal	abnormal
abnormal	abnormal
abnormal	abnormal
abnormal	normal
Difference=2/10=0.2	Difference=2/10=0.2
Rotating Frequency=1/10 =0.1	Rotating Frequency=4/10 =0.4

$$RotatingFrequency(m) = \frac{Rotation}{\mid \mathcal{D} \mid} \times 100\% \tag{4.6}$$

Where NV_m and AV_m denote the number of normal and abnormal values of m in the dataset. Normal and abnormal values are detected based on the corresponding minimum and maximum threshold values of m. *Rotation* denotes the total number of instances where two consecutive values for the medical measurement have different states. $\mid \mathcal{D} \mid$ refers to the total number of days in the dataset. Table 4.3 demonstrates two examples for difference and rotating frequency.

The proposed recommendation system was developed using MATLAB on a desktop computer configured with a 3.40 GHz Intel core i7 CPU processor with 8.00 GB RAM.

4.3 The Experimental System

Easy-to-use interfaces are developed for users to supply necessary inputs to the system and receive recommendations from the system. An intelligent recommendation system for patients with chronic diseases (IRS-CD) involves human computer interaction to receive input from human users concerning the values of the parameters that are used in the system's algorithm. Recommendations generated by the IRS-CD will be returned to users through different channels and platforms including desktops, laptops and even tablets, thus embracing the latest technological advancements for quick information dissemination. As well as being returned to patients, the results can be sent remotely to medical practitioners such as doctors and nurses so that they can keep track of the physical checkups and overall health conditions of their patients.

4.3.1 The System Architecture of IRS-CD

The major contribution of this chapter is the design of an end-user software. The software is designed based on the heterogenous prediction algorithms discussed in the previous chapters. The proposed software is a time series analysis software designed for easy use by end users in a telehealth environment. It is enriched with user friendly interfaces and interactive defaults. The software is designed to supply necessary inputs to the system and receive recommendations from the system. The work of the intelligent recommendation system can be divided into two categories:

1. Data input and preparation
2. Result analysis and recommendations

As discussed in previous sections, the time series data of patients are pre-processed to make them ready for analysis tasks. The results analysis and recommendations provision tasks are performed using the time series analysis software designed for easy use by end users in a telehealth environment. The use scenarios of the IRS-CD is described in this chapter. It contains descriptions of the software interfaces with respect to analyzing the time series data of patients and providing the appropriate recommendations to patients in relation to the necessity of medical tests taken on a daily basis.

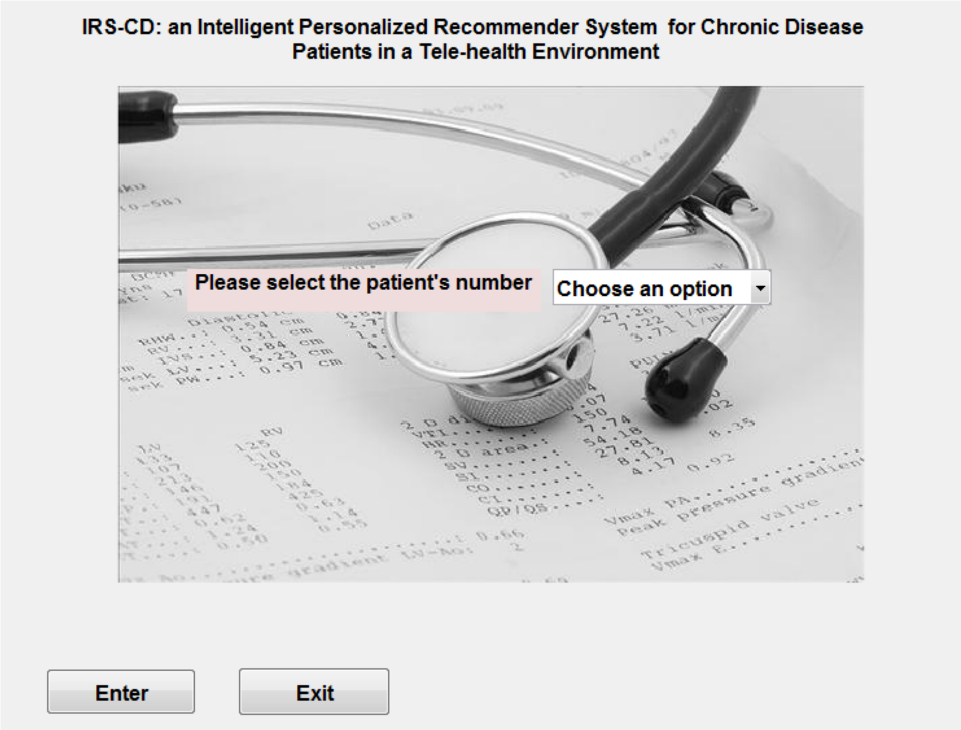
4.3.2 Managing IRS-CD System

This section presents the intelligent recommendation system equipped with an array of predictive algorithms which analyse the medical data, assess risk and provide appropriate recommendations regarding the necessity of taking a medical test in the following day based on the outcome of the prediction (Lafta, Zhang, Tao, Li and Tseng, 2016).

The IRS-CD utilizes heterogenous prediction algorithms, the basic heuristic algorithm, regression-based algorithm, hybrid algorithm, as well as a structural graph-based method to study the behavior of time series data in the time domain. In addition, the fast Fourier transformation and dual-tree complex wavelet transformation are employed to study the behavior of the time series data of chronic disease patients in the frequency domain.

All phases of the IRS-CD were implemented using Matlab. Figure 4.1 shows the first interface of the system. Here, the user of system must select the patient's number from the list and then the system moves to the main interface.

In the main interface, the user must enter the three parameters into the system and

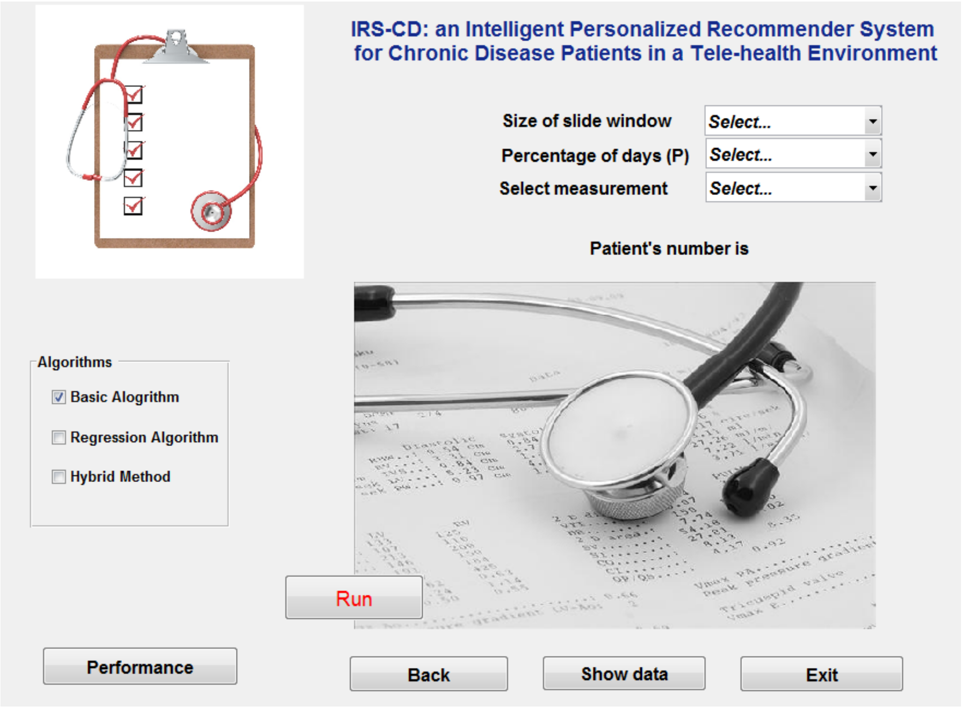


IRS-CD: an Intelligent Personalized Recommender System for Chronic Disease Patients in a Tele-health Environment

Figure 4.1: First interface to select the patient's number and enter into the system [1]

---

[1] The picture included in Figure 4.1 is downloaded from (*burnham health centre*, 2006)



IRS-CD: an Intelligent Personalized Recommender System for Chronic Disease Patients in a Tele-health Environment

Size of slide window: Select...
Percentage of days (P): Select...
Select measurement: Select...

Patient's number is

Algorithms
☑ Basic Alogrithm
☐ Regression Algorithm
☐ Hybrid Method

Figure 4.2: Main interface presenting the recommendations

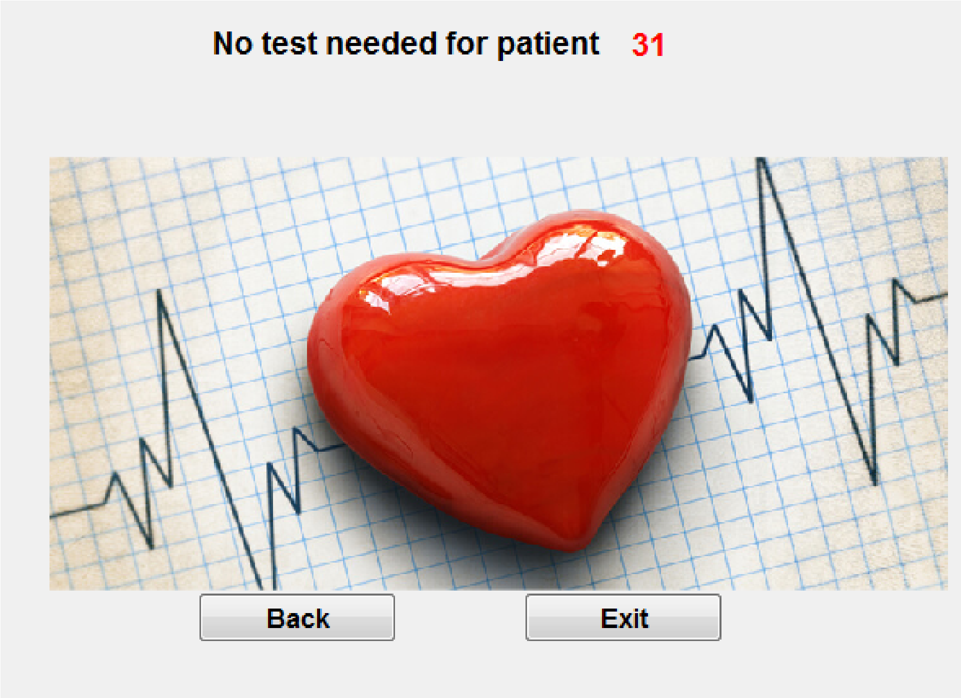


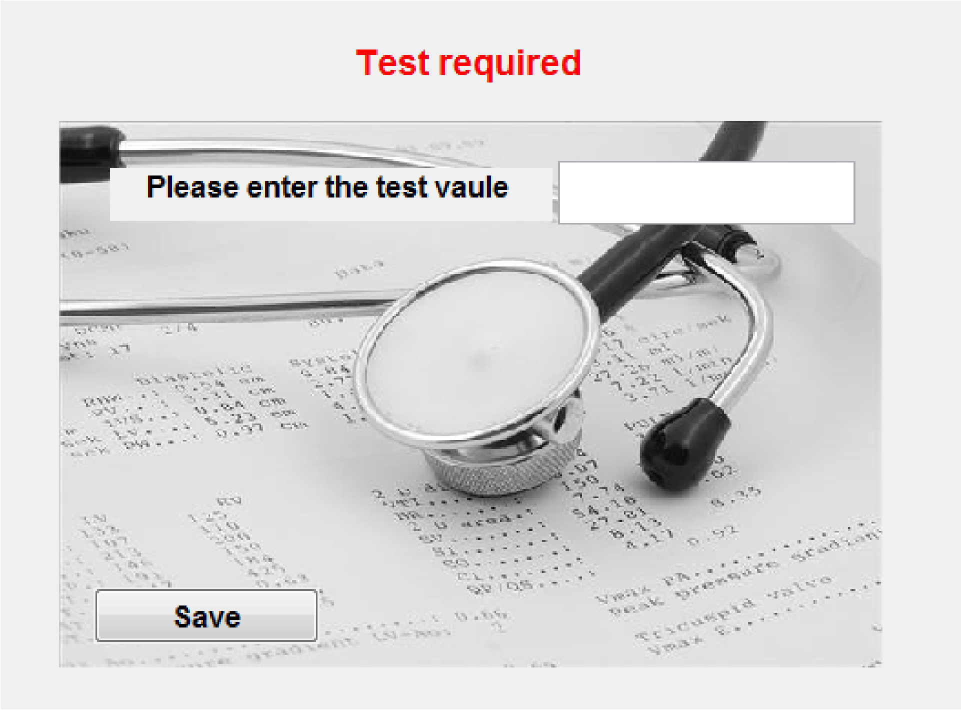
Figure 4.3: Generation of the recommendation “No test required”[2]

then select the predictive analytic methods from the list as shown in Figure 4.2. There are three parameters in the recommendation system requiring configuration before use. The first parameter is the medical measurement (e.g., heart rate, blood pressure, etc.) that the prediction is working on. The minimum (min) and maximum (max) normal values for the selected medical measurement are also specified to establish its normal range. The second parameter is the length of the sliding time window k which determines the historical data to be utilized for prediction and recommendation. The final parameter is p for determining the minimum percentage of days when the physical test is conducted by the patient for the measurement in the past k days.

Users can choose one or multiple analytical methods and, if multiple methods are selected, the IRS-CD utilizes an ensemble-based method to produce the prediction and recommendation. Through this function, the IRS can assess the patient’s status

---

[2] The picture included in Figure 4.3 is downloaded from (Sepalika, 2018)



Test required

Please enter the test vaule

Figure 4.4: Generation of the recommendation "Test required"

and then generate recommendations based on the analysis of his/her measurement readings for the past k days and decide whether a given patient needs to take a medical measurement such as the heart rate test today or not.

Here, if the patient does not need to take the test for a selected medical measurement on the following day, a recommendation of "no test required" will be generated and presented on the interface, and stored in the backend database as a part of the patient's historical records, as shown in Figure 4.3. Otherwise, a recommendation of "test required" will be generated and the patient is advised to take the medical test on the following day. Then, the recommendation system will ask the user to enter the his/her test value for today which will be stored in the system as a historical record, as shown in Figure 4.4.

4.4 The Analysis Results of Time Domain Algorithms

The three innovative predictive algorithms for short term risk assessment are proposed in this section. The algorithms – basic heuristic algorithm, regression-based algorithm and hybrid algorithm – and a structural graph-based method are used to analyze the time series data of patients in the time domain. These algorithms and a structural graph-based method are developed to provide patients in a telehealth environment with appropriate recommendations for the need to take a medical body test on the coming day. Such recommendations are established based on the prediction of the patient's conditions using their time series medical data from the past few days.

4.4.1 The Analysis Results of the Three Predictive Algorithms

In this section, the results of the experimental evaluation conducted on the proposed system are reported. The effect of two key parameters used in the algorithms, k and p were evaluated. k determines the size of the sliding windows (i.e., the number of past days involved in assessing the disease risk of the current day) used for analyzing the time series data, while p is the percentage of the days in the past k days that the patients have physically conducted the test for the measurement under study. After evaluating the effectiveness of the parameters, an investigation of the performance of the proposed system from the perspective of different measurements and patients was conducted. Then, the impact of the difference and rotating frequency on different algorithms was evaluated. After this, a comparative study on the three algorithms was undertaken to evaluate the performance of the proposed system. Finally, a correlation coefficient analysis was undertaken to assess the relationships between the performance metrics (*accuracy* and *workload saving*) and dataset characteristics (*differences* and *rotating frequency*). Both accuracy and workload savings were evaluated in all the experiments.

4.4.1.1 Effectiveness of k

First, the effectiveness of k, the size of the sliding window used in time series data analysis. Figures 4.5 and 4.6 show the averaged accuracy and saving percentages with time windows in different sizes for all algorithms. Generally speaking, for both the Tunstall and Diabetes datasets collected in the pilot studies, the best combination of accuracy and working savings was achieved when k ranged from 3 to 5, the relatively small values, meaning the risk prediction was more accurate when only the most recent days were considered.

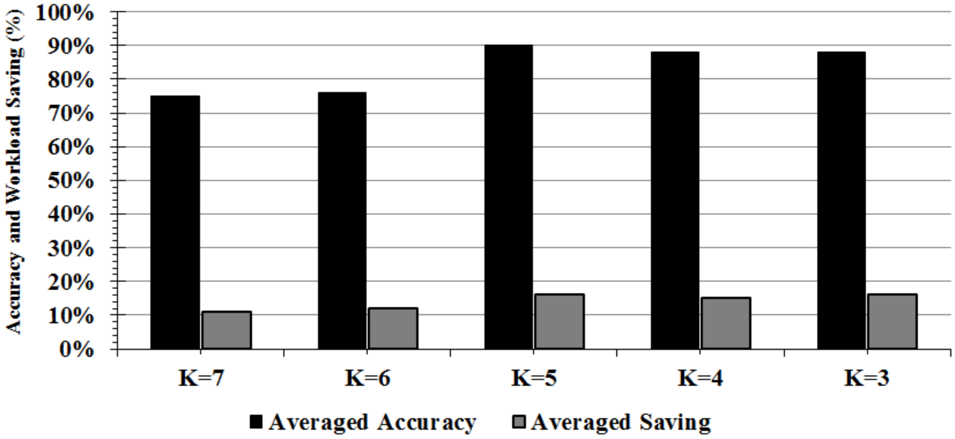


Figure 4.5: The averaged accuracy and workload savings for all algorithms under varying values of k using the Tunstall dataset

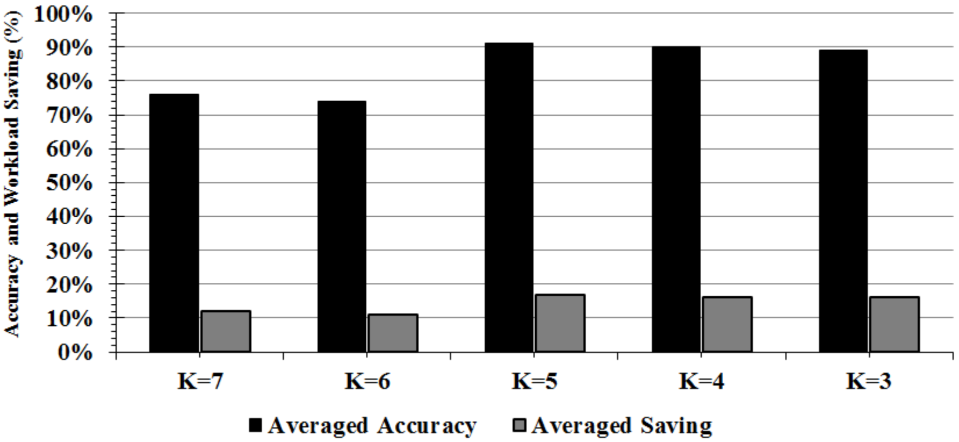


Figure 4.6: The averaged accuracy and workload savings for all algorithms under varying values of k using a Diabetes dataset

4.4.1.2 Effectiveness of p

In this experiment, the proposed system was applied to different values of p on all patients when $k = 5$. The results are presented in Figures 4.7 and 4.8. When p takes relatively large values, e.g. when p ranged from 70% to 80%, the system achieves better recommendation accuracy. The reason is that, under a fixed value of k, a larger value of p led to a larger lower-bound for the percentage of the days in the sliding window when a medical test was required. A higher percentage of the days with medical test naturally provided more information to the system and, as a result, made the system capable of providing more accurate recommendations. In addition, on the basis of observation of the experimental results, a higher value of p did not considerably affect the performance of workload saving despite better accuracy being achieved.

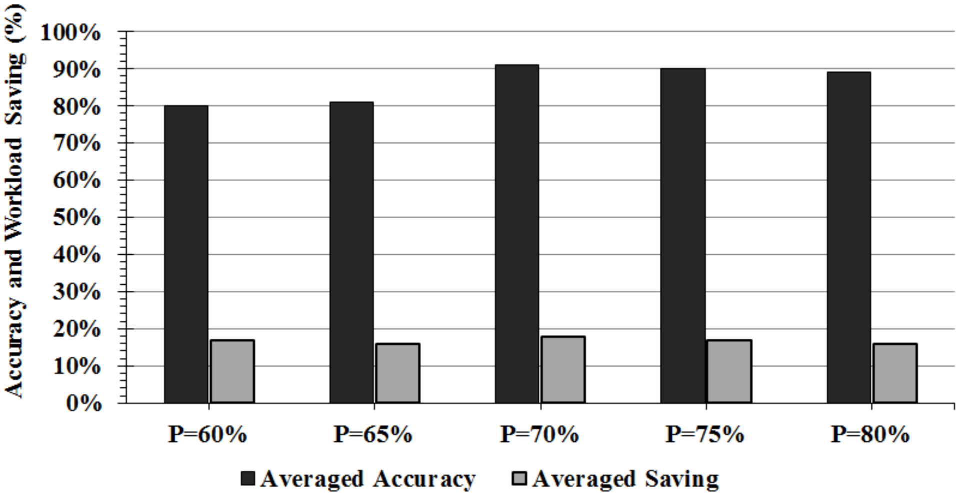


Figure 4.7: The averaged accuracy and workload savings for all algorithms under varying values of p using Tunstall dataset

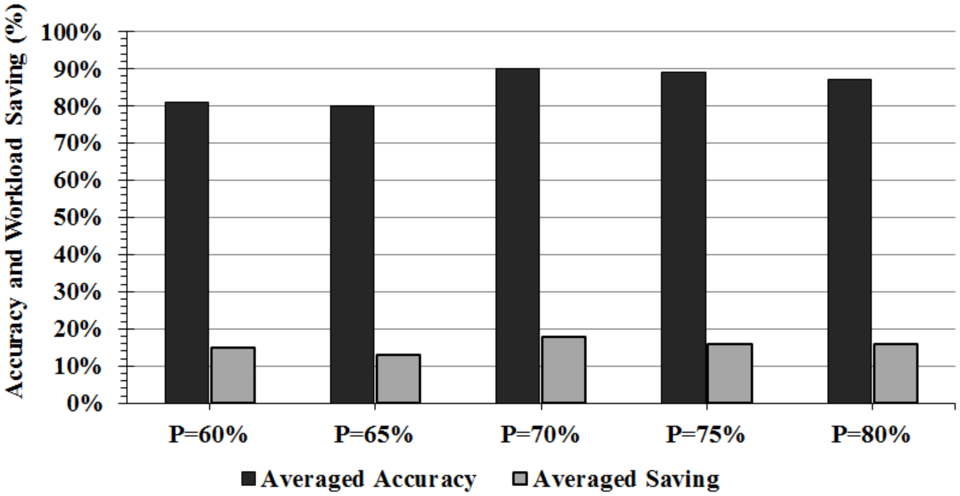


Figure 4.8: The averaged accuracy and workload savings for all algorithms under varying values of p using Diabetes dataset

4.4.1.3 Performance of the three algorithms under different measurements for all patients

In this experiment, the performance of the system is evaluated when it is applied to different measurements for all patients based on $k = 5$ for the Tunstuall dataset. Each algorithm was tested in four rounds for all patients with different medical tests (Heart rate, DBP, MAP and SO2). Figure 4.9 shows the detailed results obtained for each algorithm. From the results, we can see that the algorithms yield recommendations with varying degree of accuracy, with the heart rate and SO2 measurements registering the highest compared to others. This is because there are intrinsically stronger correlations for the neighboring readings of heart rate and SO2 measurements so the short-term prediction of risk becomes more accurate than the other measurements.

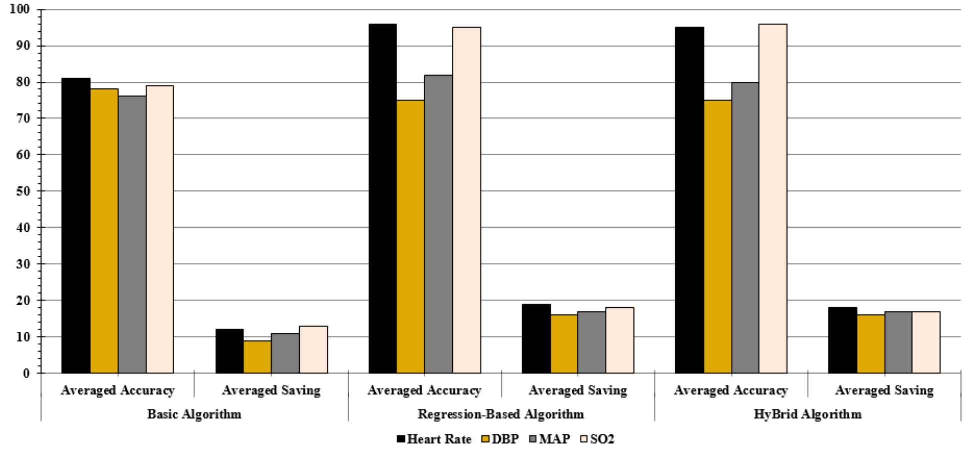


Figure 4.9: The accuracy and workload saving for different patients in Tunstall dataset

By further aggregating the results for the measurements in Figure 4.9, Figure 4.11 demonstrates the averaged accuracy and workload savings for each algorithm. The results clearly show that Regression-Based and Hybrid algorithms yield the highest accuracy and saving compared with the Basic algorithm. Therefore, the risk prediction of disease provided by the proposed system is more accurate with these two

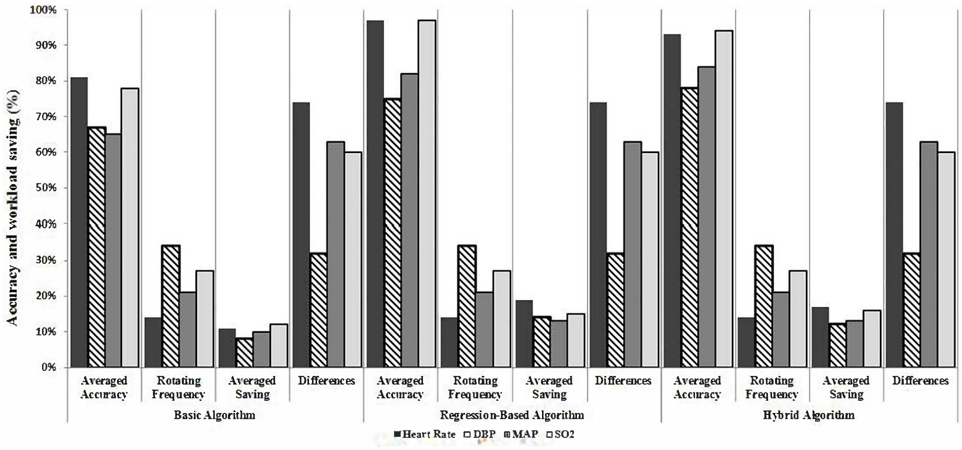


Figure 4.10: Rotating frequency analysis for each algorithm under four measurements in Tunstall dataset

algorithms. The system is capable of helping reduce, on average, 15% of workload for patients from their daily medical tests.

4.4.1.4 Difference and rotating frequency analysis

The relationship between the data characteristics measured by Difference and Rotating Frequency on the performance of the proposed system was also evaluated. Figure 4.10 shows the performance of the system and the data characteristic measurements (Difference and Rotating Frequency) of four measurements. As shown in Figure 4.10, there is a negative relationship between Accuracy and Rotating Frequency under each measurement for all algorithms. Based on the obtained results, "it is observed that the recommendations provided by the proposed system are more accurate for a measurement when Rotating Frequency is degraded for this measurement and vice visa. In addition, the value of Difference is positively correlated with that of accuracy, suggesting that the proposed system can produce more accurate recommendations when Difference values are high.

4.4.1.5 Algorithms' performance evaluation for disease risk prediction

In this experiment, three algorithms were tested to evaluate their performance for all patients in both the Tunstall and Diabetes datasets. From Figures 4.11 and 4.12, we can clearly see that there are relative differences in the values of accuracy and saving for each algorithm. General speaking, the Regression-Based and Hybrid algorithms yielded a high degree of workload saving compared with the Basic algorithm. This means that these two algorithms can help reduce patient workload by more than 15% from their daily medical tests. Recommendation accuracy of the proposed system ranged between 75% and 92% across different patients under the three algorithms.

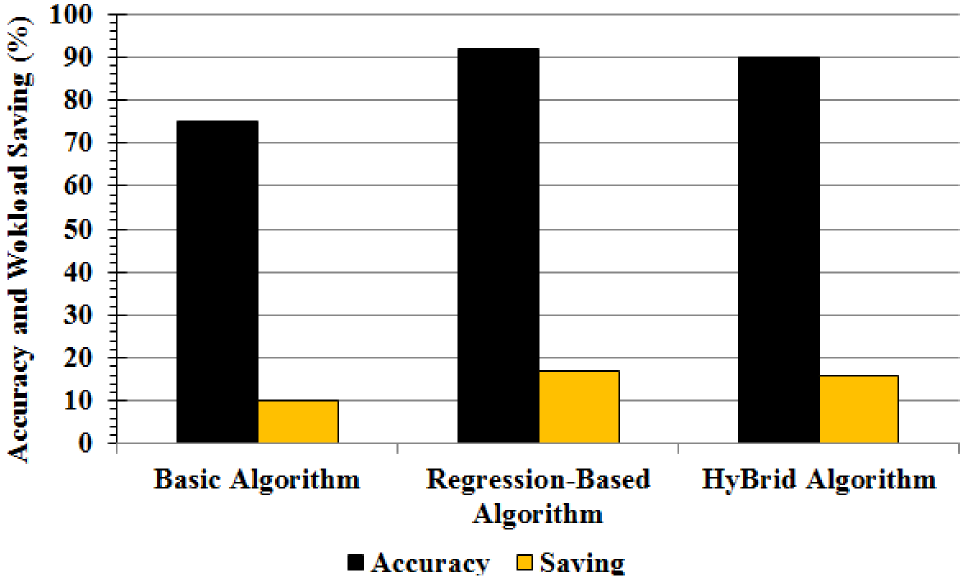


Figure 4.11: Comparison the performance of the three algorithms for all patients in Tunstall dataset

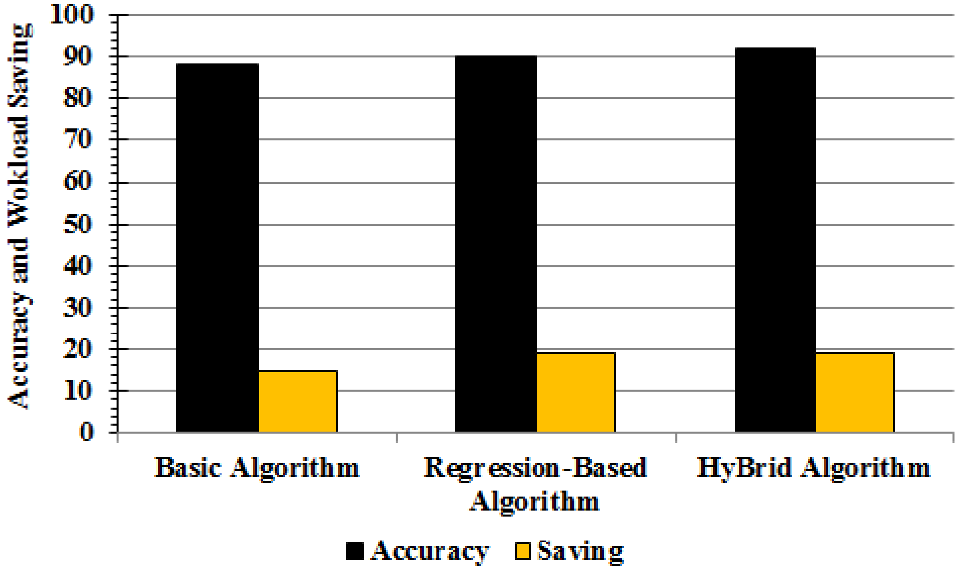


Figure 4.12: Comparison the performance of the three algorithms for all patients in Diabetes dataset

4.4.1.6 The relationship among different measurements

In this experiment, the values of Pearson's Correlation Coefficient for all the pairs of numerical measurements in the dataset are quantified. The purpose of this experiment is to investigate the correlation relationships between different numerical measurements. There are five numerical measurements in the datasets, thus the correlation coefficient for a total of 10 different pairs was evaluated. Table 4.4 presents the values of correlation coefficients of different pairs of measurements for each of the patients and the average value across all the patients. The results show that there are five pairs of measurements ({HR, SBP}, {HR, DBP}, {DBP, SBP}, {DBP, MAP} and {SBP, MAP}) exhibiting relatively strong positive correlations, ranging from 0.55 to 0.63, where other pairs of measurements did not show any strong correlations (HR is Heart rate; SBP is Systolic Blood Pressure; DBP is Diastolic Blood Pressure; MAP is Mean Arterial Pressure; SO2 is Oxygen Saturation). This experiment helps reveal the relationship between different measurements and create the foundation for an improvement of the current recommendation system so that risk prediction of a measurement can be performed on, not only its own historical data, but also on those of another measurement with which it is correlated.

Table 4.4: Correlation coefficient for 5 measurements (HR is Heart rate; SBP is Systolic Blood Pressure; DBP is Diastolic Blood Pressure; MAP is Mean Arterial Pressure; SO2 is Oxygen Saturation)

	PID-31	**PID-32**	**PID-33**	**PID-37**	**PID-39**	**PID-40**	**Average**
(HR,SBP)	0.7	0.5	0.4	0.5	0.6	0.8	0.58
(HR,DBP)	0.5	0.4	0.6	0.5	0.7	0.6	0.55
(HR,SO2)	-0.5	-0.6	-0.3	0	-0.4	-0.5	-0.38
(HR,MAP)	0.01	0.02	0.15	0.01	0.07	0.09	0.058
(DBP,SBP)	0.8	0.6	0.5	0.7	0.5	0.7	0.63
(DBP,MAP)	0.7	0.5	0.7	0.4	0.6	0.5	0.57
(DBP,SO2)	0.01	-0.01	-0.01	0.01	0.09	-0.05	0.007
(SBP,MAP)	0.7	0.6	0.5	0.4	0.5	0.7	0.567
(MAP,SO2)	0.03	0.01	0.06	0.03	0	0.04	0.028
(SBP,SO2)	0.02	0.11	-0.1	0.13	-0.4	-0.2	-0.073

4.4.1.7 The relationship between performance metrics and dataset characteristics

This experiment quantified the values of Pearson's Correlation Coefficient for the performance metrics (*Accuracy and Workload Saving*) and dataset characteristics metrics (*Difference and Rotating Frequency*). The purpose of this experiment was to investigate the correlation relationships between different performance metrics and dataset characteristics metrics. Tables 4.5 and 4.6 present the average value across all patients of correlation coefficients for different pairs of performance metrics and datasets characteristics metrics in the Tunstall and Diabetes datasets. From the obtained results, the interesting observation in Tables 4.5 and 4.6 is a relatively strong positive relationship among the Difference, Accuracy and Workload Saving, whereas the strong negative correlations clearly appeared with other pairs. From this experiment, we conclude that the system yields a higher Accuracy and Workload Saving for all the patients for those medical measurements with a higher value of Difference and a lower value of Rotating Frequency and vice visa.

Table 4.5: Correlation coefficient for performance metrics and Tunstall dataset characteristics

Performance Metrics and Characteristics	**Averaged Correlation Coefficient**
(Differences, Accuracy)	**0.7**
(Differences, Saving)	**0.6**
(Accuracy, Saving)	**0.6**
(Differences, Rotating)	**-0.8**
(Rotating, Accuracy)	**-0.7**
(Rotating, Saving)	**-0.6**

Table 4.6: Correlation coefficient for performance metrics and Diabetes dataset characteristics

Performance Metrics and Characteristics	**Averaged Correlation Coefficient**
(Differences, Accuracy)	**0.6**
(Differences, Saving)	**0.5**
(Accuracy, Saving)	**0.7**
(Differences, Rotating)	**-0.7**
(Rotating, Accuracy)	**-0.6**
(Rotating, Saving)	**-0.5**

4.4.1.8 Summary

This work has illustrated that it is possible to develop an intelligent recommender system, supported by several innovative predictive algorithms, for short-term risk assessment of patients in a telehealth environment based on the analysis of a patient's historical medical data. On the basis of the assessment results, the system provides recommendations to patients in relation to the need to take medical tests on a daily basis. The research is conducted with the aim of improving the quality of clinical evidenced-based decisions made by medical practitioners and reducing the financial and timing cost taken by patients. The work makes a theoretical contribution through the time series prediction algorithms and another contribution with the intelligent system to improve the quality of health care services. The experimental results demonstrate that the proposed system yields reasonably good recommendation accuracy and can effectively reduce the workload required in medical tests for patients.

A structural graph-based method was also applied to study the behavior of time series data in the time domain for the short-term risk assessment of patients. The results obtained are presented in the following section.

4.4.2 The Analysis Results of the Structural Graph-Based Method

4.4.2.1 Prediction accuracy with different numbers of features

First, experiments to evaluate the recommendation performance of the system under different sets of statistical features extracted from the siding windows of the datasets were carried out. Several experiments were carried out to determine the best set of graph features by which the original time series can be represented with the best form. The four graph features were tested separately to evaluate the prediction accuracy of the proposed system. Figure 4.13 shows the ranking of the statistical features based on their performance where the features are sorted in descending order based on their effectiveness in predicting the patient's condition.

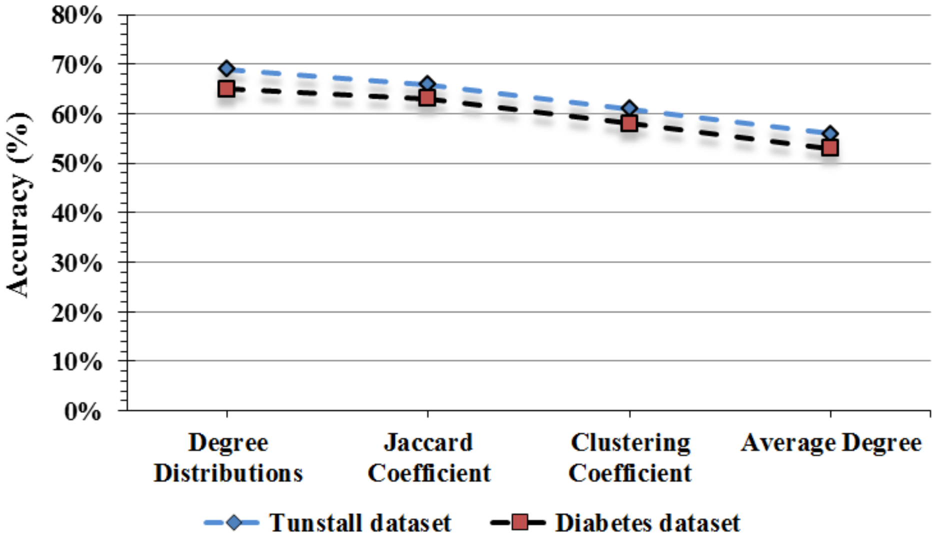


Figure 4.13: Ranking of the graph features based on their accuracy performance

1. **Two-features set**

   To determine the best combination of the two graph features, a set of experiments was designed. A two features set of graphs was picked up from the ordered list in Figure 4.13 and sent to the ensemble classifier. The six permutations of two graph features were tested. Figure 4.14 shows the performance of the proposed method based on the graphs features. Based on the obtained results, it was observed that the combination of Jaccard coefficient and degree distribution recorded the highest accuracy for both datasets compared to other combinations. These two features were able to give a promising prediction. However, the lowest accuracy rate was recorded by the pair of clustering coefficients and average degree. For further investigation, a three features set was tested in the next experiment.

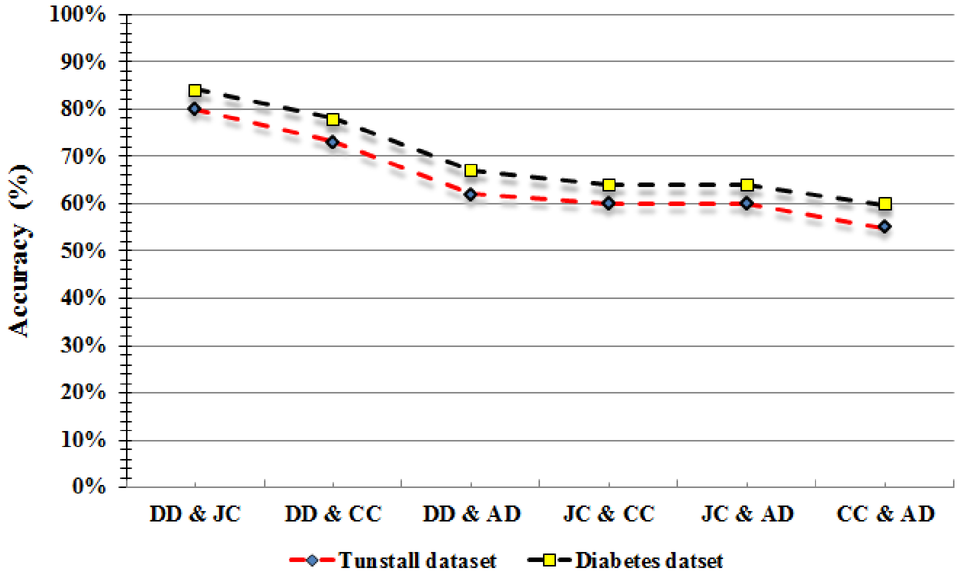


Figure 4.14: Accuracy based on two-features sets. (Note: DD=Degree Distribution; JC=Jaccard Coefficient; CC=Clustering Coefficient; AD=Average Degree).

2. **Three-features set**

To assess the method's ability to predict the status of patients with a high accuracy, the proposed method was tested using a three features set. The first three graph features in Figure 4.13 were selected. The three features were degree distribution, Jaccard coefficient and clustering coefficient. Figure 4.15 shows the performance of the proposed method using three and four features sets. The most noticeable results from this experiment were that the prediction accuracies exceeded 94% for both datasets compared with the sets of two features. For more accurate results, different experiments were designed with different data sizes. The results showed a stability in the performance of the proposed method. Another three features set was also tested, however, the results confirmed that the three features set of degree distribution, Jaccard

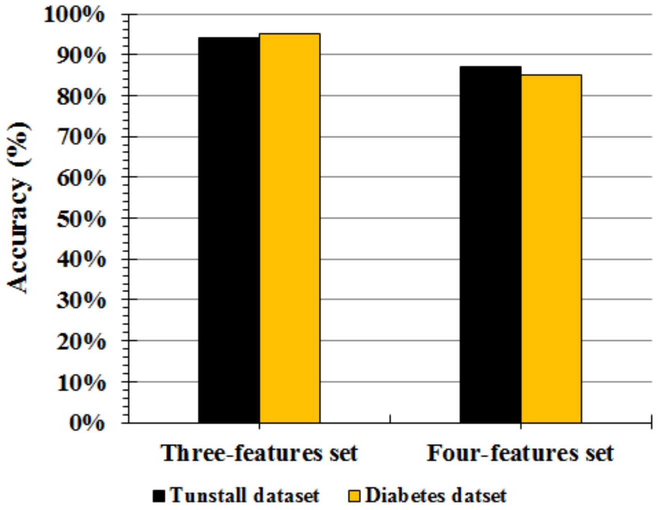


Figure 4.15: Accuracy based on three and four features sets

coefficient and clustering coefficient was the best combination of the graphs features to provide the recommendation accurately.

Four features was also tested and investigated. Based on the results in Figure 4.15, the prediction accuracy of the proposed method was lower compared with the three features set. A rate of 86% using all the graph features (degree distribution, Jaccard coefficient, clustering coefficient and average degree) was achieved. The combination of the first three features of degree distribution, Jaccard coefficient and clustering coefficient were considered to achieve the best accuracy.

4.4.2.2 Prediction accuracy with different sizes of sliding windows

The impact of window size is also studied. In this experiment, the best window size was investigated to obtain the desired prediction accuracy. From the obtained results, it is clear that there is a positive relationship between the selected size of sliding window and the predictive performance of the proposed system. It was found that, when the number of nodes in a graph was increased due to the increasing the size of a sliding window, the proposed method generated more accurate recommendations. To determine the optimum size of a slide, a set of experiments was conducted with different window sizes. The model performance was found to be improved by increasing the size of the sliding window (the number of nodes). This is because the characteristics of time series data were clearly presented when the number of graph nodes was increased. As a result, the proposed model was tested with different window sizes at 7 days, 10 days, 15 days and 20 days. In these experiments, the three features set of degree distribution, Jaccard coefficient and clustering coefficient were considered. The four medical attributes of Heart Rate, Diastolic Blood Pressure (DBP), Mean Arterial Pressure (MAP), and Oxygen Saturation (SO2) were used in the following experiments.

1. **Sliding window of 7 days**

   A sliding window of 7 days was used to test the predictive performance of the proposed method. For each day, the sliding window was represented by a node in a graph. The three structural properties of the graphs were extracted and considered the key features to represent each window. The metrics of accuracy, workload saving and risk for all the graphs were calculated to verify the performance of the proposed method. Table 4.7 presents the metrics of accuracy, workload saving and risk for each measurement in the Tunstall dataset.

   Based on the obtained results, it was noticed that the performance of the proposed method was not good enough to predict the patient's condition as the number of the graph nodes was not enough to reflect the behaviours of the time series data. To tackle this issue, the number of nodes in each graph was increased by considering a new window size. In the next experiment, the influence of a 10 days window size was discussed.

2. **Sliding window of 10 days**

Table 4.7: Performance evaluation based on a sliding windows of 7 days

Dataset	**Measurement**	**Accuracy(%)**	**Saving(%)**	**Risk(%)**
Tunstall	Heart Rate	86.37	60.34	05.21
	DBP	85.30	57.18	05.40
	MAP	87.70	61.33	05.00
	SO2	84.30	55.44	05.90
Diabetes	BG	83.90	55.30	05.70

The time series data were segmented into windows using a sliding window of 10 days and then each window was transferred into a graph. As previously mentioned, 10 days sliding windows were considered to improve the accuracy of the proposed method and give more accurate recommendations. One of the interesting findings was that the proposed method yielded a high performance using 10 days sliding window compared with the window size of 7 days. The performance of the proposed method significantly improved due to the increased number of graph nodes. The graphs nodes were found to reflect big differences between the patient states which include whether he/she requires medical a test or not. Table 4.8 shows the results obtained by the proposed method after considering the window size of 10 days.

Based on the obtained results in Table 4.8, it is interesting to note that the accuracy values for all measurements improved by more than 5% compared to the results in Table 4.7. In addition, using a window size of 10 days did not considerably affect the performance of workload saving, although the accuracy and risk were increased.

3. **Sliding window of 15 days**

For further investigation, a window size of 15 days was adopted to test the performance of the proposed method. In this experiment, the size of window was increased to 15 days. Table 4.9 presents the metrics of accuracy, workload

Table 4.8: Performance evaluation based on a sliding window of 10 days

Dataset	**Measurement**	**Accuracy(%)**	**Saving(%)**	**Risk(%)**
Tunstall	Heart Rate	92.75	59.80	03.95
	DBP	91.40	58.77	04.50
	MAP	90.55	60.65	04.80
	SO2	91.60	55.50	04.20
Diabetes	BG	90.85	60.00	04.90

saving and risk for all the measurements using sliding window size of 15 days. Based on results in Table 4.9, the average accuracy of the proposed method exceeded 94% across the different measurements. The obtained results proved that the size of the window has the potential to significantly impact the accuracy of the prediction for all measurements. The connectivity among the graph nodes (clustering coefficients) are strong enough to reveal the differences between time series data.

Different window sizes including 20, 25 and 30 days were also tested and evaluated in this study. There were no significant differences compared with the results obtained using the 15 days sliding windows. Thus, the optimal window size is 15 days because it reflects the actual behavior of the time series data.

Table 4.9: Performance evaluation based on a sliding windows of 15 days

Dataset	**Measurement**	**Accuracy(%)**	**Saving(%)**	**Risk(%)**
Tunstall	Heart Rate	94.80	62.30	02.60
	DBP	93.80	59.50	03.60
	MAP	93.90	61.40	03.40
	SO2	94.60	61.80	02.90
Diabetes	BG	94.60	61.20	02.70

4.4.2.3 Performance evaluation based on a single classifier as well as ensemble model

In this experiment, the performance of the system using a sliding window of 15 and the three graph features set based on the previous information was evaluated. Table. 4.10 shows the results of comparison among the ensemble classifier and the individual classifiers. Based on these results, the system performance using individual classifier was between 80% and 85% across different measurements. The maximum accuracy of 85% was obtained by the LS-SVM, while the minimum accuracy of 80% was gained by Naive Byes. We can notice that, although the proposed system was conducted with different classifiers, there is not a big fluctuation in its performance and the accuracy values of those classifiers are quite close. A 'gold' solution for improving the performance of the proposed method and decreasing the error rate, is to combine multi-classifiers to classify the extracted features.

Table 4.10: Performance evaluation based on a single classifier and an ensemble model

Classifier	Accuracy(%)	Saving(%)	Risk(%)
LS-SVM	85.30	61.10	04.30
Neural Network	83.50	60.80	05.10
Naive Bayes	80.40	60.95	04.90
Ensemble model	**94.27**	**61.25**	**03.01**

However, in this work, an ensemble machine learning was used to classify the graphs features. The recommendation system achieved a better prediction accuracy compared with the individual classifiers; with an increase of 12%. As mentioned above, each classifier was trained and conducted with the datasets separately and were then combined according to an appropriate criteria. By comparing the results in Table. 4.10, we can observe that the performance of the proposed system was improved when the ensemble machine learning was adopted.

Furthermore, the performance of the proposed system for a single classifier as well as ensemble model were evaluated by the receiver operating characteristics (ROC) curve. The ROC plot is a graph of true positive rate ($y - axis$) against false positive rate ($x - axis$) (Fawcett, 2006). The ROC provides a better measure of the recommendation accuracy.

The evaluation results for all the classifiers and the ensemble model were supported using ROC curve. From Figure 4.16, it is clear that the ROC for the ensemble model is 0.94, which indicates that the proposed method is the best method for providing accurate recommendations to patients. The second best result was obtained with the LS-SVM classifier, while the Naive Bayes recorded the lowest area under ROC curve among the four classifiers.

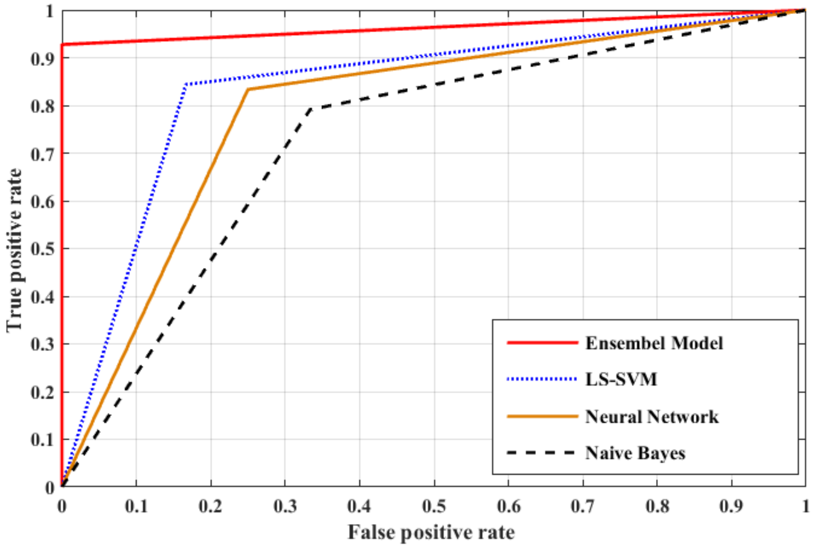


Figure 4.16: The ROC plot of the three classifiers and the ensemble model

For more investigation, the execution time of the proposed model was calculated based on the ensemble classifier as well as individual classifiers. Figure 4.17 shows the complexity time for each individual classifier and the ensemble model. We observed that the ensemble model took more time to complete the training and prediction than the individual base classifier. This is reasonable as the ensemble model needs to aggregate the results from the base classifiers to generate the weights and produce the final recommendation. The ensemble model sacrifices a little on the execution time to achieve better recommendation effectiveness for patients. Additionally, the training stage can be performed off-line so that it will not adversely affect the efficiency in generating recommendations for patients during the prediction stage.

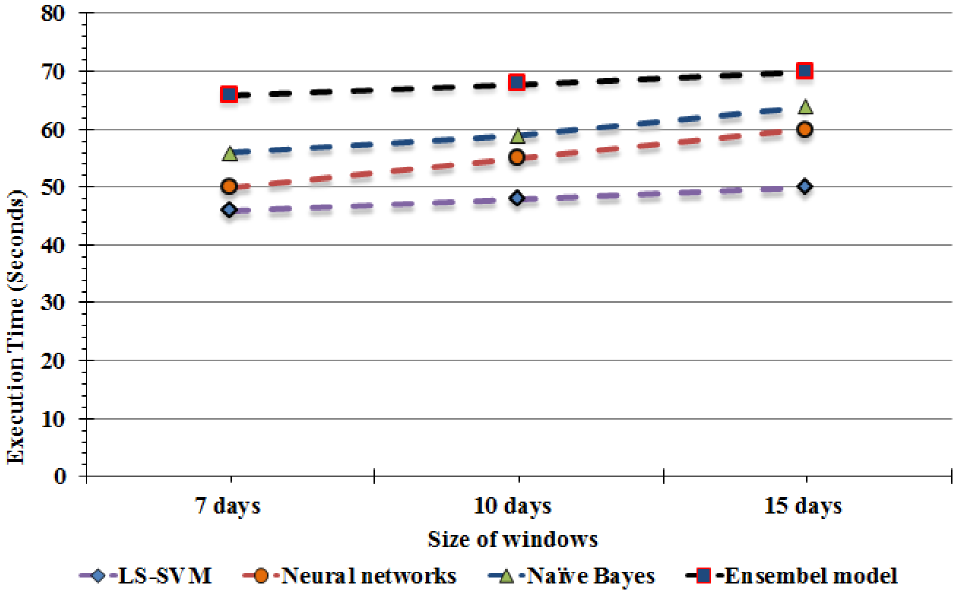


Figure 4.17: Comparison of the execution time between the classifiers and the ensemble model under different sliding windows

4.4.2.4 Summary

This work proposes a recommendation system supported by the structural graph properties and advanced machine learning ensemble for short-term disease risk prediction and medical test recommendation in a telehealth environment for chronic diseases patients. This study applies the the structural graph, which effectively represents the medical time series data and enters the extracted statistical features to the ensemble model to generate the accurate, reliable recommendations for chronic disease patients. Three popular and capable classifiers, Least Square-Support Vector Machine, Neural Network, and Naive Bayes, were used to construct the ensemble framework.

The experimental results showed that the proposed system using a sliding windows of 15 days with the optimal statistical features set produced by the structural graph properties yielded a better predictive performance for both datasets. The results also showed that the proposed system using the ensemble classifier with optimal features set can correctly predict up to 94% of the subjects across all measurements. The system is also more effective than the individual base classifiers used in the ensemble model and outperforms the previously proposed approaches to solve the same problem. The evaluation establishes that the recommendation system is effective in improving the quality of clinical evidence-based decisions and reduces the time costs incurred by patients taking their daily medical test, thereby improving their overall life quality.

In summary, three innovative predictive algorithms and a structural graph-based method have been proposed to study the behavior of time series data in the time domain for short-term risk assessment of patients in a telehealth environment based on an analysis of a patient's historical medical data.

Furthermore, two decomposition methods, the fast Fourier transformation and dual-tree complex wavelet transformation, were applied to study the behavior of the time series data of patients in the frequency domain. The detailed experimental results of these decomposition methods are presented in the following section.

4.5 The Analysis Results of the Frequency Domains Techniques

In this section, two decomposition techniques – the fast Fourier transformation and dual-tree complex wavelet transformation – are used to study the behavior of time series data in the frequency domain for short-term risk assessment of patients in a telehealth environment based on the analysis of a patient's historical medical data.

4.5.1 The Evaluation of Fast Fourier Transformation

4.5.1.1 Recommendation effectiveness under different statistical features

First, experiments to evaluate the recommendation performance of the system under different sets of statistical features extracted from the siding windows of the Tunstall dataset were conducted.

Eight statistical features were tested separately to evaluate the prediction accuracy of the proposed system. Figure 4.18 shows the ranking of the statistical features based on their performance with features sorted in ascending order based on their effectiveness in predicting a patient's condition. In this experiment, the extracted statistical features from all 28 frequency bands were used as the input for training the base classifiers in the ensemble model.

An experiment to investigate the recommendation effectiveness of the system under different numbers of features was also conducted. To this end, the performance of the system under different feature sets whose number of features ranged from two to eight were tested and recorded thoroughly. The results are presented in Table 4.11. Based on these results, it is observed that there is a positive correlation between

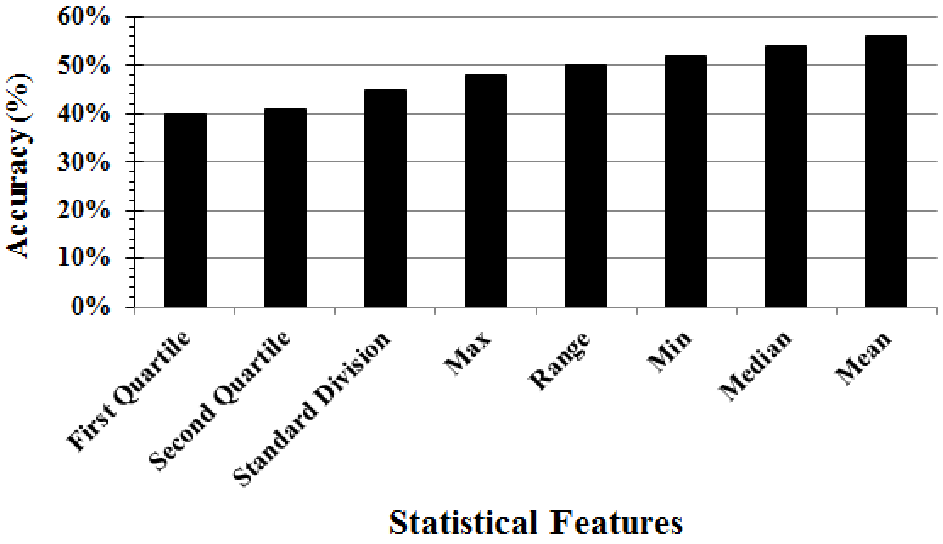


Figure 4.18: Ranking of the statistical features based on their accuracy performance

the number of extracted statistical features and the recommendation performance of the proposed system in terms of all three performance metrics used. The system becomes more accurate when the number of statistical features used increases.

Finally, the system's performance was evaluated under three different feature sets containing three, six and eight statistical features respectively based on the sorted feature information presented in Figure 4.18. Only the top three, six and eight set features in the Figure were selected into the three sets. The detailed statistical features selected for the three evaluated sets are as follows:

Three-feature set: X_{Mean}, X_{Med} and X_{Min}

Six-feature set: $X_{Mean}, X_{Med}, X_{Min}, X_{RG}, X_{Max}$ and X_{SD}

Eight-feature set: X_{Mean}, X_{Med}, X_{Min}, X_{RG}, X_{Max}, X_{SD}, X_{FQ} and X_{SQ}

For each of the three feature sets evaluated, system performance was further evaluated when only five and all the 28 frequency bands are used for each sliding window,

Table 4.11: The average performance of the proposed model under different numbers of features

Feature sets	Accuracy (%)	Saving (%)	Risk (%)
2-feature set	86.80	62.00	05.20
3-feature set	87.00	62.12	05.00
4-feature set	92.50	64.00	04.50
5-feature set	92.70	64.20	04.20
6-feature set	93.00	64.50	04.00
7-feature set	94.80	64.80	03.00
8-feature set	94.83	64.80	02.30

indicating the impact of the choice of frequency bands on the performance of the system. The five frequency bands used were $\alpha, \beta, \gamma, \delta$ and θ, respectively. Under this setup, the total number of features extracted for each sliding window in different experimental setups can be calculated as $N_b \times N_f$, where N_b and N_f represent the number of frequency bands and statistical features selected, respectively.

The detailed experiment results are presented in Table 4.12, Table 4.13 and Table 4.14 for three, six and eight-feature sets, respectively. First, it can be seen that the recommendation performance of the system, for all the four medical measurements, is noticeably improved compared with the performance when using only five frequency bands. This is because using all 28 frequency bands can better capture the characteristics of each sliding window than using just five. Second, the results from the three tables reveal a similar trend; that the performance of the system in terms of all three performance metrics is generally improved when the number of the statistical features used is increased. For example, under 28 frequency bands, the recommendation accuracy and risk of the system, when using eight statistical features, is approximately 2.4% and 42.9% better than using six statistical features across different medical measurements while maintaining roughly the same level of workload saving at the same time. Therefore, to achieve the best possible performance, it is suggested to use all eight statistical features for all the 28 frequency bands to generate recommendations for patients. Table 4.15 shows the averaged performance of the proposed model based on an eight-feature set for both datasets. Based on the results obtained from Table 4.15, we can see that, for both datasets, we can achieve a very high accuracy and workload saving as well as a very low risk. For instance, when using 28 frequency bands and eight statistical features, the system can achieve an accuracy over 94% and a workload saving over 64% with a risk lower than 2.6%, indicating that the recommendation system is highly accurate and able to significantly reduce the workload for chronic heart disease patients undertaking their daily medical tests with a very low health risk.

Table 4.12: The model performance using a three-feature set

Tunstall dataset(%)

Measurement	Classifiers	No. of bands	Accuracy (%)	Saving (%)	Risk (%)
Heart Rate	Neural Network	5	71.60	55.54	09.72
		28	72.60	56.56	09.90
	LS-SVM	5	76.49	61.55	07.30
		28	77.78	60.20	07.20
	Naive Bayes	5	72.85	54.55	09.60
		28	71.60	56.60	09.70
	Ensemble model	5	**86.44**	**61.60**	**5.70**
		28	**87.30**	**62.50**	**5.20**
DBP	Neural Network	5	70.10	54.30	09.90
		28	71.30	55.88	09.10
	LS-SVM	5	75.44	62.51	08.10
		28	77.20	60.10	07.80
	Naive Bayes	5	69.20	52.30	09.60
		28	70.36	60.60	07.50
	Ensemble model	5	**85.50**	**60.40**	**5.90**
		28	**88.40**	**61.88**	**5.10**
MAP	Neural Network	5	69.90	50.20	10.50
		28	70.10	53.50	09.80
	LS-SVM	5	73.55	59.40	09.40
		28	75.78	60.90	08.95
	Naive Bayes	5	72.20	58.60	09.95
		28	73.40	60.20	09.80
	Ensemble model	5	**84.50**	**60.20**	**6.10**
		28	**86.40**	**60.60**	**5.95**
SO2	Neural Network	5	70.75	60.50	09.85
		28	71.50	60.80	09.40
	LS-SVM	5	77.60	62.65	06.70
		28	78.20	63.40	06.10
	Naive Bayes	5	70.80	55.30	09.50
		28	71.30	56.40	09.30
	Ensemble model	5	**85.80**	**63.40**	**5.90**
		28	**86.75**	**63.50**	**5.20**

Table 4.13: The model performance using a six-feature set

Tunstall dataset(%)

Measurement	Classifiers	No. of bands	Accuracy (%)	Saving (%)	Risk (%)
Heart Rate	Neural Network	5	73.50	59.50	09.30
		28	74.60	59.80	08.80
	LS-SVM	5	78.20	62.60	06.80
		28	79.40	63.44	06.20
	Naive Bayes	5	75.30	55.20	08.30
		28	75.90	55.60	08.10
	Ensemble model	5	**91.20**	**63.50**	**4.60**
		28	**92.50**	**63.20**	**4.20**
DBP	Neural Network	5	72.70	50.80	09.70
		28	73.60	54.40	09.25
	LS-SVM	5	78.50	64.60	06.75
		28	80.40	64.80	06.10
	Naive Bayes	5	72.55	57.40	09.90
		28	73.40	58.30	09.50
	Ensemble model	5	**90.10**	**64.10**	**4.95**
		28	**92.80**	**65.70**	**4.05**
MAP	Neural Network	5	71.55	55.30	09.40
		28	72.50	56.40	09.10
	LS-SVM	5	75.60	62.54	08.25
		28	78.54	64.25	07.30
	Naive Bayes	5	73.30	63.70	09.95
		28	75.30	65.35	08.70
	Ensemble model	5	**91.60**	**64.30**	**4.20**
		28	**92.70**	**64.80**	**4.10**
SO2	Neural Network	5	75.50	61.20	08.50
		28	77.10	61.90	07.60
	LS-SVM	5	80.20	64.10	06.60
		28	81.10	64.80	06.15
	Naive Bayes	5	72.50	58.40	09.10
		28	73.50	60.50	08.95
	Ensemble model	5	**90.50**	**64.20**	**4.95**
		28	**92.50**	**64.60**	**4.30**

Table 4.14: The model performance using an eight-feature set

Tunstall dataset(%)

Measurement	Classifiers	No. of bands	Accuracy (%)	Saving (%)	Risk (%)
Heart Rate	Neural Network	5	80.40	60.60	06.80
		28	81.50	61.70	06.40
	LS-SVM	5	82.20	63.55	06.20
		28	83.30	64.20	06.00
	Naive Bayes	5	80.55	59.80	06.70
		28	80.90	60.20	06.60
	Ensemble model	5	**93.40**	**64.60**	**2.80**
		28	**94.80**	**64.30**	**2.60**
DBP	Neural Network	5	81.40	61.75	06.60
		28	82.10	62.30	06.30
	LS-SVM	5	82.20	64.10	06.20
		28	83.50	64.80	05.80
	Naive Bayes	5	80.40	60.20	07.80
		28	80.80	60.50	07.50
	Ensemble model	5	**94.25**	**63.30**	**2.95**
		28	**94.85**	**64.10**	**2.40**
MAP	Neural Network	5	79.40	58.30	08.10
		28	80.30	59.20	07.90
	LS-SVM	5	80.65	60.50	07.70
		28	81.45	63.60	06.60
	Naive Bayes	5	78.50	60.40	08.90
		28	79.10	61.45	08.80
	Ensemble model	5	**93.20**	**63.25**	**3.20**
		28	**94.20**	**63.70**	**2.60**
SO2	Neural Network	5	82.10	64.10	06.50
		28	83.40	64.80	05.70
	LS-SVM	5	84.20	65.30	04.90
		28	84.90	65.70	04.20
	Naive Bayes	5	81.30	62.50	06.75
		28	82.50	63.20	06.20
	Ensemble model	5	**94.80**	**64.75**	**2.40**
		28	**95.50**	**65.80**	**1.90**

Table 4.15: Model performance average based on an eight-feature set using Tunstall and Diabetes datasets

Dataset	Classifiers	Accuracy (%)	Saving (%)	Risk (%)
Tunstall	Neural Network	81.28	62.00	06.57
	LS-SVM	83.28	64.57	05.65
	Naive Bayes	80.82	61.33	07.27
	Ensemble model	**94.66**	**64.21**	**2.50**
Diabetes	Neural Network	80.60	60.50	06.80
	LS-SVM	83.00	64.00	05.10
	Naive Bayes	79.90	60.00	07.00
	Ensemble model	**94.40**	**64.00**	**2.80**

4.5.1.2 Efficiency Comparison

An efficiency study was also conducted to compare the execution time of the base classifiers and the ensemble model under different feature sets. Figure 4.19 and Figure 4.20 present the execution time of the base classifiers and the ensemble model in their training and prediction stages.

From the results, we observe that: 1) The execution time of the training stage is higher than that of the prediction stage for all the models evaluated; 2) All the base classifiers and the ensemble model feature an approximately linear execution time in both the training and prediction stages under different feature sets, ensuring the necessary efficiency for producing fast recommendations for patients; 3) Of all models, the ensemble model takes more time to complete the training and prediction than the individual base classifier. This is reasonable as the ensemble model needs to aggregate the results from the base classifiers to generate the weights for them and produce the final recommendation. The ensemble model sacrifices a little on the execution time to achieve better recommendation effectiveness for patients. Additionally, the training stage can be performed off-line so that it will not adversely affect the efficiency in generating recommendations for patients during the prediction stage.

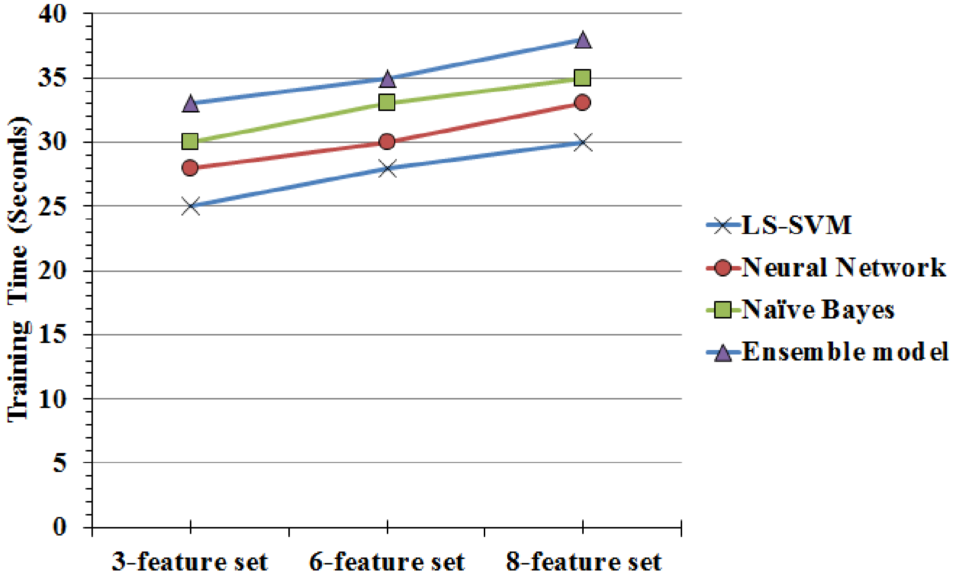


Figure 4.19: Comparison of the training time between the base classifiers and the ensemble model

4.5.1.3 Summary

This work proposes a recommendation system supported by a machine learning ensemble model with the fast Fourier transformation for short-term disease risk prediction and medical test recommendation in a telehealth environment for patients suffering from chronic disease such as heart disease and diabetes. This study applies the fast Fourier transformation, which effectively analyzes the medical time series data and enters the extracted statistical features to the ensemble model to generate accurate and reliable recommendations for chronic disease patients. Three popular and capable classifiers, Neural Network, Least Square-Support Vector Machine and

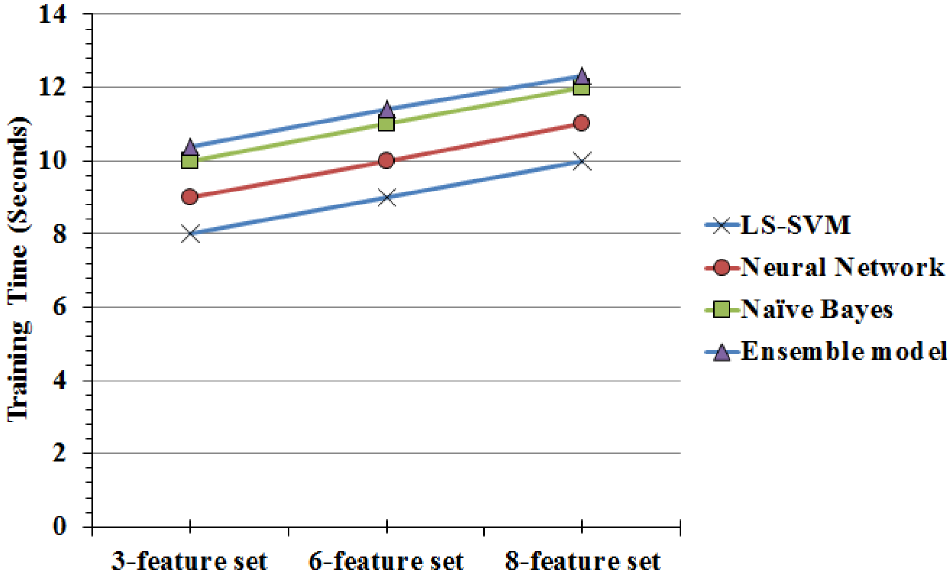


Figure 4.20: Comparison of the prediction time between the classifiers and the ensemble model

Naive Bayes are used to construct the ensemble framework.

The experimental results show that the proposed system, using all the eight statistical features produced by the fast Fourier transformation, yielded a better predictive performance for the patient's condition compared with the other feature sets. The results also show that the system is more effective than the individual base classifiers used in the ensemble model and outperforms the previously proposed approaches to solve the same problem. This evaluation establishes that the recommendation system is effective in improving the quality of clinical evidence-based decisions and reduces the time costs incurred by patients with chronic disease in taking their daily medical test, thereby improving their overall life quality.

The study also applied another technique, namely a dual-tree complex wavelet transformation, to study the behavior of time series data in the frequency domain for the short-term risk assessment of patients. The results obtained are presented in the next section.

4.5.2 The Evaluation of Dual-Tree Complex Wavelet Transformation

4.5.2.1 Recommendation effectiveness under different sub-bands and statistical features

The first experiment was designed to evaluate the recommendation performance of the proposed system under various DTCWT sub-bands. The optimum number of features from these sub-bands were selected and extracted based on the performance of the proposed system. To be more rigorous in statistical analysis, a t-test was performed to assess the statistical significance of the results in this work. The p-values were obtained by comparing the features extracted from the sub-bands. The t-test was conducted separately for all sub-bands (y1, y2, y3, y4 and z4). Table 4.16 shows the corresponding values of accuracy, workload saving, risk and p values obtained from each sub-band based on the features. From the associated p-value analysis, it is clear that the high frequency sub-bands (i.e., y1, y2, and y3) provide a significant difference ($p < 0.003$) towards good recommendation performance, indicating that the high frequency sub-bands (i.e., y1, y2, and y3) are able to produce more accurate recommendations compared to the low-frequency sub-bands (i.e., y4 and z4). Hence, features from the high frequency sub-bands (i.e., y1, y2, and y3) were selected to represent the time series data of the patient, while the features from the low-frequency sub-bands (i.e., y4 and z4) were ignored.

Different sub-bands sets (combinations) were tested and recorded thoroughly to identify the best combination of sub-bands for system design. These results presented in Table 4.17 show that the performance of the proposed system improves significantly when a combination of the three high-frequency sub-bands (y1, y2, and y3) was used. In this experiment, the extracted statistical features from the coefficients of y1, y2, and y3 sub-bands were used as the input to the base classifiers as well as to the

Table 4.16: Prediction performance using single sub-bands

Names of the used sub-bands	Accuracy (%)	Saving (%)	Risk (%)	*p*-value
y1	86.18	62.95	05.50	3.3359E-3
y2	89.30	63.85	05.10	2.2058E-5
y3	90.40	64.95	04.95	2.4268E-6
y4	82.45	60.70	06.70	0.01049
z4	81.15	60.80	06.20	0.35026

Table 4.17: Prediction performance using combinations of sub-bands

Names of the used sub-bands	Accuracy (%)	Saving (%)	Risk (%)	*p*-value
y1,y2	94.20	64.20	02.50	6.2245E-4
y1,y3	94.90	65.40	02.20	3.4129E-5
y2,y3	95.45	65.55	01.90	4.6328E-6
y1,y2,y3	96.70	66.50	01.25	2.8493E-11

ensemble model.

However, to study the relationship between the number of the extracted features and the recommendation effectiveness of the system, several experiments were conducted using different sets of extracted features. First, the six statistical features were tested separately using the high frequency sub-bands. Figure 4.21 shows the order of the statistical features based on their performance. The features were then sorted in ascending order based on their potential patient's condition. In this experiment, the extracted statistical features from the high-frequency sub-bands (i.e., y1, y2, and y3) were used as the input for the ensemble model.

Furthermore, to produce accurate recommendations, several experiments under different numbers of features were conducted. The performance of the system under

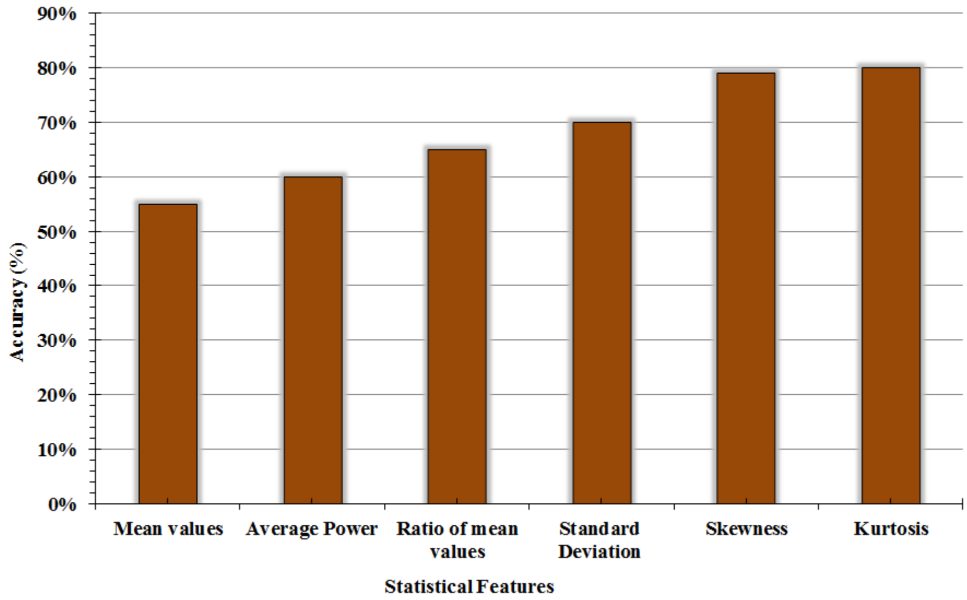


Figure 4.21: Ranking of the statistical features based on their accuracy performance

different feature sets was tested and recorded thoroughly. The results obtained are presented in Table 4.18. Based on these results, a positive correlation is observed between the number of the extracted statistical features and the recommendation performance of the proposed system in terms of all of the three performance metrics used. The system was found to be more accurate when the number of statistical features used increased.

Finally, to investigate the recommendation effectiveness of the proposed system under different numbers of features, the performance of the system was evaluated under three different feature sets containing two, four and six statistical features respectively based on the sorted feature information presented in Figure 4.21. Only the top two, four and six set features in Figure 4.21 were selected in the three sets. The detailed statistical features selected for the three evaluated sets are as follows:

Two-feature set: Kurtosis and Skewness

Four-feature set: Kurtosis, Skewness, Standard Deviation and Ratio of mean values

Table 4.18: The average performance of the proposed system under different numbers of features

Feature sets	Accuracy (%)	Saving (%)	Risk (%)
2-feature set	90.37	63.90	04.56
3-feature set	91.00	64.35	04.30
4-feature set	94.05	64.06	02.60
5-feature set	94.70	64.50	02.20
6-feature set	96.12	65.48	01.56

Table 4.19: The model performance using a two-feature set

Measurement	Classifiers	Accuracy (%)	Saving (%)	Risk (%)
Heart Rate	LS-SVM	87.25	61.35	05.60
	Neural Network	85.80	60.55	06.00
	Naive Bayes	86.82	60.10	05.90
	Ensemble model	**90.20**	**63.50**	**04.70**
DBP	LS-SVM	88.44	63.51	05.00
	Neural Network	84.10	60.30	06.75
	Naive Bayes	83.20	61.30	06.90
	Ensemble model	**90.50**	**63.82**	**04.90**
MAP	LS-SVM	86.55	64.85	05.90
	Neural Network	82.90	59.35	07.00
	Naive Bayes	85.20	62.80	06.00
	Ensemble model	**90.30**	**64.90**	**04.20**
SO2	LS-SVM	86.20	64.65	05.75
	Neural Network	86.50	63.50	05.85
	Naive Bayes	85.70	62.30	05.90
	Ensemble model	**90.50**	**63.40**	**04.45**

Six-feature set: Kurtosis, Skewness, Standard Deviation, Ratio of mean values, Average power and Mean values

Table 4.20: The model performance using a four-feature set

Measurement	Classifiers	Accuracy (%)	Saving (%)	Risk (%)
Heart Rate	LS-SVM	91.60	63.55	03.95
	Neural Network	90.30	62.65	04.50
	Naive Bayes	90.75	62.25	04.45
	Ensemble model	**94.10**	**63.40**	**02.60**
DBP	LS-SVM	92.20	63.31	03.60
	Neural Network	91.80	62.74	03.80
	Naive Bayes	90.66	63.15	04.50
	Ensemble model	**94.60**	**64.54**	**02.50**
MAP	LS-SVM	92.25	62.20	03.55
	Neural Network	89.30	60.35	04.80
	Naive Bayes	90.75	62.10	04.40
	Ensemble model	**93.80**	**64.90**	**02.95**
SO2	LS-SVM	92.57	64.35	03.30
	Neural Network	90.80	62.22	04.35
	Naive Bayes	91.50	62.80	03.95
	Ensemble model	**93.70**	**63.40**	**02.35**

For each of the three feature sets evaluated, the performance of the system was further evaluated when only the three high-frequency sub-bands were used for each sliding window, indicating the impact of the choice of sub-bands on the performance of the system. The three sub-bands used were y1, y2, and y3 respectively.

Furthermore, the performance of the proposed system was evaluated by the receiver operating characteristics (ROC) curve. The ROC curve is a suitable metric for studying the dependency of sensitivity and specificity. The relationships between the true positive rate and false positive rate were investigated using the ROC. The ROC curve is represented by a graph in which the false positive rate is plotted on the x-axis and the true positive rate is plotted on the y-axis. The perfect point in the ROC is represented by the point (0,1) (Fawcett, 2006).

Table 4.21: The model performance using a six-feature set

Measurement	Classifiers	Accuracy (%)	Saving (%)	Risk (%)
Heart Rate	LS-SVM	94.55	63.95	02.55
	Neural Network	93.15	63.20	02.90
	Naive Bayes	92.30	62.85	03.60
	Ensemble model	**96.90**	**65.88**	**01.20**
DBP	LS-SVM	94.66	64.80	02.60
	Neural Network	94.20	64.30	02.80
	Naive Bayes	92.45	62.20	03.50
	Ensemble model	**96.20**	**64.90**	**01.20**
MAP	LS-SVM	94.25	64.80	02.75
	Neural Network	93.30	64.20	02.74
	Naive Bayes	91.75	62.30	03.80
	Ensemble model	**95.60**	**65.77**	**02.00**
SO2	LS-SVM	94.57	64.95	02.40
	Neural Network	92.80	62.65	03.10
	Naive Bayes	93.50	63.50	02.70
	Ensemble model	**95.80**	**65.40**	**01.85**

Table 4.22: Model performance average based on a six-feature set using Tunstall and Diabetes datasets

Dataset	Classifiers	Accuracy (%)	Saving (%)	Risk (%)
Tunstall	LS-SVM	94.50	64.63	2.57
	Neural Network	93.36	63.58	2.88
	Naive Bayes	92.05	62.71	03.04
	Ensemble model	**96.12**	**65.48**	**1.56**
Diabetes	LS-SVM	94.00	64.50	02.60
	Neural Network	93.10	63.20	02.90
	Naive Bayes	92.20	62.10	03.30
	Ensemble model	**96.00**	**65.10**	**1.80**

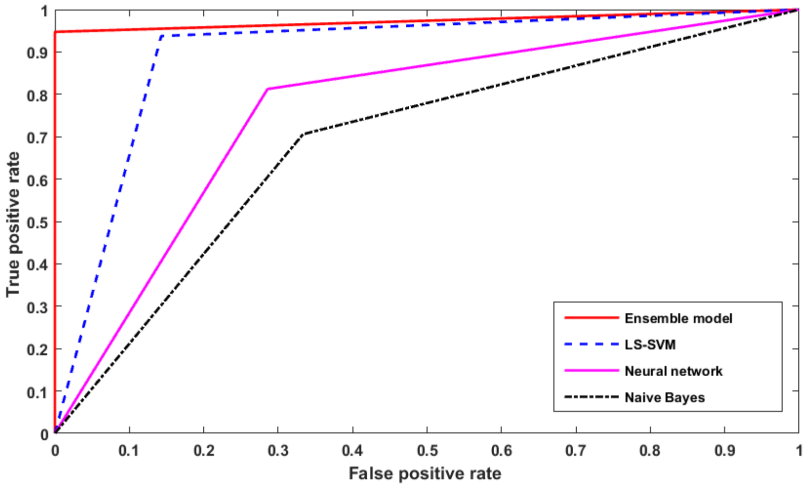


Figure 4.22: The ROC plot of the classifiers and the ensemble model

The evaluation results for all the classifiers and the ensemble model were supported using the ROC curve. From Figure 4.22, it is clear that the biggest area under the ROC curve was constructed by the ensemble model. The second best results were obtained with the LS-SVM classier, while the Naive Bayes recorded the lowest area under ROC curve among the four classifiers. The results of the area under the curve of 0.96 were reported from the ensemble model. Based on these results, the ensemble model was shown to be the best classifier for short-term disease risk prediction.

The detailed experiment results are presented in Table 4.19, Table 4.20 and Table 4.21 for the two, four and six-feature sets, respectively. First, the results from the three tables reveal a similar trend that the performance of the system in terms of all the three performance metrics, is generally improved when the statistical features used are increased. For example, the recommendation accuracy, workload saving and risk of the system when using six statistical features is approximately 2.2%, 2.2% and 47.3% better than using four statistical features across different medical measurements. Therefore, to achieve the best possible performance, it is suggested that all six statistical features be used for the three high-frequency sub-bands when generating recommendations for patients. Table 4.22 shows the averaged performance of the proposed model based on a six-feature set for both datasets. It is clear that the recommender system has achieved a significant accuracy and workload saving with a low risk. For instance, when using six statistical features, the system can achieve an accuracy over 96% and a workload saving over 65% while the risk rate is around 1.56%. These results demonstrate that the recommendation system is highly accurate and able to significantly reduce the workload for chronic heart disease patients when undertaking their daily medical tests with a very low health risk.

4.5.2.2 Effectiveness comparison with previous approaches

The previous subsection compared the performance of the proposed ensemble model with individual classifiers. The experimental results show that the proposed system outperformed the base classifiers in terms of recommendation performance.

This section compares the proposed system with our previous work that tackles exactly the same problem as presented in this work. In that previous work, a basic heuristic approach was developed based on some fundamental heuristic rules for producing appropriate recommendations for patients with chronic heart disease (Lafta et al., 2015). The basic heuristic approach was then combined with the regression-based prediction algorithm and a hybrid method (Lafta, Zhang, Tao, Li, Tseng, Luo and Chen, 2016) to tackle the same task. A fast Fourier transformation-coupled with a machine learning based ensemble model was used to provide appropriate recommendations to patients suffering chronic diseases (Lafta et al., 2017). Table 4.23 presents the comparison results. The results revealed that the proposed system achieved the best accuracy of the four approaches. In this system, the accuracy performance was improved from 94% to 96% while the workload saving experienced a small improvement from 63% to more that 65%. The recommendation risk of the system is also lower than the three competitive approaches.

Table 4.23: Recommendation performance comparison with previous methods

Tunstall dataset

Method	Techniques used	Accuracy(%)	Saving(%)	Risk(%)
(Lafta et al., 2015)	Basic heuristic algorithm	86	10	8
(Lafta, Zhang, Tao, Li, Tseng, Luo and Chen, 2016)	Basic heuristic algorithm, Regression-based algorithm and Hybrid algorithm	91	15	5
(Lafta et al., 2017)	Fast Fourier transformation and ensemble model	94	63	3
Proposed method	Dual-tree complex wavelet transformation and ensemble model	**96**	**65**	**1**

4.5.2.3 Efficiency comparison

In this experiment, comparisons were made to evaluate the base classifiers with the ensemble model in term of execution time. Figure 4.23 and Figure 4.24 present the execution time of the individual classifiers and the ensemble model.

The results show that: 1) The execution time of the training stage is higher than that of the prediction stage for all models including the base classifiers and the ensemble model; 2) All models present an approximately linear execution time in both the training and prediction stages under different feature sets, ensuring the necessary efficiency for producing fast recommendations for patients; 3) Of all models, including the base classifiers and the ensemble model, the ensemble model takes more time than the individual base classifier to complete the training and prediction. This is because the ensemble model needs to aggregate the results from the base classifiers to generate the weights for them and then produce the final recommendation. The ensemble model sacrifices a little on the execution time to achieve better recommendation effectiveness for patients. Additionally, the training stage can be performed offline so it does not adversely affect the efficiency in generating recommendations for patients during the prediction stage.

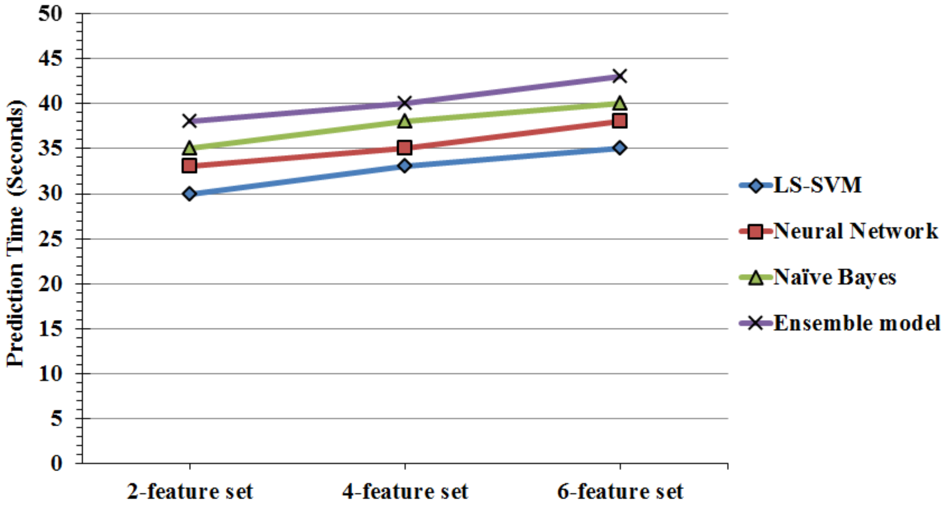


Figure 4.23: Comparison of the training time between the base classifiers and the ensemble model

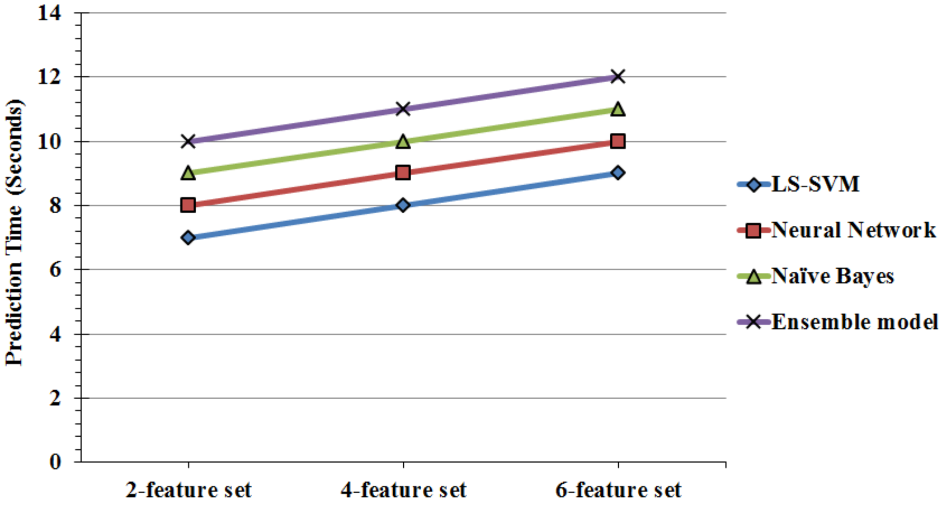


Figure 4.24: Comparison of the prediction time between the classifiers and the ensemble model

4.5.2.4 Summary

The experimental results demonstrate that the proposed system using high-frequency sub-bands, i.e., y1, y2 and y3 yielded a better predictive performance for the patient's condition compared with the other sub-bands. The findings show that using a combination of six statistical features gives a better predictive performance compared with the other feature sets. In addition, the proposed system is more effective than the individual base classifiers used in constructing the ensemble model and outperforms the previously proposed approaches when solving the same problem. Based on the experimental evaluations, the proposed recommendation system can be used as an effective tool to improve the quality of clinical evidence-based decisions by which the time and cost associated with the daily medical test can be reduced. The experimental evaluation of this system establishes that the proposed recommendation system is effective in improving the quality of clinical evidence-based decisions and reducing the time costs incurred by chronic heart disease patients undertaking their daily medical test, thereby improving their general quality of life.

Chapter 5

Conclusions and Future Work

The main purpose of this study is to develop an intelligent recommendation system supported by a set of predictive algorithms to predict and assess the short-term risk for patients with chronic diseases. The time series data of patients were investigated in both the time domain and the frequency domain. This research has proposed and applied three novel predictive algorithms, the basic heuristic algorithm, regression-based algorithm and hybrid algorithm, and also developed three techniques including a structural graph-based method, the fast Fourier transformation, and dual-tree complex wavelet transformation, which are better suited for the times series data of chronic diseases patients than currently established methods, in providing patients in a telehealth environment with appropriate recommendations for taking a medical body test on the coming day. First, two real-life time series telehealth datasets were collected from chronic diabetes and heart disease patients. Second, the time series of data were pre-processed by converting them into smaller overlapped sliding windows based on the size of the sliding window used in the data analysis. Third, six new methods were developed for short-term disease risk prediction to provide chronic disease patients with appropriate recommendations about the need to take a medical test on the coming day based on an analysis of their medical data. Finally, user friendly interfaces were developed to supply the necessary inputs to the system and receive recommendations from the system. The experimental results show that the proposed system yields very good recommendation accuracy and offers an effective way to reduce the risk of incorrect recommendations as well as the workload for patients with chronic diseases who undergo body tests most days. The results conclusively show that the proposed system is a promising tool for analyzing time series medical data and for providing accurate and reliable recommendations to patients suffering from chronic diseases.

5.1 Contributions

The significant achievements and contributions of this thesis are presented in Chapter 5 which introduces three innovative predictive algorithms for short-term disease risk prediction. In addition, three techniques have been developed to pre-process the time series data of patients with chronic diseases. Details of the research contributions for each method are summarized as follows:

In the First section of Chapter 5, the time domains analysis algorithms for short-time disease risk prediction are presented. First, three innovative predictive algorithms are proposed for the short-term risk assessment of patients in a telehealth environment based on the analysis of a patient's historical medical data. This work is considered one of the first studies to use the medical measurements of patients in the prediction of short-term disease risk in the telehealth environment. The system aims to improve the quality of clinical evidence-based decisions and reduce the financial and timing cost taken by patients. The system is based on a time series prediction algorithm which predicts short-term disease risk for patients and facilitates the generation of recommendations for whether or not a particular patient needs to take the physical body test on the current date using the telehealth facility. In addition, this system is believed to contribute to the better monitoring, assessment and management of disease risk for patients suffering chronic diseases such as heart disease. The work makes a theoretical contribution with the time series prediction algorithm and applied contribution through an intelligent system to improve the quality of health care services. The results have been published in one conference paper and one journal paper:

- Lafta, R., Zhang, J., Tao, X., Li, Y. and Tseng, V. S. (2015), "An intelligent recommender system based on short-term risk prediction for heart disease patients.", *Proceeding of the Web IEEE/WIC/ACM International Conference on Intelligence and Intelligent Agent Technology (WI-IAT) in Singapore*, pp. 102-105.
- Lafta, R., Zhang, J., Tao, X., Li, Y., Tseng, V. S., Luo, Y. and Chen, F. (2016), "An intelligent recommender system based on predictive analysis in telehealthcare environment.", *Web Intelligence*, Vol. 14, *IOS Press*, pp. 325-336.

Next, the system was further developed using a structural graph with a machine learning ensemble model to provide patients in a telehealth environment with appropriate recommendations. This study applies the the structural graph, which effectively represents the medical time series data and enters the extracted statistical features into the ensemble model to generate accurate, reliable recommendations for chronic heart disease patients. Three popular and capable classifiers, Least Square-Support Vector Machine, Neural Network, and Naive Bayes, were used to construct the ensemble framework.

The experimental results showed that the proposed system using sliding windows of 15 days with the optimal statistical features set produced by the structural graph properties yielded a better predictive performance for all measurements. The results also showed that the system using the ensemble classifier with optimal features set can correctly predict up to 94% of the subjects across all measurements. The system is also more effective than the individual base classifiers used in the ensemble model and outperforms the previously proposed approaches developed to solve the same problem. The evaluation establishes that the recommendation system is effective in improving the quality of clinical evidence-based decisions and reduces the time costs incurred by patients undertaking their daily medical test, thereby improving their overall life quality.

Finally, the frequency domains analysis techniques for short-time disease risk prediction were applied. In this section, two decomposition methods (fast Fourier transformation and dual tree complex wavelet transformation) were proposed to extract frequency information to predict the patient's condition. In the fast Fourier transformation method, the system's recommendation accuracy and risk when using eight statistical features, was approximately 2.4% and 42.9% better than using six statistical features across different medical measurements while maintaining roughly the same level of workload saving at the same time. Therefore, this method achieved a better performance using all eight statistical features for all 28 frequency bands for generating recommendations for patients. The system achieves a very high accuracy and workload saving as well as a very low risk. For instance, when using 28 frequency bands and eight statistical features, the system achieves an accuracy over 94% and a workload saving over 63% with a risk is lower than 2.6%, indicating that the recommendation system is highly accurate and able to significantly reduce the workload for chronic heart disease patients undertaking their daily medical tests with a very low health risk. The obtained results have been published in one conference paper and one journal paper:

- Lafta, R., Zhang, J., Tao, X., Li, Y., Abbas, W., Luo, Y., Chen, F. and

Tseng, V. S. (2017), “A fast fourier transform-coupled machine learning-based ensemble model for disease risk prediction using a real-life dataset.”, *Proceeding of the Pacic-Asia Conference on Knowledge Discovery and Data Mining, South Korea*, pp. 654-670.

- Zhang, J., Lafta, R., Tao, X., Li, Y., Zhu, X., Luo, Y. and Chen, F. (2017), “Coupling a fast fourier transformation with a machine learning ensemble model to support recommendations for heart disease patients in a telehealth environment.”, *IEEE Access*, Vol.5, pp. 10674 - 10685.

The second decomposition method is a dual-tree complex wavelet transformation. This system was developed using the the dual-tree complex wavelet with a machine learning ensemble model to provide patients in a telehealth environment with appropriate recommendations for the need to take a medical body test on the coming day. Such recommendations are established based on the prediction of patients' health conditions using their time series medical data from the past few days. The experimental results showed that the proposed system using high-frequency sub-bands, y1, y2 and y3, yielded a better predictive performance for the patient's condition compared with the other sub-bands. The findings showed that using a combination of six statistical features gives a better predictive performance compared with the other feature sets. In addition, the proposed system is more effective than the individual base classifiers used to construct the ensemble model and outperforms the previously proposed approaches to solve the same problem. Based on the experimental evaluations, the proposed recommendation system can be used as an effective tool to improve the quality of clinical evidence-based decisions by which the time and cost associated with the daily medical test can be reduced. The experimental evaluation of this system establishes that the proposed recommendation system is effective in improving the quality of clinical evidence-based decisions and reduces the time costs incurred by patients undertaking their daily medical test, thereby improving their overall quality of life.

In summary, the main contributions of this thesis are the intelligent recommender system which is supported by a set of innovative algorithms and techniques that predict and assess the short-term risk for patients with chronic diseases and then provide the appropriate recommendation related to the need to take a medical test body on the coming day. The proposed system is a promising tool for analyzing time series medical data and providing accurate and reliable recommendations to patients suffering from chronic diseases in a telehealth environment.

5.2 Future Research Directions

There are several directions for future work from this thesis. The first direction is to evaluate the proposed system using additional appropriate datasets which preferably have a large number of data records. The generated recommendations would be more accurate and reliable when dealing with a such large dataset.

We are also interested in applying other ensemble techniques, such as boosting and Adaboost, to produce recommendations and conduct a comparative study of those different ensemble models. These machine learning ensembles could be very successful in improving the predictive efficiency of certain classifiers for artificial and real-world datasets.

In the future work, we are interested in using the Multi Criteria Decision Making (MCDM) method to select a combination of classifiers for building the machine learning ensembles as selecting an inappropriate classier in building the ensemble could affect the classification accuracy negatively. Applying the MCDM will help in selecting the best classifiers from the available machine learning algorithms.

In addition, to increase the predictive ability of the proposed system, we will find relationships among different measurements used in the system such as the relationship between heart rate and blood pressure. This practice will help the proposed system in decision making based on multi measurements instead of a single measurement.

Furthermore, we are interested to apply a new metric, such as impact of wrong recommendations, that may improve the performance of the proposed system.

Finally, given the generality of the proposed model in dealing with medical time series data, we will explore the possibility of applying the system to support telehealth care for patients suffering from other types of diseases.

References

Ahmed, K., Jesmin, T. and Rahman, M. Z. (2013), 'Early prevention and detection of skin cancer risk using data mining', *International Journal of Computer Applications* **62**(4).

Ali, S. M., Giordano, R., Lakhani, S. and Walker, D. M. (2016), 'A review of randomized controlled trials of medical record powered clinical decision support system to improve quality of diabetes care', *International journal of medical informatics* **87**, 91–100.

Alwan, A., MacLean, D. R., Riley, L. M., d'Espaignet, E. T., Mathers, C. D., Stevens, G. A. and Bettcher, D. (2010), 'Monitoring and surveillance of chronic non-communicable diseases: progress and capacity in high-burden countries', *The Lancet* **376**(9755), 1861–1868.

Aoki, T., Iijima, H., Tada, T., Kumada, T., Nishimura, T., Nakano, C., Kishino, K., Shimono, Y., Yoh, K., Takata, R. et al. (2017), 'Prediction of development of hepatocellular carcinoma using a new scoring system involving virtual touch quantification in patients with chronic liver diseases', *Journal of gastroenterology* **52**(1), 104–112.

Artameeyanant, P., Sultornsanee, S. and Chamnongthai, K. (2015), Classification of electromyogram using weight visibility algorithm with multilayer perceptron neural network, *in* 'Proceedings of 7th International Conference on Knowledge and Smart Technology (KST)', IEEE, pp. 190–194.

Bai, Y., Han, X., Chen, T. and Yu, H. (2015), 'Quadratic kernel-free least squares support vector machine for target diseases classification', *Journal of Combinatorial Optimization* **30**(4), 850–870.

Bashir, S., Qamar, U. and Khan, F. H. (2015), 'Bagmoov: A novel ensemble for heart disease prediction bootstrap aggregation with multi-objective optimized voting', *Australasian physical & engineering sciences in medicine* **38**(2), 305–323.

Beaney, K. E., Cooper, J. A., Shahid, S. U., Ahmed, W., Qamar, R., Drenos, F., Crockard, M. A. and Humphries, S. E. (2015), 'Clinical utility of a coronary heart disease risk prediction gene score in uk healthy middle aged men and in the pakistani population', *PLoS One* **10**(7), e0130754.

Bernhardt, B. C., Bonilha, L. and Gross, D. W. (2015), 'Network analysis for a network disorder: the emerging role of graph theory in the study of epilepsy', *Epilepsy & Behavior* **50**, 162–170.

Blondel, V. D., Gajardo, A., Heymans, M., Senellart, P. and Van Dooren, P. (2004), 'A measure of similarity between graph vertices: Applications to synonym extraction and web searching', *SIAM review* **46**(4), 647–666.

Boateng, D., Agyemang, C., Beune, E., Meeks, K., Smeeth, L., Schulze, M. B., Addo, J., Aikins, A. d.-G., Galbete, C., Bahendeka, S. et al. (2018), 'Cardiovascular disease risk prediction in sub-saharan african populationscomparative analysis of risk algorithms in the rodam study', *International Journal of Cardiology* **254**, 310–315.

Boccaletti, S., Latora, V., Moreno, Y., Chavez, M. and Hwang, D.-U. (2006), 'Complex networks: Structure and dynamics', *Physics reports* **424**(4), 175–308.

Braamse, A. M., Jean, C. Y., Visser, O. J., Heymans, M. W., van Meijel, B., Dekker, J. and Syrjala, K. L. (2016), 'Developing a risk prediction model for long-term physical and psychological functioning after hematopoietic cell transplantation', *Biology of Blood and Marrow Transplantation* **22**(3), 549–556.

Breiman, L. (1996), 'Bagging predictors', *Machine learning* **24**(2), 123–140.

Brentan, B. M., Luvizotto Jr, E., Herrera, M., Izquierdo, J. and Pérez-García, R. (2017), 'Hybrid regression model for near real-time urban water demand forecasting', *Journal of Computational and Applied Mathematics* **309**, 532–541.

burnham health centre (2006), `https://www.burnhamhealthcentre.co.uk/tests-and-results.aspx`. (Accessed on 10/07/2017).

Carbone, M., Sharp, S. J., Flack, S., Paximadas, D., Spiess, K., Adgey, C., Griffiths, L., Lim, R., Trembling, P., Williamson, K. et al. (2016), 'The uk-pbc risk scores: Derivation and validation of a scoring system for long-term prediction of end-stage liver disease in primary biliary cholangitis', *Hepatology* **63**(3), 930–950.

Chang, C.-D., Wang, C.-C. and Jiang, B. C. (2011), 'Using data mining techniques for multi-diseases prediction modeling of hypertension and hyperlipidemia by common risk factors', *Expert systems with applications* **38**(5), 5507–5513.

Chatfield, C. (2016), *The analysis of time series: an introduction*, CRC press.

Chatterjee, S. and Hadi, A. S. (2015), *Regression analysis by example*, John Wiley & Sons.

Cooley, J. W., Lewis, P. A. and Welch, P. D. (1969), 'The fast fourier transform and its applications', *IEEE Transactions on Education* **12**(1), 27–34.

Das, A. B., Bhuiyan, M. I. H. and Alam, S. S. (2016), 'Classification of eeg signals using normal inverse gaussian parameters in the dual-tree complex wavelet transform domain for seizure detection', *Signal, Image and Video Processing* **10**(2), 259–266.

Deo, R. C., Wen, X. and Qi, F. (2016), 'A wavelet-coupled support vector machine model for forecasting global incident solar radiation using limited meteorological dataset', *Applied Energy* **168**, 568–593.

Diykh, M. and Li, Y. (2016), 'Complex networks approach for eeg signal sleep stages classification', *Expert Systems with Applications* **63**, 241–248.

Diykh, M., Li, Y. and Wen, P. (2016), 'Eeg sleep stages classification based on time domain features and structural graph similarity', *IEEE Transactions on Neural Systems and Rehabilitation Engineering* **24**(11), 1159–1168.

Diykh, M., Li, Y. and Wen, P. (2017), 'Classify epileptic eeg signals using weighted complex networks based community structure detection', *Expert Systems with Applications* **90**, 87–100.

Druss, B. G., Marcus, S. C., Olfson, M., Tanielian, T., Elinson, L. and Pincus, H. A. (2001), 'Comparing the national economic burden of five chronic conditions', *Health Affairs* **20**(6), 233–241.

Famili, A., Shen, W.-M., Weber, R. and Simoudis, E. (1997), 'Data preprocessing and intelligent data analysis', *Intelligent data analysis* **1**(1-4), 3–23.

Fang, Z. and Wang, J. (2014), 'Efficient identifications of structural similarities for graphs', *Journal of Combinatorial Optimization* **27**(2), 209–220.

Fawcett, T. (2006), 'An introduction to roc analysis', *Pattern recognition letters* **27**(8), 861–874.

Gao, H., Jian, S., Peng, Y. and Liu, X. (2016), 'A subspace ensemble framework for classification with high dimensional missing data', *Multidimensional Systems and Signal Processing* pp. 1–16.

Geng, H., Lu, T., Lin, X., Liu, Y. and Yan, F. (2015), 'Prediction of protein-protein interaction sites based on naive bayes classifier', *Biochemistry research international* **2015**.

Hawkes, A. L., Patrao, T. A., Atherton, J., Ware, R. S., Taylor, C. B., ONeil, A., Foreman, R. and Oldenburg, B. F. (2013), 'Effect of a telephone-delivered coronary heart disease secondary prevention program (proactive heart) on quality of life and health behaviours: primary outcomes of a randomised controlled trial', *International journal of behavioral medicine* **20**(3), 413–424.

He, Y. and Evans, A. (2010), 'Graph theoretical modeling of brain connectivity', *Current opinion in neurology* **23**(4), 341–350.

Huang, F., Wang, S. and Chan, C.-C. (2012), Predicting disease by using data mining based on healthcare information system, *in* 'Granular Computing (GrC), 2012 IEEE International Conference on', IEEE, pp. 191–194.

Huang, X. and Lai, W. (2006), 'Clustering graphs for visualization via node similarities', *Journal of Visual Languages & Computing* **17**(3), 225–253.

Jain, A. K., Murty, M. N. and Flynn, P. J. (1999), 'Data clustering: a review', *ACM computing surveys (CSUR)* **31**(3), 264–323.

Ji, Y., Xu, P. and Ye, Y. (2015), 'Hvac terminal hourly end-use disaggregation in commercial buildings with fourier series model', *Energy and Buildings* **97**, 33–46.

Jung, C., Yang, Q., Sun, T., Fu, Q. and Song, H. (2017), 'Low light image enhancement with dual-tree complex wavelet transform', *Journal of Visual Communication and Image Representation* **42**, 28–36.

Kahn, M. (1994), *AIM-94 data set provided by Michael Kahn, MD, PhD, Washington University, St. Louis, MO,US.*

Kim, J.-K., Lee, J.-S., Park, D.-K., Lim, Y.-S., Lee, Y.-H. and Jung, E.-Y. (2014), 'Adaptive mining prediction model for content recommendation to coronary heart disease patients', *Cluster computing* **17**(3), 881–891.

Kingsbury, N. G. (1998), The dual-tree complex wavelet transform: a new technique for shift invariance and directional filters, *in* 'Proc. 8th IEEE DSP workshop', Vol. 8, Utah, p. 86.

Kiplangat, D. C., Asokan, K. and Kumar, K. S. (2016), 'Improved week-ahead predictions of wind speed using simple linear models with wavelet decomposition', *Renewable Energy* **93**, 38–44.

Kovacs, P., Samiee, K. and Gabbouj, M. (2014), On application of rational discrete short time fourier transform in epileptic seizure classification, *in* 'Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)', IEEE, pp. 5839–5843.

Krishnaiah, V., Narsimha, D. G. and Chandra, D. N. S. (2013), 'Diagnosis of lung cancer prediction system using data mining classification techniques', *International Journal of Computer Science and Information Technologies* **4**(1), 39–45.

Kuh, D. and Shlomo, Y. B. (2004), *A life course approach to chronic disease epidemiology*, number 2, Oxford University Press.

Kurosaki, M., Hiramatsu, N., Sakamoto, M., Suzuki, Y., Iwasaki, M., Tamori, A., Matsuura, K., Kakinuma, S., Sugauchi, F., Sakamoto, N. et al. (2012), 'Data mining model using simple and readily available factors could identify patients at high risk for hepatocellular carcinoma in chronic hepatitis c', *Journal of hepatology* **56**(3), 602–608.

Lafta, R., Zhang, J., Tao, X., Li, Y., Abbas, W., Luo, Y., Chen, F. and Tseng, V. S. (2017), A fast fourier transform-coupled machine learning-based ensemble model for disease risk prediction using a real-life dataset, *in* 'Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining, South Korea', pp. 654–670.

Lafta, R., Zhang, J., Tao, X., Li, Y. and Tseng, V. S. (2015), An intelligent recommender system based on short-term risk prediction for heart disease patients, *in* 'Proceedings of International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), IEEE/WIC/ACM, Singapore', pp. 102–105.

Lafta, R., Zhang, J., Tao, X., Li, Y. and Tseng, V. S. (2016), Irs-hd: an intelligent personalized recommender system for heart disease patients in a tele-health environment, *in* 'Proceedings of 12th International Conference on Advanced Data Mining and Applications (ADMA), Gold Coast, QLD, Australia', pp. 803–806.

Lafta, R., Zhang, J., Tao, X., Li, Y., Tseng, V. S., Luo, Y. and Chen, F. (2016), An intelligent recommender system based on predictive analysis in telehealthcare environment, *in* 'Web Intelligence', Vol. 14, IOS Press, pp. 325–336.

Lang, S. (2017), 'Cognitive eloquence in neurosurgery: Insight from graph theoretical analysis of complex brain networks', *Medical hypotheses* **98**, 49–56.

Lee, M.-H., Yang, H.-I., Liu, J., Batrla-Utermann, R., Jen, C.-L., Iloeje, U. H., Lu, S.-N., You, S.-L., Wang, L.-Y. and Chen, C.-J. (2013), 'Prediction models of long-term cirrhosis and hepatocellular carcinoma risk in chronic hepatitis b patients: Risk scores integrating host and virus profiles', *Hepatology* **58**(2), 546–554.

Li, M., Chen, W. and Zhang, T. (2017), 'Automatic epileptic eeg detection using dt-cwt-based non-linear features', *Biomedical Signal Processing and Control* **34**, 114–125.

Li, S., Tang, B. and He, H. (2016), 'An imbalanced learning based mdr-tb early warning system', *Journal of medical systems* **40**(7), 164.

Li, X., Hu, X., Jin, C., Han, J., Liu, T., Guo, L., Hao, W. and Li, L. (2013), 'A comparative study of theoretical graph models for characterizing structural networks of human brain', *International journal of biomedical imaging* **2013**.

Li, Y., Wen, P. P. et al. (2011), 'Clustering technique-based least square support vector machine for eeg signal classification', *Computer methods and programs in biomedicine* **104**(3), 358–372.

Lu, H.-Y., Huang, C.-Y., Su, C.-T. and Lin, C.-C. (2014), 'Predicting rotator cuff tears using data mining and bayesian likelihood ratios', *PloS one* **9**(4), e94917.

Lu, J.-F., Chi, M.-J. and Chen, C.-M. (2014), 'Advocacy of home telehealth care among consumers with chronic conditions', *Journal of clinical nursing* **23**(5-6), 811–819.

Machac, J., Weiss, A., Winters, S. L., Barecca, P. and Gomes, J. A. (1988), 'A comparative study of frequency domain and time domain analysis of signal-averaged electrocardiograms in patients with ventricular tachycardia', *Journal of the American College of Cardiology* **11**(2), 284–296.

Mertins, A. (1999), 'Signal analysis: Wavelets, filter banks, time-frequency transforms and applications, english (revised edition)'.

Micheloyannis, S., Pachou, E., Stam, C. J., Vourkas, M., Erimaki, S. and Tsirka, V. (2006), 'Using graph theoretical analysis of multi channel eeg to evaluate the neural efficiency hypothesis', *Neuroscience letters* **402**(3), 273–277.

Miraglia, F., Vecchio, F. and Rossini, P. M. (2017), 'Searching for signs of aging and dementia in eeg through network analysis', *Behavioural brain research* **317**, 292–300.

Moghaddass, R., Rudin, C. and Madigan, D. (2016), 'The factorized self-controlled case series method: An approach for estimating the effects of many drugs on many outcomes', *Journal of Machine Learning Research* **17**(185), 1–24.

Mohktar, M. S., Redmond, S. J., Antoniades, N. C., Rochford, P. D., Pretto, J. J., Basilakis, J., Lovell, N. H. and McDonald, C. F. (2015), 'Predicting the risk of exacerbation in patients with chronic obstructive pulmonary disease using home telehealth measurement data', *Artificial intelligence in medicine* **63**(1), 51–59.

Nihtyanova, S. I., Schreiber, B. E., Ong, V. H., Rosenberg, D., Moinzadeh, P., Coghlan, J. G., Wells, A. U. and Denton, C. P. (2014), 'Prediction of pulmonary complications and long-term survival in systemic sclerosis', *Arthritis & rheumatology* **66**(6), 1625–1635.

Nilashi, M., bin Ibrahim, O., Ahmadi, H. and Shahmoradi, L. (2017), 'An analytical method for diseases prediction using machine learning techniques', *Computers & Chemical Engineering* .

Njie, G. J., Proia, K. K., Thota, A. B., Finnie, R. K., Hopkins, D. P., Banks, S. M., Callahan, D. B., Pronk, N. P., Rask, K. J. and Lackland, D. T. (2015), 'Clinical decision support systems and prevention: a community guide cardiovascular disease systematic review', *American journal of preventive medicine* **49**(5), 784–795.

Nobel, L., Mayo, N. E., Hanley, J., Nadeau, L. and Daskalopoulou, S. S. (2014), 'Myrisk_stroke calculator: A personalized stroke risk assessment tool for the general population', *Journal of Clinical Neurology* **10**(1), 1–9.

Odan, F. K. and Reis, L. F. R. (2012), 'Hybrid water demand forecasting model associating artificial neural network with fourier series', *Journal of Water Resources Planning and Management* **138**(3), 245–256.

O'Neil, A., Taylor, B., Sanderson, K., Cyril, S., Chan, B., Hawkes, A. L., Hare, D. L., Jelinek, M., Venugopal, K. and Atherton, J. J. (2014), 'Efficacy and feasibility of a tele-health intervention for acute coronary syndrome patients with depression: Results of the moodcare randomized controlled trial', *Annals of Behavioral Medicine* **48**(2), 163–174.

Panzica, F., Varotto, G., Rotondi, F., Spreafico, R. and Franceschetti, S. (2013), 'Identification of the epileptogenic zone from stereo-eeg signals: a connectivity-graph theory approach', *Frontiers in neurology* **4**.

Payne, B. A., Hutcheon, J. A., Ansermino, J. M., Hall, D. R., Bhutta, Z. A., Bhutta, S. Z., Biryabarema, C., Grobman, W. A., Groen, H., Haniff, F. et al. (2014), 'A risk prediction model for the assessment and triage of women with hypertensive disorders of pregnancy in low-resourced settings: the minipiers (pre-eclampsia integrated estimate of risk) multi-country prospective cohort study', *PLoS medicine* **11**(1), e1001589.

Polat, K. and Güneş, S. (2007), 'Breast cancer diagnosis using least square support vector machine', *Digital Signal Processing* **17**(4), 694–701.

Reddy, B. K., Delen, D. and Agrawal, R. K. (2018), 'Predicting and explaining inflammation in crohns disease patients using predictive analytics methods and electronic medical record data', *Health informatics journal* p. 1460458217751015.

Sabibullah, M., Shanmugasundaram, V. and Priya, R. (2013), 'Diabetes patients risk through soft computing model', *International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)* **2**(6), 60–65.

Samiee, K., Kovacs, P. and Gabbouj, M. (2015), 'Epileptic seizure classification of eeg time-series using rational discrete short-time fourier transform', *IEEE transactions on Biomedical Engineering* **62**(2), 541–552.

Samuel, O. W., Asogbon, G. M., Sangaiah, A. K., Fang, P. and Li, G. (2017), 'An integrated decision support system based on ann and fuzzy_ahp for heart failure risk prediction', *Expert Systems with Applications* **68**, 163–172.

Sánchez, A. S., Iglesias-Rodríguez, F. J., Fernández, P. R. and de Cos Juez, F. (2016), 'Applying the k-nearest neighbor technique to the classification of workers according to their risk of suffering musculoskeletal disorders', *International Journal of Industrial Ergonomics* **52**, 92–99.

Sarsoh, J. T., Hashem, K. M. et al. (2012), 'Classifying of human face images based on the graph theory concepts', *Global Journal of Computer Science and Technology* **12**(13-F).

Şen, B., Peker, M., Çavuşoğlu, A. and Çelebi, F. V. (2014), 'A comparative study on classification of sleep stage based on eeg signals using feature selection and classification algorithms', *Journal of medical systems* **38**(3), 18.

Sepalika (2018), 'Heart rate variability measurement: The secret test for impending heart attacks?', `https://www.sepalika.com/heart/heart-rate-variability`. (Accessed on 10/03/2018).

Sharawardi, N. A., Choo, Y.-H., Chong, S.-H., Muda, A. K. and Goh, O. S. (2014), Single channel semg muscle fatigue prediction: An implementation using least square support vector machine, *in* 'Proceedings of Fourth World Congress on Information and Communication Technologies (WICT)', pp. 320–325.

Singh, Y. N. and Gupta, P. (2007), Quantitative evaluation of normalization techniques of matching scores in multimodal biometric systems, *in* 'Proceedings of International Conference on Biometrics, Berlin, Heidelberg', pp. 574–583.

Stam, C. J. and Reijneveld, J. C. (2007), 'Graph theoretical analysis of complex networks in the brain', *Nonlinear biomedical physics* **1**(1), 3.

Suykens, J. A. and Vandewalle, J. (1999), 'Least squares support vector machine classifiers', *Neural processing letters* **9**(3), 293–300.

Swami, P., Gandhi, T. K., Panigrahi, B. K., Bhatia, M., Santhosh, J. and Anand, S. (2017), 'A comparative account of modelling seizure detection system using wavelet techniques', *International Journal of Systems Science: Operations & Logistics* **4**(1), 41–52.

Tayefi, M., Tajfard, M., Saffar, S., Hanachi, P., Amirabadizadeh, A. R., Esmaeily, H., Taghipour, A., Ferns, G. A., Moohebati, M. and Ghayour-Mobarhan, M. (2017), 'hs-crp is strongly associated with coronary heart disease (chd): A data mining approach using decision tree algorithm', *Computer methods and programs in biomedicine* **141**, 105–109.

Thong, N. T. and Le, H. S. (2015), 'Hifcf: An effective hybrid model between picture fuzzy clustering and intuitionistic fuzzy recommender systems for medical diagnosis', *Expert Systems with Applications* **42**(7), 3682–3701.

Tsai, C.-L., Chen, W. T. and Chang, C.-S. (2016), 'Polynomial-fourier series model for analyzing and predicting electricity consumption in buildings', *Energy and Buildings* **127**, 301–312.

Valentini, G. and Masulli, F. (2002), Ensembles of learning machines, *in* 'Italian Workshop on Neural Nets', Springer, pp. 3–20.

Van Ma, L., Kim, J., Park, S., Kim, J. and Jang, J. (2016), 'An efficient session_weight load balancing and scheduling methodology for high-quality telehealth care service based on webrtc', *The Journal of Supercomputing* **72**(10), 3909–3926.

Vecchio, F., Miraglia, F., Piludu, F., Granata, G., Romanello, R., Caulo, M., Onofrj, V., Bramanti, P., Colosimo, C. and Rossini, P. M. (2017), 'small world architecture in brain connectivity and hippocampal volume in alzheimers disease: a study via graph theory from eeg data', *Brain imaging and behavior* **11**(2), 473–485.

Voyant, C., Notton, G., Kalogirou, S., Nivet, M.-L., Paoli, C., Motte, F. and Fouilloy, A. (2017), ‘Machine learning methods for solar radiation forecasting: A review’, *Renewable Energy* **105**, 569–582.

Vural, C. and Yildiz, M. (2010), ‘Determination of sleep stage separation ability of features extracted from eeg signals using principle component analysis’, *Journal of medical systems* **34**(1), 83–89.

Wang, J., Qiu, S., Xu, Y., Liu, Z., Wen, X., Hu, X., Zhang, R., Li, M., Wang, W. and Huang, R. (2014), ‘Graph theoretical analysis reveals disrupted topological properties of whole brain functional networks in temporal lobe epilepsy’, *Clinical Neurophysiology* **125**(9), 1744–1756.

Yang, J.-G., Kim, J.-K., Kang, U.-G. and Lee, Y.-H. (2014), ‘Coronary heart disease optimization system on adaptive-network-based fuzzy inference system and linear discriminant analysis (anfis–lda)’, *Personal and ubiquitous computing* **18**(6), 1351–1362.

Yao, X. (1999), ‘Evolving artificial neural networks’, *Proceedings of the IEEE* **87**(9), 1423–1447.

Ye, C., Fu, T., Hao, S., Zhang, Y., Wang, O., Jin, B., Xia, M., Liu, M., Zhou, X., Wu, Q. et al. (2018), ‘Prediction of incident hypertension within the next year: Prospective study using statewide electronic health records and machine learning’, *Journal of medical Internet research* **20**(1), e22.

Yeh, D.-Y., Cheng, C.-H. and Chen, Y.-W. (2011), ‘A predictive model for cerebrovascular disease using data mining’, *Expert Systems with Applications* **38**(7), 8970–8977.

Yu, B., Jia, B., Ding, L., Cai, Z., Wu, Q., Law, R., Huang, J., Song, L. and Fu, S. (2016), ‘Hybrid dual-tree complex wavelet transform and support vector machine for digital multi-focus image fusion’, *Neurocomputing* **182**, 1–9.

Zhang, J., Lafta, R., Tao, X., Li, Y., Zhu, X., Luo, Y. and Chen, F. (2017), ‘Coupling a fast fourier transformation with a machine learning ensemble model to support recommendations for heart disease patients in a telehealth environment’, *IEEE Access* pp. 10674–10685.