



PERSONALISED BLOOD GLUCOSE LEVEL PREDICTION MODEL USING NON-CGM DATA

TAVA MANGGAI MUTHUSAMY

A thesis submitted in fulfilment of the
requirements for the award of the degree of
MASTER OF SCIENCE IN DATA SCIENCE & BUSINESS ANALYTICS

ASIA PACIFIC UNIVERSITY OF TECHNOLOGY & INNOVATION (APU)

AUGUST 2019

DECLARATION OF THESIS CONFIDENTIALITY

Author's full name: **TAVA MANGGAI MUTHUSAMY**

IC No./Passport No.: **820113-14-5008**

Thesis/Project title: **PERSONALISED BLOOD GLUCOSE LEVEL PREDICTION MODEL USING NON-CGM DATA**

I declare that this thesis is classified as:

- ☐ CONFIDENTIAL
☐ RESTRICTED
☒ OPEN ACCESS

I acknowledged that Asia Pacific University of Technology & Innovation (APU) reserves the right as follows:

1. The thesis is the property of Asia Pacific University of Technology & Innovation (APU).
2. The Library of Asia Pacific University of Technology & Innovation (APU) has the right to make copies for the purpose of research only.
3. The Library has the right to make copies of the thesis for academic exchange.

Author's Signature:

Date: 23 August 2019

Supervisor's Name: **DR. MANOJ JAYABALAN**

Date: 23 August 2019

Signature:

DECLARATION OF SUPERVISOR(S)

“We hereby declare that We have read this thesis and in our
opinion this thesis is sufficient in terms of scope and quality for the award
of the degree of
Master of Science in Software Engineering”

Name of Supervisor: **DR. MANOJ JAYABALAN**

Signature:

Date: 23 August 2019

Name of Supervisor (II) **PROF. DR. IR. VINESH THIRUCHELVAM**

Signature:

Date: 23 August 2019

DECLARATION OF ORIGINALITY AND EXCLUSIVENESS

I declare that this thesis entitled
PERSONALISED BLOOD GLUCOSE LEVEL PREDICTION MODEL USING NON-
CGM DATA
is the result of my own research work except as cited in the references.
This thesis has not been accepted for any degree and it is not concurrently
submitted in candidature of any other degree.

Signature:

Name: Tava Manggai Muthusamy

Date: 23 August 2019

ACKNOWLEDGMENT

The completion of this project has given me much pleasure and great amount of knowledge. Firstly, I would like to extend my deepest gratitude to my supervisor, Dr. Manoj Jayabalan for his continuous support, encouragement and trust. His contribution in providing guidance and encouragement throughout the year, greatly helped me to complete this project exceptionally well. A special thanks also goes to my second supervisor, Prof. Dr. Ir. Vinesh Thiruchelvam for his valuable insights and motivation especially during the project presentations.

I would also like to thank all my lecturers from Asia Pacific University of Technology and Innovation (APU) for sharing their knowledge and equipping me with the right skills and abilities to complete this project successfully. Last but not least, I am grateful for my classmates and friends who continuously provided valuable suggestions and moral support throughout this project.

ABSTRACT

With the increase in the prevalence of diabetes and complications related to the disease, it has become imperative for every diabetic patient to self-monitor the progress of blood glucose levels closely. However, due to the impracticability of continuously monitoring blood glucose levels in a free-living condition, blood glucose monitoring is often neglected. Thus, having an effective blood glucose prediction model would help patients improve self-management of diabetes and allow patients to take preventive actions in case of any drastic changes in blood glucose levels. There are many researches on blood glucose prediction models that have been developed and tested using CGM data. Nevertheless, only a few research studies have utilized non-CGM data for the purpose. Furthermore, these researches also relied on a large number of variables i.e. meal, physical activity, insulin dosage, sleep and etc to derive reliable predictions. However, in real life application, it is impractical and inconvenient for the patient, who will be obliged to record these information continuously. Hence, this research intended to develop a blood glucose prediction model using only non-CGM data. As non-CGM data are self-collected, the data obtained were in non-continuous form and contain missing values at certain times of the day making the prediction task to be even more challenging. To complement this issue of sparse or missing data, a deep learning model was introduced in this study. The LSTM model developed in this study proved its capability to handle the long-term temporal dependencies in time series data. They were able to utilize the patterns of missing data to achieve reliable prediction results. The LSTM model delivered an overall RMSE values of 2.87 and 2.57 for 30 minutes and 60 minutes prediction respectively. This significantly lower RMSE value compared to past studies proved that besides providing more reliable prediction accuracy, the model developed in this study is less complex and can be easily implemented as it only requires one type of input data which is the self-monitored blood glucose records. This study proves that this prediction model can serve as a cost effective and a convenient solution for the majority of diabetics patients around the world who do not use a CGM device for monitoring of blood glucose levels.

TABLE OF CONTENTS

DECLARATION OF THESIS CONFIDENTIALITY	ii
DECLARATION OF SUPERVISOR(S)	iii
DECLARATION OF ORIGINALITY AND EXCLUSIVENESS	iv
ACKNOWLEDGMENT	v
ABSTRACT	vi
TABLE OF CONTENTS	vii
LIST OF TABLES	x
LIST OF FIGURES	xi
CHAPTER 1: INTRODUCTION	1
1.1 Background	1
1.2 Problem Statement	3
1.3 Aim of the Study	4
1.4 Objectives of the Study	4
1.5 Research Questions	4
1.6 Significance of the Study	4
1.6 Scope of the Study	5
1.7 Structure of the Report	5
CHAPTER 2: LITERATURE REVIEW	6
2.1 Introduction	6
2.2 Method	6
2.3 Blood Glucose Level Prediction	6
2.3.1 Types of blood glucose prediction models	7

2.3.2	Types of data used in blood glucose prediction models	11
2.3.3	Comparison of Techniques	13
2.3.4	Performance Evaluation	13
2.4	Summary	13
CHAPTER 3: RESEARCH METHODOLOGY		18
3.1	Introduction	18
3.2	Research Approach	18
3.2.1	Data Collection	18
3.2.2	Description of Data Set	20
3.2.3	Data Pre-processing	20
3.2.4	Exploratory Data Analysis	21
3.2.5	Model Development	21
3.2.6	Model Evaluation	24
3.3	Summary	24
CHAPTER 4: IMPLEMENTATION		25
4.1	Introduction	25
4.2	Data Pre-Processing	25
4.2.1	Dataset Description	25
4.2.2	Variables Selection	26
4.2.3	Data Transformation	27
4.2.4	Data Exploration	27
4.2.5	Treatment of Outliers	30
4.2.6	Data Standardization	32
4.2.7	Treatment of Missing Values	34
4.3	Data Preparation	35
4.3.1	Transform Timeseries to Supervised Learning	35
4.3.2	Transform Timeseries to Scale	36
4.4	Data Partitioning	36

4.5	Model Development	37
4.6	Summary	38
CHAPTER 5: RESULTS AND ANALYSIS		39
5.1	Introduction	39
5.2	Personalised Model's Prediction Result and Analysis	39
5.2.1	Results of Model P01	39
5.2.2	Results of Model P02	40
5.2.3	Results of Model P03	41
5.2.4	Results of Model P04	41
5.2.5	Results of Model P06	42
5.2.6	Results of Model P07	43
5.3	Overall Prediction Result and Analysis	44
5.4	Result Comparison with Previous Related Works	47
5.5	Summary	48
CHAPTER 6: DISCUSSION AND CONCLUSIONS		49
6.1	Introduction	49
6.2	Discussion and Conclusions	49
6.3	Importance and Contributions of the Study	51
6.4	Future Recommendations	52
REFERENCES		53
APPENDIX A: ETHICAL APPROVAL OF RESEARCH PROJECT		56
APPENDIX B: TURNITIN SIMILARITY REPORT		63
APPENDIX C: LOG SHEETS FOR SUPERVISORY SESSION		65

LIST OF TABLES

Table 2.1: Summary of Studies on Blood Glucose Prediction Models	15
Table 3.1: Description of Variables in Data Set	20
Table 4.1: Description of Variables in Dataset	26
Table 4.2: Summary of dataset for each selected subject	26
Table 4.3: Visual Output of Blood Glucose Records for Each Research Subject	27
Table 4.4: Boxplot of Patient's Blood Glucose Records	31
Table 4.5: Summary of Missing Values Before and After Group Mean Imputation	35
Table 4.6: Summary of total observations in training and testing datasets	36
Table 5.1: RMSE Result for Prediction Model P01	40
Table 5.2: RMSE Result for Prediction Model P02	40
Table 5.3: RMSE Result for Prediction Model P03	41
Table 5.4: RMSE Result for Prediction Model P04	42
Table 5.5: RMSE Result for Prediction Model P06	42
Table 5.6: RMSE Result for Prediction Model P07	43
Table 5.7: Overall RMSE Result	44
Table 5.8: Comparison of Previous Related Works with This Study	47

LIST OF FIGURES

Figure 3.1: Framework for the Blood Glucose Prediction Model Development	19
Figure 3.2: Information Flow in RNN	22
Figure 3.3: LSTM Loop Unrolled	23
Figure 3.4: Internal Network of LSTM	23
Figure 4.1: Average Blood Glucose Level of Research Subjects by Hour	30
Figure 4.2: Example of dataset after the standardization of timing to 30 minutes interval	33
Figure 4.3: Example of dataset after the standardization of timing to 60 minutes interval	33
Figure 5.1: Prediction Result Plot for P01	40
Figure 5.2: Prediction Result Plot for P02	40
Figure 5.3: Prediction Result Plot for P03	41
Figure 5.4: Prediction Result Plot for P04	42
Figure 5.5: Prediction Result Plot for P06	43
Figure 5.6: Snippet of Prediction Result for P07	43
Figure 5.7: Prediction Result Plot for P07	44
Figure 5.8: Total Number of BG Records by Individual Research Subjects	45
Figure 5.9: Number of BG Records by Hour for Individual Research Subject	46

CHAPTER 1

INTRODUCTION

1.1 Background

Diabetes Mellitus is a serious and complex disease that is gaining prevalence globally. There is a rapid increase in diabetes among the adult population in recent years due to the effects of modern lifestyle and dietary intake. The latest report from the International Diabetes Federation (IDF) in 2017 indicate 425 million people are known to be living with diabetes currently and these figures are projected to increase to 629 million by the year 2045.

There are two main types of diabetes, known as type 1 and type 2. Type 1 diabetes is diagnosed as an autoimmune disease. This happens when the body's own defence system, attacks the cells in the pancreas that produces insulin. Type 2 diabetes occurs when the body is resistant to the insulin produced or is unable to produce enough insulin to maintain a normal blood glucose level. Among the two, type 2 diabetes is the more common form of diabetes and accounts for at least 90 percent of all diabetes cases (*IDF Diabetes Atlas*, 2017).

Diabetes is diagnosed by an elevated blood sugar level of more than 126 milligrams per decilitre (mg/dl) or 7.0 millimoles per litre (mmol/L) when fasting; or more than 200 mg/dl or 11.1mmol/L at any time during the day (Diabetes, 2012). Hyperglycaemia is a condition when a person's blood glucose level becomes too high. This leads to long-term complication such as diabetic ketoacidosis, a serious condition that leads to severe illness or death. Blood glucose levels that are too low, are known as hypoglycaemia. This condition causes severe symptoms such as body weakness, confusion, dizziness, excessive sweating, shaking, and if not treated in time, will lead to seizures, coma or death.

With the increase in the prevalence of diabetes, it has become imperative for every diabetic patient to monitor the progress of glucose levels closely. Stringent self-management of diabetes often helps in lessening the impact of this critical disease. Recommended practices for self-management of diabetes include monitoring of dietary intake, physical activity, and examinations of blood glucose

levels throughout the day. The ability to predict drastic changes in blood glucose level before they occur is imperative to prevent conditions such as hypoglycaemia and hyperglycaemia. Accurate predictions would allow plenty of time for a patient to take preventive action. Besides improving overall blood glucose levels of a patient, this would greatly help in ensuring the patient's safety.

Thus, over the years, many research works have focused on developing a robust and accurate blood glucose prediction model. Yet, the complexity of the issue and the growing prevalence of diabetes, is continuously opening doors for more sophisticated research in this field. It was apparent during the literature review that, many of the notable research work in the field, made use of Continuous Glucose Monitor (CGM) data as the main data input for their prediction model. CGM is a wearable device with a sensor under the skin to track glucose levels throughout day and night. Thus, CGM can collect data in high frequency (i.e. every 5 minutes) and able to produce huge amounts of data if collected over a period of time. Although CGM technology has the potential to revolutionize diabetes care because of the real-time feedback it provides about blood glucose levels; several challenges remain to be addressed. The high cost of the devices, limitations in approved clinical uses, and insurance coverage for the technology have limited the adoption rate of CGM devices among diabetic patients (Burge *et al.*, 2008). Hence, with the majority of diabetic patients not using CGM for self-monitoring of blood glucose levels, there is a critical need for blood glucose level prediction models that are built based on non-CGM data. Very few attempts have been made to develop blood glucose prediction model using non-CGM data. The complexities in obtaining and modelling non-CGM data suggest ample opportunities for new discoveries and improvements. Thus, this study intends to address the challenges of building blood glucose prediction model based on non-continuous or sparse data (i.e. non-CGM data).

1.2 Problem Statement

Over the years, numerous researchers have developed and tested the performance of blood glucose prediction models using CGM data and other variables such as insulin, meal consumption and physical activity as data inputs (Rollins *et al.*, 2010; Georga *et al.*, 2011; Balakrishnan *et al.*, 2013; Bunescu *et al.*, 2013; Plis *et al.*, 2014; Zecchin *et al.*, 2014; Contreras *et al.*, 2017; Hidalgo *et al.*, 2017; Fiorini *et al.*, 2017; Contreras *et al.*, 2018). As CGM is a wearable device that records blood glucose levels automatically and frequently; CGM data are easily obtained in high frequency and huge amounts serving as valuable input for the prediction models. And the other variables used, contribute more weight to these type of prediction models.

On the contrary, limited researches have utilized non-CGM data for their prediction models due to the inaccessibility of it. Non-CGM data are obtained using self-monitoring devices. Thus, it is difficult to manually collect huge amounts of data using this method. Furthermore, the data obtained are often in non-continuous form and contain missing values at certain times of the day making the prediction task to be even more challenging.

Some researchers attempted studies using non-CGM data and other similar variables by requesting patients to manually record measurements of blood glucose levels, insulin, meals, sleep, physical activity and etc on a smartphone app for a period of time. Although they were able to make reliable predictions, it is impractical for a patient to be recording all their daily activities (i.e. meals, sleep, physical activity) at frequent periods of the day, every day. Certainly, having this additional information would contribute to improving the accuracy of the prediction algorithm. However, in real-life application, it is impractical and inconvenient for the patient, who will be obliged to record this information continuously.

In light of this, this study intends to develop a prediction model using only non-CGM data (i.e. self-recorded blood glucose records) and to evaluate its performance in predicting blood glucose levels based on a 30 minutes and 60 minutes prediction horizon.

1.3 Aim of the Study

The main aim of this research is to develop a personalised blood glucose prediction model using only non-CGM data. The goal of this research is to contribute to the vast majority of diabetic patients that do not use CGM for self-monitoring of blood glucose levels.

1.4 Objectives of the Study

1. To investigate the performance of existing blood glucose prediction models developed using non-CGM data
2. To develop a personalised prediction model using only non-CGM data
3. To evaluate the performance of the proposed blood glucose prediction model

1.5 Research Questions

For the purpose of this research, the following questions will be addressed:

1. What are the main variables considered in existing blood glucose level prediction models?
2. What are the techniques and methodologies used to predict blood glucose levels using non-CGM data?
3. How a deep learning model can be designed and trained to predict blood glucose level based on only non-CGM data?
4. How the proposed prediction model can be tested and evaluated?

1.6 Significance of the Study

The findings of this research will contribute greatly to the benefit of society, considering that diabetes is currently a growing issue. This research will be beneficial to diabetic patients by providing more accurate blood glucose predictions and enabling them to monitor their blood glucose level on a more regular basis without the hassle of carrying devices. With the effective use of this model, there is a high potential for diabetic patients to avoid the consequences of being in hypoglycaemia or hyperglycaemia conditions. Moreover, this blood glucose prediction model could also assist the medical practitioners when prescribing medication and doses for diabetic patients; as the model would provide a better view of the patient's future condition. With meticulous and effective implementation, this model could be a stepping stone to revolutionize the

medical industry. For researchers, this research will help uncover new methods that can be used for non-CGM data based blood glucose prediction models that many have yet to explore. In addition, this study would be among the pioneer researches that developed and tested deep learning model for blood glucose prediction. Thus, this study would serve as a future reference for researchers embarking on studies that use deep learning models for prediction.

1.6 Scope of the Study

The scope of this research will be limited to predicting blood glucose levels based on the individual patient's historical data. The model would not be able to warn or alert patients in case of any high and low (i.e. hyperglycaemic or hypoglycaemic) situations. The research subjects will be limited to adult patients above 30 years old as the mortality rate caused by diabetes is higher among this age group (World Health Organization (WHO), 2016). Both male and female research subjects will be considered in this research and the data set will be classified based on gender.

1.7 Structure of the Report

The project report is organized into six chapters and the brief description of each chapter are as follows. The first chapter known as the introduction chapter introduces the research background in brief and describes the problem statement for this research study. The main aim, objective, scope and significance of this study are also described in this chapter. The second chapter gives a detailed review of past relevant literature related to this research topic. The types of prediction models, various techniques and datasets used in other research works for similar purpose are critically analysed. The third chapter focuses on research methodology in which the research approach of this study is highlighted. The dataset used in this study will be described and details on the data pre-processing methods are given. The techniques involved in developing the model and evaluation methods used are also discussed in this chapter. The fourth chapter discusses in detail the steps taken and the techniques applied in the model implementation phase. And the results achieved are presented in the fifth chapter. The performance of the developed models are critically analysed and insights generated from the analysis are presented in this chapter. Finally, the study is concluded in the sixth chapter, with an overall summary and discussion of the findings, achievements and contributions of this study.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

Having a blood glucose level prediction model that is able to predict a diabetic patient's current glucose level based on historical data would greatly be an advantage in cases where patients do not have immediate access to the blood glucose monitor. Furthermore, it is impractical to continuously monitor blood glucose levels using this type of device throughout the day i.e., while working, driving, traveling and etc.

Thus, many research works have focused on developing various models and new methods to predict blood glucose levels more accurately. These studies have highlighted the tremendous potential of this research field. In this chapter, several important works that are closely related to the proposed study are discussed with a focus on the types of prediction models, types of data, features and techniques used in these models.

2.2 Method

For the purpose of this literature review, databases namely, ScienceDirect, PubMed and Google Scholar were searched for studies published from the year 2010 to 2019 using the search term 'blood glucose prediction'. The articles' abstracts were reviewed and the articles that met the inclusion criteria were chosen for this review. The inclusion criteria is that the primary aim of the article had to be the development of blood glucose level prediction model for diabetic patients (type 1 and/or type 2). Articles were excluded if they included only review of pre-existing prediction models or if they were predicting the diabetes disease rather than the blood glucose levels.

2.3 Blood Glucose Level Prediction

Blood glucose level prediction is performed by inputting specific data into a prediction model and the model would analyse these data to study the correlation between them, understand the dynamics of blood glucose and use this information to predict a new glucose level. Some of the common

variables that are often inputted into this model are historical blood glucose levels, dietary intake, physical activity, insulin, and oral medicine dosage levels.

2.3.1 Types of blood glucose prediction models

Blood glucose level prediction models can be categorized into three main categories: physiological models, data-driven models and hybrid models (Oviedo *et al.*, 2017). The three widely used types of blood glucose prediction models are discussed as follows.

2.3.1.1 Physiological Model

Physiological models require the support of experts with wide knowledge and a thorough understanding of the subject as it considers input factors such as reaction to insulin, meal absorption rates and glucose metabolism in a person's body. Other relevant parameters may be included as deemed necessary by the experts. The study of the physiological process for regulating glucose levels requires modelling techniques known as compartmental modelling. Three compartments that are widely identified for this technique are meal absorption dynamics, insulin dynamics, and glucose dynamics (Lehmann & Deutsch, 1998; Mougiakakou *et al.*, 2005).

The performance of physiological models is often evaluated by comparing the model's prediction results with predictions from diabetes experts in parallel. In a study conducted by Bunesco *et al.* (2013), the authors used three diabetes experts to assess and label an evaluation dataset with a 30-minute and 60-minute prediction. The physiological model from the study was found to outweigh the diabetes experts' predictions. This outstanding result emphasises the need for a good blood glucose level prediction model as it is able to provide a more accurate outcome compared to medical practitioners' common knowledge. Another research by Plis *et al.* (2014), also displayed that their physiological model was able to outperform the results of other experts in similar research and was able to provide reliable predictions.

It can be observed from these two-research works that the physiological models are rather challenging. The major obstacle in this type of model is the requirement of physiological parameters as input, prior to being able to make predictions. These parameters require expert

interventions and are complicated to be identified as each diabetic patient react to the parameters differently. Hence, this model is not chosen to be experimented in this study.

2.3.1.2 Data-Driven Model

Data-driven models solely depend on external data input for making predictions without requiring any knowledge about the physiology of a person's body (Fiorini *et al.* 2017). They are the most commonly applied model for blood glucose prediction. Related works using data-driven models show that some researchers tried to supply the models with as much data, while some tried to use as little data as possible.

In an interesting research by Rollins *et al.* (2010), the authors considered 24 data inputs for their prediction model. The data input consists of three nutrient variables, 20 activity and stress-related variables, and time of the day. This made the model more challenging as each independent effect has to be considered. The outcome of this research showed evidence that it is possible, using an extensive set of data which consist of food, activity, and stress-related information, to accurately model blood glucose concentration for individuals. This research contributes greatly to future researchers in this field to understand each variable's mechanism and utilize them to improve prediction accuracy.

Meanwhile, Georga *et al.* (2011) used three compartmental data models in their study, as the data collected from patients (i.e. food, insulin) were non-uniformly sampled. They were known as the 1) meal model that focuses on carbohydrate absorption), 2) insulin model that focuses on the absorption of administered insulin and 3) exercise model that learns the impact of exercise on glucose-insulin metabolism. However, their result only showed small improvements in prediction accuracy compared to previous studies.

In the same year, Gani *et al.* (2011), elevated blood glucose research to new heights by suggesting that a universal data-driven model can be developed based on only one patient's data input. The authors believed that this input can be applied to a data-driven model which will then be able to predict glucose concentrations for other patients irrelevant of their diabetes type. The authors conducted this study using a hypothesis that the dynamics of blood glucose regulation within 30

minutes, is similar for different diabetic individuals. The result of the study showed that the prediction model does not depend on a specific individual, their diabetes type or CGM device used. This finding has a significant impact clinically and in practical terms, as it shows a possibility of a universal, individual and independent predictive model, which can significantly reduce the hassle of model development by only using one research subject as input.

While the previous authors believed in generalizing one standard model for all patients, authors Li & Fernando (2016) and Fiorini *et al.* (2017) assumed otherwise. In their research, they aimed to tackle blood glucose prediction by building personalized prediction models according to individual patient characteristics. To be able to produce a personalized blood glucose prediction models for individuals, the authors Li & Fernando (2016), utilized a population data from a pooled database to group similar patients. This data is then used in a three-stage evolution model which is a time series regression model based on personal history, a regression model of pooled panel data (PPD) and a regression model of pre-clustered personalized data. Although their prediction precision result was relatively low at 42 percent, this research demonstrated that it is able to remedy the data sparsity problem of the existing models.

In another relevant research by Nguyen & Rokicki (2018), the authors intended to test data-driven model's performance on sparse and non-continuous data by using non-CGM data in their model. Many machine learning algorithms were tested in the study, and results revealed that Random Forest and Extra Trees ensemble-based models were the most suitable models for this case, as they could account for the outliers as well as overfitting problems when data are limited. In other recent researches by Hamdi *et al.* (2018) and Ben Ali *et al.* (2018), the authors aimed to use only one data input for their prediction model. CGM data collected from 12 type 1 diabetic patients in free-living condition was the only input used in their study. Compared to past similar researches, their proposed method delivered better prediction accuracy. This is a breakthrough for blood glucose prediction research, since developing a prediction model based on only CGM data is very convenient.

It can be observed from these researches that data-driven models are very flexible and can be easily enhanced using various forecasting algorithms and techniques to improve its prediction capability.

The studies reviewed highlight the tremendous potential of the data-driven models. Some key possibilities that were noted are:

- i. a universal model can be build based on a single subject's data
- ii. a model can be personalised based on individual subject's attributes
- iii. a model can be build based on sparse or non-continuous data
- iv. a model can be build based on only one data input

Hence, this study also intends to utilize the efficacy of data-driven model to achieve its intended objective.

2.3.1.3 Hybrid Model

Some challenges noted in both physiological and data-driven models to perform independently have brought upon the third model known as the hybrid model. The hybrid model combines the strategies used in both physiological and data-driven models to enhance the overall performance and prediction accuracy (Contreras *et al.*, 2017). A hybrid model combines a few different models to make a more meaningful prediction. For example, at the data pre-processing stage, a physiological model that examines glucose and insulin dynamics may be integrated with a second model that examines meal absorption dynamics. This then can be combined with other models that may add value such as the physical activity of a patient. The outputs of these combinations are then inputted into a data-driven model to derive an improved prediction (Contreras *et al.*, 2018). There are several significant research works using hybrid models that have been carried out in the past. In a study by Balakrishnan *et al.* (2013), the authors developed a hybrid and personalized blood glucose prediction model. The hybrid structure consisted of three different classes of models. The three models were a mechanistic model to assess meal absorption dynamics; an empirical model for insulin absorption kinetics; and a transfer function model for prediction of personalized blood glucose. The validation of the prediction results showed that the hybrid models were effective in capturing the blood glucose dynamics of tested subjects.

Contreras *et al.* (2017) developed a new method that uses symbolic regression through Grammatical Evolution (GE) to assess a patient's glucose dynamics. This hybrid model using the new GE approach was supplied with information of insulin intake and glucose absorption rates. Another research by the same authors Contreras *et al.* (2018), developed a hybrid model using data

gathered by a sensor-augmented therapy and a fitness band. The model was supplied with information of insulin, physical activity and glucose absorption rates to predict blood glucose values. The results obtained from both this research works were found to be very encouraging.

The use of GE algorithm on a hybrid model was further explored in another research by Hidalgo *et al.* (2017). In this experiment, the authors proposed an enhanced version of the GE algorithm and tree-based genetic programming (GP). The enhanced GE uses an optimized grammar and the GP uses a three-compartment model for gauging carbohydrate and insulin dynamics. This experiment's result showed that the GE variant performed well for shorter prediction horizon, while the GP variant performed better for the longer prediction horizons. In overall, the performance of proposed GE and GP methods in a hybrid model was found to be better than or equal to other compared works of literature in the research.

As the hybrid models tend to be highly structural and more complex, Zecchin *et al.* (2014) suggested the use of a jump Neural Network (NN) prediction algorithm for this type of model. Jump NN was found to be flexible and could be easily constructed with fewer data inputs. Their study showed that the jump NN model was able to predict glucose concentration accurately and the results were comparable to previous researches.

2.3.2 Types of data used in blood glucose prediction models

There are many types of data that have been explored by researchers for the purpose of blood glucose level prediction. The main data would be the blood glucose records collected using CGM or self-monitoring devices. Besides those, data such as dietary intake, medication and physical activity have also been used to make a prediction.

2.3.2.1 CGM Data

Majority of the studies reviewed in this paper used CGM data for their study. CGM is able to collect data in high frequency (i.e. every 5 minutes) and able to produce huge amounts of data if collected over a period of time. Thus, this type of data serves as a valuable asset for the blood glucose prediction models and was able to deliver good prediction accuracy regardless of forecasting techniques used. Authors Hamdi *et al.* (2018) and Ben Ali *et al.* (2018) proved that reliable

prediction can be produced even if only one data input (i.e. CGM data) is supplied to the prediction model.

2.3.2.2 Non-CGM Data

Data that are obtained using self-monitoring devices are known as the non-CGM data. From this review, it can be noted that there is still a lack of prediction models that are built using non-CGM data. The challenge of obtaining self-monitored blood glucose records from patients is the main reason that hinders this type of study. Nevertheless, authors (Li & Fernando, 2016; Nguyen & Rokicki, 2018)) overcame this challenge by using smartphone applications to collect records from patients. However, it was noted from these two research works that data-driven models are often data-hungry and the performance of these models build based on sparse data suggest lots of room for improvement.

2.3.2.3 Other Variables in the Dataset

Majority of the studies in this review used the dietary intake of a patient as one of the data inputs for the model. In terms of dietary intake, some study considered only the carbohydrate consumption amount from a meal (Nguyen & Rokicki, 2018; Contreras *et al.*, 2017; Hidalgo *et al.*, 2017; Zecchin *et al.*, 2014); meanwhile, others used complete meal information (Balakrishnan *et al.*, 2013; Bunesco *et al.*, 2013; Plis *et al.*, 2014).

Information of physical activity of a patient and medication details (i.e. insulin dosage) of patients were also considered by a few studies. Besides these common variables, there was also one study that considered the stress level of a patient as data input (Rollins *et al.*, 2010). And another two studies considered total sleeping hours as a data input for their model (Fiorini *et al.*, 2017; Li & Fernando, 2016).

It can be noted that many different types of variables have been used in these studies. Some models were supplied with as many as 24 data inputs while some models relied on only one data input. Hence, a clear indication of the contribution of these variables to the prediction model could not be identified from this review.

2.3.3 Comparison of Techniques

Time-series forecasting and machine learning are the two popular methods that have been widely adopted into many of the prediction models and have delivered good results. Techniques used in each study are listed in Table 2.1. Support Vector Regression (SVR), ARIMA, NN and GE stood out as the most frequently used techniques compared to the rest. RF models delivered one of the lowest error values (Nguyen & Rokicki, 2018). However, there was no direct correlation between a certain chosen method to an accuracy level as each study used a different approach, different sets of variables and data. Hence, a fair comparison of techniques could not be made. Nevertheless, the approaches used in these studies asserted that time series forecasting and machine learning algorithms are suitable for predicting blood glucose levels and it should be continuously explored and optimized in the future for continuous improvement.

2.3.4 Performance Evaluation

Assessing the capability of the prediction models is imperative to ensure the performance metrics of a prediction model are met. The performance metrics signify the quality of the prediction model and its functional capability. The most popular measurement that was used in these researches to quantify the errors was the Root Mean Square Error (RMSE) calculation. Majority of the studies highlighted in this literature mainly used the 30 minutes and 60 minutes prediction horizon test for accuracy evaluation. A few studies also included the 15 minutes and 120 minutes horizon test for performance evaluation (Contreras et al., 2017; Ben Ali et al., 2018). It was found that, based on RMSE measurements, whenever a timing comparison was performed in a study; as the prediction horizon timing increased, prediction accuracy tends to decrease. In other words, most models were able to predict more accurately in a shorter time horizon.

2.4 Summary

Table 2.1 shows the summary of studies on blood glucose predictions models that were reviewed in this chapter. It can be noted that various prediction models have been extensively analysed and experimented using various techniques, input variables and under different conditions to improve prediction accuracy. They have achieved astounding breakthroughs and constantly displayed promising results. With almost half a billion people in the world being diabetic, it is crucial to continuously improve and strengthen the existing blood glucose prediction models. A solid blood

glucose prediction model that can be personalised and could be developed on a minimal amount of data will be a great advantage to lessen the impact of this critical disease. Although many good prediction models were reviewed in this study, the observations made indicate a lack of personalised, non-CGM data based blood glucose prediction models. As the vast majority of diabetic patients do not use CGM, this will be a cutting-edge invention for self-management of diabetes; which is the main goal of this study.

Table 2.1: Summary of Studies on Blood Glucose Prediction Models

Citation of Study	Diabetes Type (No of Patients)	Inputs	Model	Pre-processing Method	Prediction Technique	Result - RMSE (mmol/L)	
						Prediction Horizon	
						30 mins	60 mins
Bunescu <i>et al.</i> (2013)	Type 1 (5)	CGM data, daily events insulin	Physiological	Feature engineering	SVR without ARIMA & without feature selection	1.27	2.34
					SVR with ARIMA & without feature selection	1.23	2.29
					SVR without ARIMA & with feature selection	1.09	2.01
					SVR with ARIMA & with feature selection	1.08	1.98
Plis <i>et al.</i> (2014)	Type 1 (5)	CGM data, daily events insulin	Physiological	EKF	SVR	1.26	1.99
					ARIMA	1.38	2.2
Rollins <i>et al.</i> (2010)	Type 2 (1)	CGM data, meal, activity, stress, time of the day	Data-driven	Forward feature selection	Block-Oriented Wiener Network	-	-
Georga <i>et al.</i> (2011)	Type 1 (7)	CGM data, meal, insulin, exercise	Data-driven	-	SVR	0.90	1.38
Gani <i>et al.</i> (2011)	Type 1 (27) Type 2 (7)	CGM data	Data-driven	-	AR	3.3 (on average)	
Fiorini <i>et al.</i> (2017)	Type 1 (72) Type 2 (34)	CGM data, insulin, meal, exercise, sleep	Data-driven	-	ARIMA	0.93	2.06
					KF	3.24	3.51
					KRR	0.86	1.87

					LSTM	1.46	3.06
Li & Fernando (2016)	-	Blood glucose record, insulin, meal, exercise, sleep	Data-driven	-	SVM	3.82	
					Decision Tree	2.28	
					Random Forest	2.20	
					Pooled Panel Data	2.19	
					Pre-clustered (2-cluster)	2.04	
					Pre-clustered (3-cluster)	1.87	
					Pre-clustered (5-cluster)	1.84	
					Pre-clustered (9-cluster)	1.73	
					Pre-clustered (43-cluster)	1.53	
Nguyen & Rokicki (2018)	Type 1 (8) Type 2 (1)	Blood glucose record, carbohydrate, insulin, physical activity	Data-driven	Sanity filter and stability filter	Simple Baselines		25.71
					AVG		12.96
					Context AVG		12.53
					ARIMA		13.88
					LSTM (10 iterations)		10.41
					LSTM (100 iterations)		19.24
					RandomForest (RF)		12.05
					Extra Trees (ET)		12.15
					RF + Sanity Filter		8.80
					ET + Sanity Filter		9.01
					RF + Sanity + Stability Filter		8.71
					RF + Stability Filter		7.57
Hamdi <i>et al.</i> (2018)	Type 1 (12)	CGM data	Data-driven	-	DE - SVR	0.60	0.72

Ben Ali <i>et al.</i> (2018)	Type 1 (12)	CGM data	Data-driven	-	ANN	0.38	0.50
Contreras <i>et al.</i> (2017)	Type 1(100)	CGM data, carbohydrate, insulin	Hybrid	-	GE	1.18 (120 min)	
Contreras <i>et al.</i> (2018)	Type 1 (6)	CGM data, insulin, physical activity	Hybrid	Feature Engineering	GE	1.18	1.74
Hidalgo <i>et al.</i> (2017)	Type 1 (10)	CGM data, insulin, carbohydrate	Hybrid	Data aggregation and feature engineering	GP GE	-	-
Zecchin <i>et al.</i> (2014)	Type 1 (10)	CGM data, carbohydrate	Hybrid	-	Jump Neural Network	0.92	-
Balakrishnan <i>et al.</i> (2013)	Type 1 (34)	CGM data, insulin, meal, exercise	Hybrid	-	TFM	-	-

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Introduction

This chapter introduces the research approach adopted for this study. The complete research methodology framework that is used to develop the prediction model is provided and explained in this chapter. The theoretical explanations behind each stage of the research framework and the justifications for selecting the approaches is discussed in detail. The stages that are discussed in this chapter include the data collection method, data pre-processing techniques, data exploration analysis, model development process and the model evaluation criteria.

3.2 Research Approach

The schematic representation of the overall research methodology proposed in this study is shown in Figure 3.1. There are five key stages in this research framework, namely the data collection, data pre-processing, exploratory data analysis, model development and model evaluation. Each of these key stages and the methodologies applied are explained in the subsequent sections.

3.2.1 Data Collection

The first phase of this research study is the data collection process. The dataset for this research is sourced from several diabetic patients from around the world with the support of social media platforms. All subjects are above 30 years of age and consist of both type 1 and type 2 diabetics. Subjects' self-recorded blood glucose level data were obtained from them on a volunteer basis with their consent for research. The data set that is collected for this research purpose, consist of the patient's age, diabetes type, gender, and time-based blood glucose level records. Data from the subjects are not continuous, in other words, the blood glucose levels were not recorded on a fix interval periods. The blood glucose level values are found to be recorded at random timing for different days. Data were collected from 10 volunteers in total. It was ensured that a single subject has at least 60 blood glucose records in total to be qualified for this study. And the blood glucose records must be self-monitored data, in other words, not obtained using a CGM device. Based on these criteria, six subjects were shortlisted as the qualified candidates for this research study. The data collected from all subjects were anonymized. Henceforth, all research subjects will be referred by a uniquely assigned patient ID.

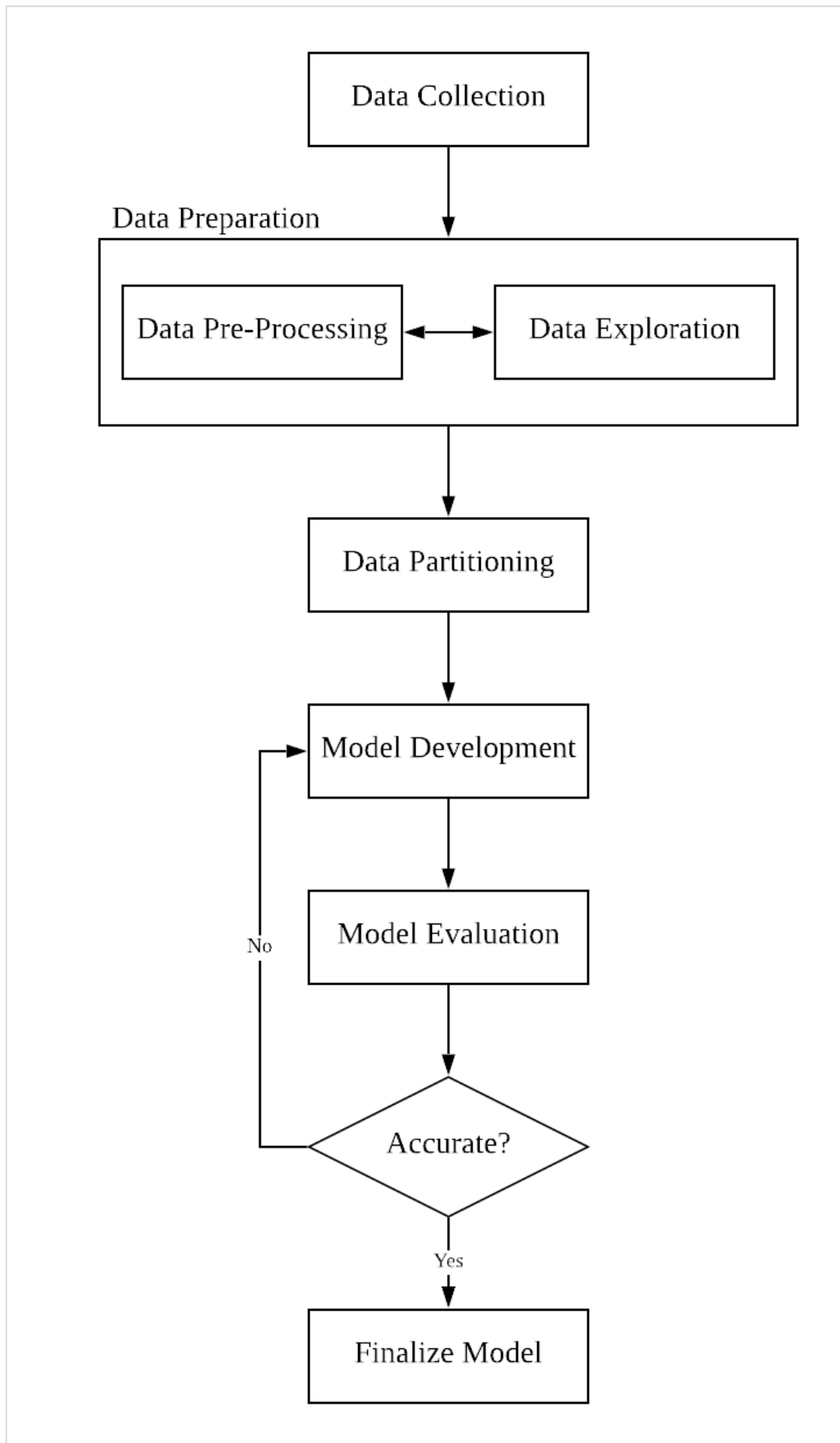


Figure 3.1: Framework for the Blood Glucose Prediction Model Development

3.2.2 Description of Data Set

The final dataset to be used in this study consists of over 1900 observations collectively from all research subjects and 8 variables describing the characteristics of the diabetic subjects. Table 3.1 describes the variables present in the data set used in this study.

Table 3.1: Description of Variables in Data Set

Attributes	Description
Patient ID	A unique identifier of a research subject
Diabetes Type	Type 1 or Type 2 Diabetes
Age Group	Research subject's age categorized in age groups
Gender	Research subject's gender
Country	The country research subject is from
Date	Date blood glucose is recorded
Time	Time blood glucose is recorded
BG Records	Blood glucose record in mg/dl or mmol/L unit

3.2.3 Data Pre-processing

Data pre-processing is an integral part of this study. Data pre-processing involves tasks such as data integration, data cleaning, data transformation, and data reduction. Data pre-processing enriches the data to be the best fit for the model. This would help to improve the accuracy of the model outcome. Hence, the data collected for this study would be pre-processed to fit the requirements of this research. The data pre-processing for this study involves variable selection task, data transformation and data cleansing.

Firstly, the data collected from research subjects will be analysed to identify the most appropriate variables to be used in the model development process. Some of the variables such as the diabetes type, country and age of the research subjects may only be used for the purpose of exploration and a better understanding of the researcher. Meanwhile, variables such as date, time and BG records will be used as the main input for the prediction model. Once the appropriate variables are selected, data will be processed to identify any inconsistencies. It can be noted that the value for variable such as the BG records are given in two different units, thus, this variable would require transformation. A standard unit will be adopted and the data will be transformed to be in a uniform manner.

Data cleaning task will then be performed to fix the missing values, smooth noisy data, identify or remove outliers, and resolve the inconsistencies in the dataset. As the dataset used in this study are manual records of research subjects, from manual observation, not many missing values are noticed in this dataset. This has to be confirmed with a further detailed exploration of data. Missing values or outliers found through the exploration phase has to be treated accordingly. Missing values or outliers found in dataset equalling to less than five percent of total data will be ignored and removed from observations. However, if a substantial amount of missing values or outliers found in data, data imputation techniques such as the mean imputation will be applied for the observations.

3.2.4 Exploratory Data Analysis

Exploratory Data Analysis (EDA) is the step taken to perform preliminary investigations on data to discover patterns, spot irregularities, test hypothesis and check assumptions with the support of statistics, visual and graphical representations (Jebb *et al.*, 2017). The primary objective of using the EDA process in this research is to have a better understanding of data in hand and to identify abnormalities. Having a good grasp of data would greatly help in strengthening the performance of the prediction model. This task will produce some visual outputs (i.e. graphs and charts), that would provide a basis for further data pre-processing if deemed necessary. The data pre-processing and EDA process will be performed simultaneously to ensure all issues found in data are addressed.

3.2.5 Model Development

As the blood glucose records which are the main data input in this research, is a series of data points indexed based on time order, the dataset used in this study is considered as a time-series data. And times series analysis is a statistical technique that is used to identify systematic patterns or trends in time series data. In a profound research back in 1999, the authors Bremer & Gough made clear that the dynamics of blood glucose have a detectable structure; therefore the glucose level can be predicted by making use of its recent historical data. Since that work, several studies have considered time series forecasting models for blood glucose prediction (Gani *et al.*, 2011; Bunesu *et al.*, 2013; Fiorini *et al.*, 2017) and they have demonstrated good prediction ability when dealing with continuous time-based inputs such as CGM data.

However, as this research intends to use non-continuous data, addressing the issue of missing data in certain times of the day in a time series analysis is considered to be even more challenging (Che *et al.*, 2018). Hence, a deep learning model is introduced to address the time series analysis challenges. Deep learning models have the capability to handle the long-term temporal dependencies in time series. They are also known to utilize the patterns of missing data to achieve better prediction results (Gamboa, 2017).

Recurrent Neural Networks (RNN) have been making great progress in becoming the best theoretical model for time-series analysis (Che *et al.*, 2018). In an RNN, the information cycles through a loop. When it makes a decision, it takes into consideration the current input and also what it has learned from the inputs it received previously. Figure 3.2 illustrates the information flow in an RNN. The process of producing an output and copying the output by looping back into the network is known as the internal memory of RNN.

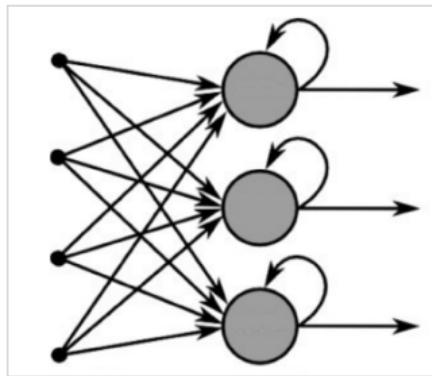


Figure 3.2: Information Flow in RNN

Long Short Term Memory (LSTM) network is an extension of RNN that contains special units called memory cells that are capable of learning long-term dependencies in data (Hochreiter & Uergen Schmidhuber, 1997). LSTM networks have internal contextual state cells that act as long-term or short-term memory cells. The output of the LSTM network is modulated by the state of these cells. This is a key feature of a LSTM model as it is able to make predictions based on the historical context of inputs, rather than only the very last input. In the context of this study, blood glucose level predictions also cannot be solely based on the last input as the current blood glucose level of a subject is heavily influenced by the historical data in previous hours or days according to the patient.

LSTM network can achieve this as it has the capability to keep the contextual information of inputs by integrating a loop that allows information to flow from one step to the next. The unrolled loop of a LSTM network is illustrated in Figure 3.3. This looping mechanism enables the network to make predictions conditioned by the past experience and trend of the patient.

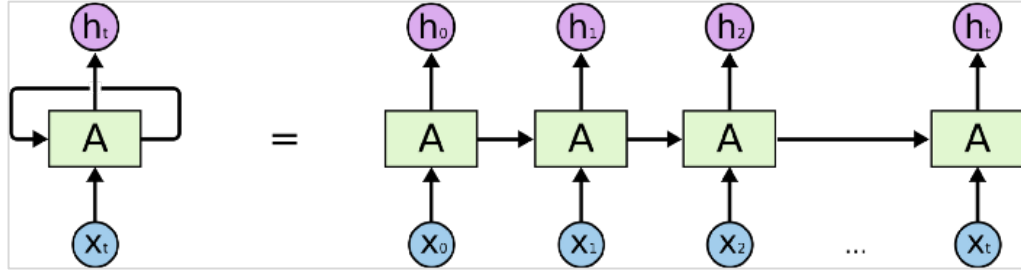


Figure 3.3: LSTM Loop Unrolled

Furthermore, the memory cells that lie within the LSTM network as illustrated in Figure 3.4, functions as a gated cell, where the cell decides whether to store or delete information based on the weights of the input data. Thus, over time and training, the LSTMs are able to capture time-dependencies in the data and find their relationship to the model output (Fiorini *et al.*, 2017). This feature is very useful for this study, as the research subject's data is not continuous and contain many missing values for specific time periods. Thus, it is expected that the LSTM model used in this study, would be able to recognize the pattern of missing values and decide to use or ignore the information of missing values when making predictions.

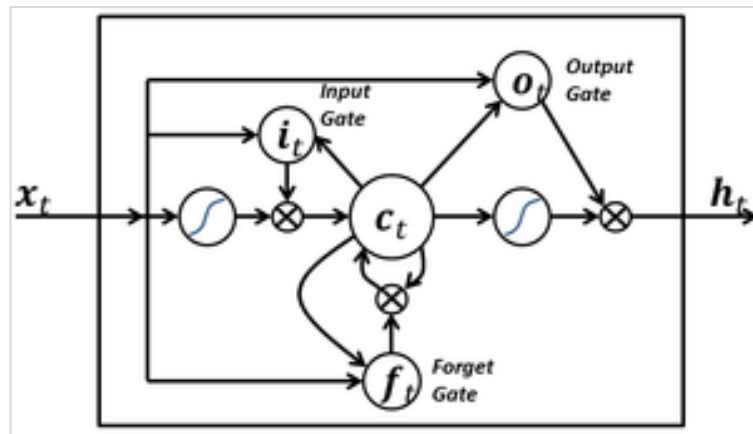


Figure 3.4: Internal Network of LSTM

Considering the features and suitability of the model for the type of dataset used in this study, an RNN-LSTM based deep learning model would be developed in this study and its ability to predict blood glucose levels will be evaluated.

3.2.6 Model Evaluation

The dataset that has gone through the pre-processing and exploration will be divided into two groups known as training and testing data. 80 percent of the data will be used as training data, and the remaining 20 percent will be used as test data. The model would be initially fitted using the training data, for it to learn the patterns in the dataset and to train it to make predictions accordingly. Once the model is fully trained, testing data is supplied to the model and it would be expected to make predictions according to the attributes that exist in the testing data.

The predicted blood glucose level values will be compared with the expected blood glucose levels values to measure the performance of the model. Root Mean Square Error (RMSE) calculation will be used to measure the performance of the model. RMSE can be defined as a measure of the differences between the predicted and actual values of the target variable. A smaller RMSE value will be desired as the smaller is the RMSE value, the more accurate is the prediction (Ben Ali *et al.*, 2018).

The results obtained would be analysed and the model parameters may be tuned for further improvement. The hyperparameter tuning of the model may be repeated and the prediction model will continuously be trained and improved until the desired accuracy is achieved. Related works in this literature showed that 30 and 60 minutes prediction horizon were the commonly used measurement and provided more reliable predictions compared to a longer time horizon. Thus, the predicted result in this study will be tested based on a 30 minutes and 60 minutes prediction horizon.

3.3 Summary

This chapter described the research approach adopted for this study. The five main stages in the research framework, namely, data collection, data pre-processing, data exploration, model development and the model evaluation were given emphasis. The theoretical concepts and motives behind each stage of the data analysis and model development were described in this chapter. Idealistically, an RNN-LSTM based deep learning model would be developed and tested in this study using the data acquired and processed in the initial stages of this research. The concepts and justifications behind techniques proposed have been thoroughly described in this chapter and the specific details of implementation will be discussed in the following chapter.

CHAPTER 4

IMPLEMENTATION

4.1 Introduction

This chapter focuses on each stage of the model implementation with a detailed explanation of the adopted approaches and techniques in this study. The dataset that is used for the prediction model and the pre-processing techniques that were applied will be discussed in detail. Many data pre-processing practices such as data transformation, data exploration, data standardization, variables selection, treatment of outliers, treatment of missing values and etc were applied in this study. A detailed explanation of steps involved in each of this data pre-processing methods is provided. As the prediction model developed in this study has some specific requirements, the pre-processed data is then accordingly prepared to meet the requirements of the model. The data preparation steps involved and the data partitioning method that was applied is also discussed in this chapter. Finally, the steps involved in developing the LSTM model, the parameters that were defined for the model and the evaluation method applied to test the performance of the model is also described in detail.

4.2 Data Pre-Processing

Data pre-processing is an integral step to model development in order to ensure the model fits the requirements of this study. During initial analysis, it was found that the data is not clean and some inconsistencies in data were noticed. Hence, further detailed exploration was performed on the dataset and the pre-processing steps involved in preparing the data to be fit for the prediction model is explained in subsequent sections.

4.2.1 Dataset Description

As this study intends to develop personalised models, the data for each subject were handled individually. Among the six subjects selected for this study, each subject's dataset consists of eight variables and a varying number of observations, as each subject's blood glucose records were collected at different periods of time. The eight variables in the dataset are described in Table 4.1 and the duration of data for different subjects are summarized in Table 4.2.

Table 4.1: Description of Variables in Dataset

No.	Attributes	Description
1	Patient ID	A unique identifier of a research subject
2	Diabetes Type	Type 1 or Type 2 Diabetes
3	Age Group	Research subject's age categorized in age groups
4	Gender	Research subject's gender
5	Country	The country research subject is from
6	Date	Date blood glucose is recorded
7	Time	Time blood glucose is recorded
8	BG Records	Blood glucose record in mg/dl or mmol/L unit

Table 4.2: Summary of dataset for each selected subject

No.	Patient ID	Number of Observations	Duration of BG Record	
			From	To
1	P01	93	21 December 2017	23 January 2018
2	P02	568	31 January 2018	8 August 2018
3	P03	226	27 December 2017	27 November 2018
4	P04	98	14 February 2018	12 May 2018
5	P06	131	14 March 2018	25 June 2018
6	P07	847	31 December 2017	1 May 2018

4.2.2 Variables Selection

Although the dataset for each subject consist of eight variables, the first five variables as shown in Table 4.1 are merely for the researcher's study and understanding but are not required for the model development. As this study intends to develop a personalised model only using non-CGM data; only the Date, Time and BG Records of each patient are used as the input variables for the prediction model.

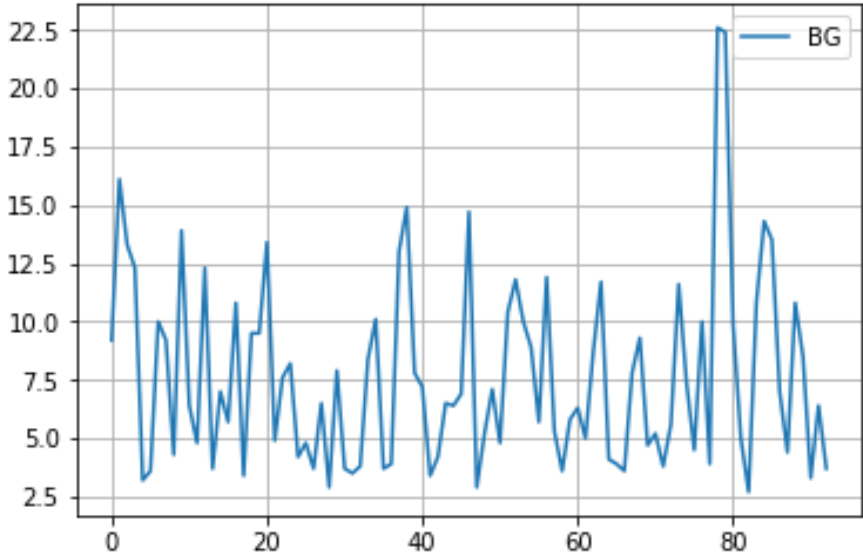
4.2.3 Data Transformation

As the subjects in this study came from different countries and regions, the data collected from them were not in a uniform manner. The date and time were recorded in different time zone formats and the key attribute blood glucose records were found in two different measurement units i.e. mmol/L and mg/dl. The international standard way of measuring blood glucose levels is in mmol/L (millimoles per litre). However, in some countries such as the United States and Germany, the measurements are recorded in mg/dl (milligrams per decilitre). Hence, in this study, the international standard way of measurement is adopted and the records of patients were standardized to mmol/L units. As for the time variable, 24-hour time format was adopted.

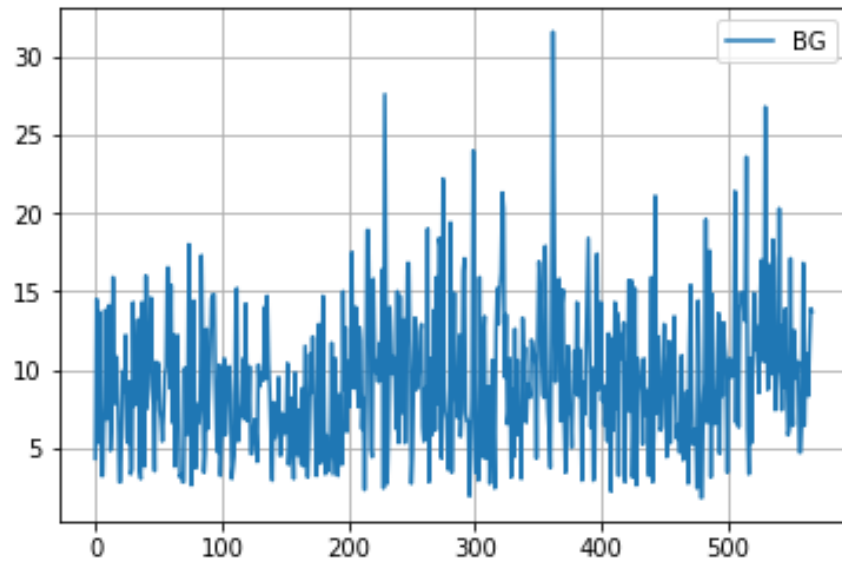
4.2.4 Data Exploration

In order to better understand the data, the trend of blood glucose of records of each patient and to identify abnormalities in data, exploratory data analysis was performed. Each subject's data were explored individually with the support of Matplotlib library in Python and Tableau software. The visual output of blood glucose records of each patient is shown in Table 4.3.

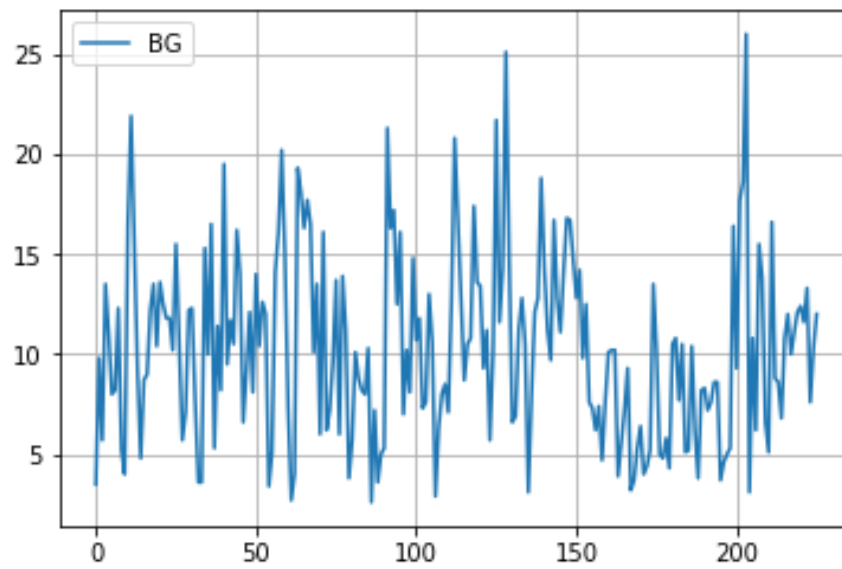
Table 4.3: Visual Output of Blood Glucose Records for Each Research Subject

Patient ID	Visual Output of Blood Glucose Records
P01	

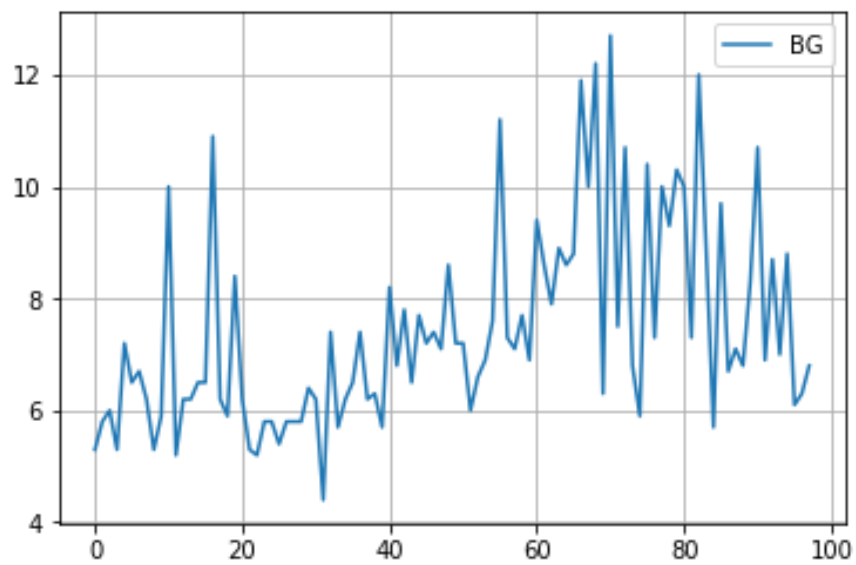
P02

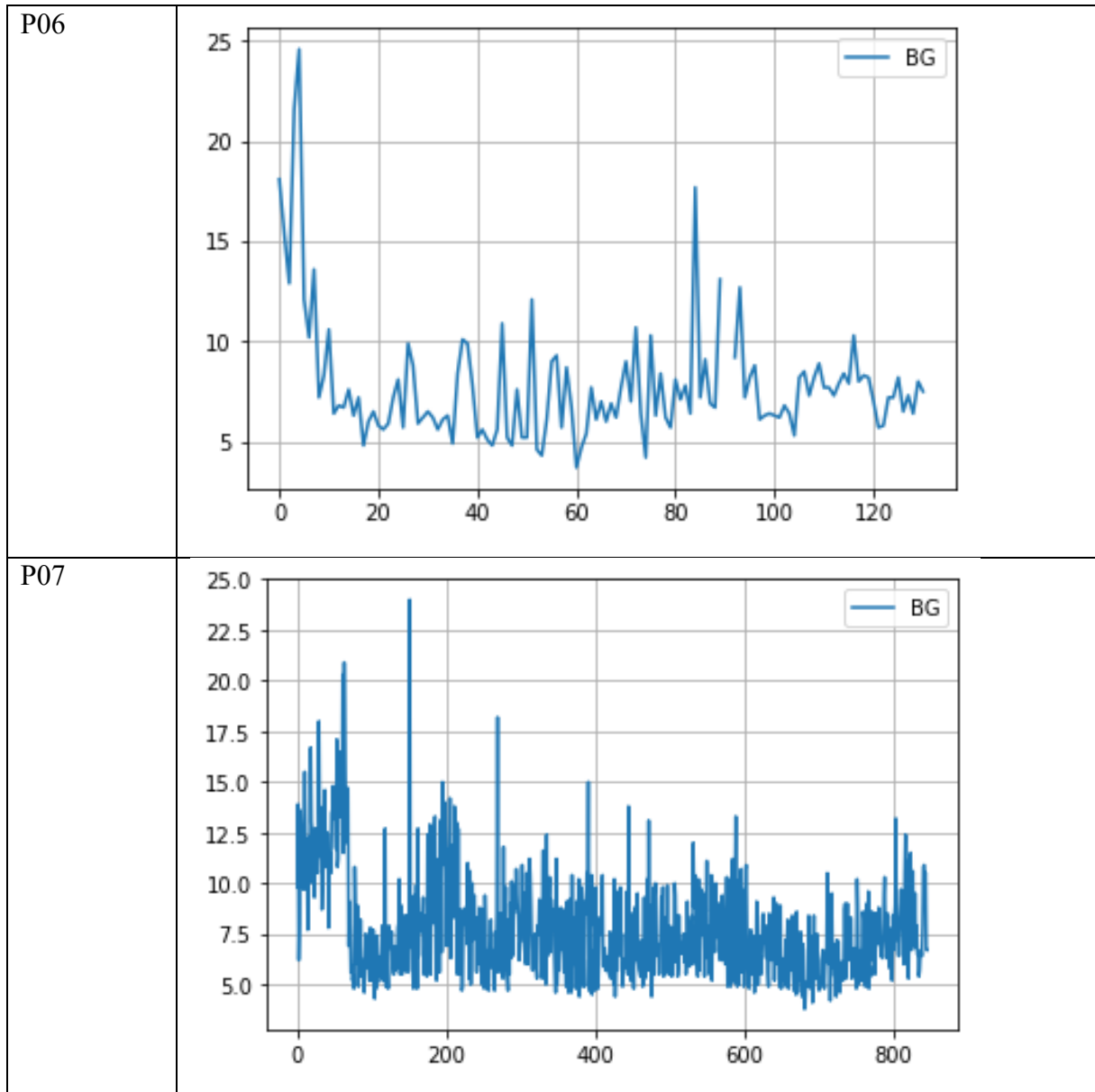


P03



P04





It was noted from this initial observation that some values in the blood glucose records are extremely high and not within the usual range of the particular patient. These values could have been misrecorded or occasional hyperglycaemia incidents. As the actual reason for the record could not be determined, they are considered as outliers and treated in order not to mislead the training of the model. The outlier treatment process is described in the next section.

In addition to individual research subject's data exploration, a comparative analysis was also performed to detect any similar trend or pattern among research subjects that can be useful for this study. All the six patient's average blood glucose levels were analysed by the hour as illustrated in Figure 4.1. No apparent trend was noticed among the subjects. Peak and low values of each patient are at different hours for different patients. The average blood glucose

value differs from one patient to another. Subjects P03 and P07 seem to have a more stable blood glucose levels throughout the day, meanwhile, the rest of the subjects' blood glucose level fluctuates more frequently. This finding confirms that having one universal model would not be a good fit for all patients as they do not share any common pattern. Hence, personalised models as proposed in this study will be a better fit for the research subjects.

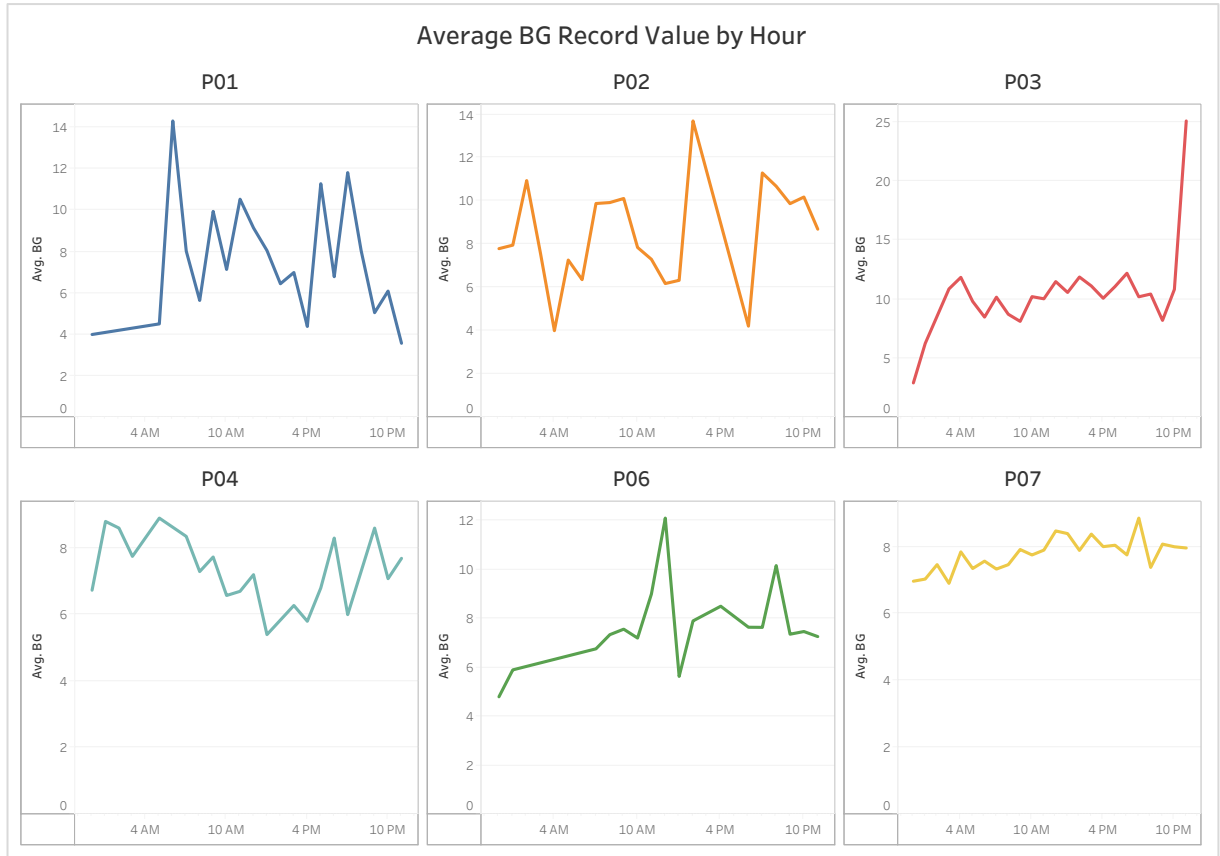
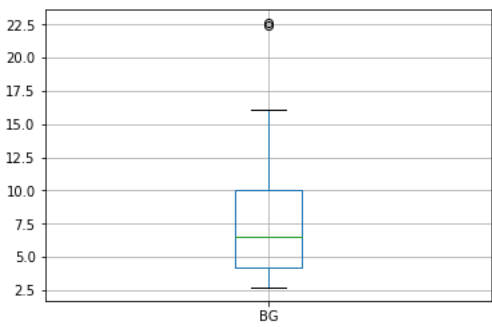
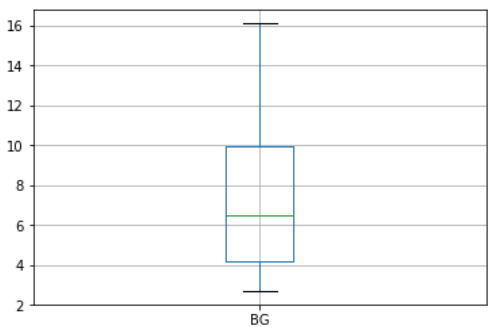
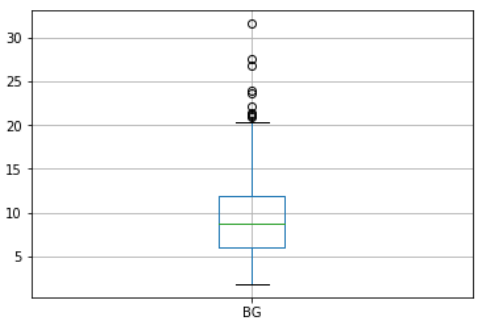
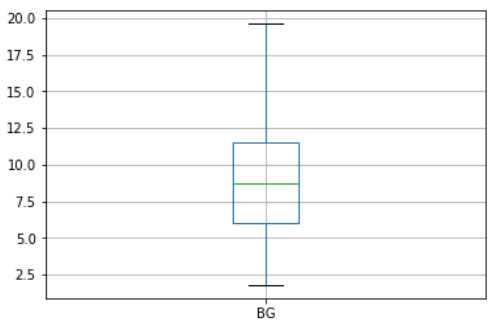
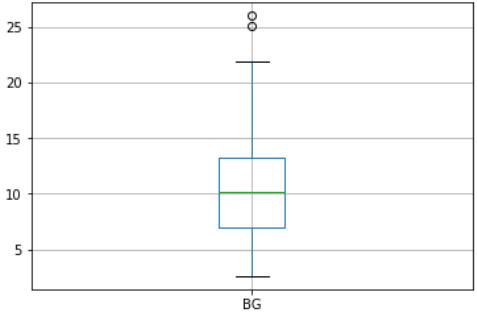
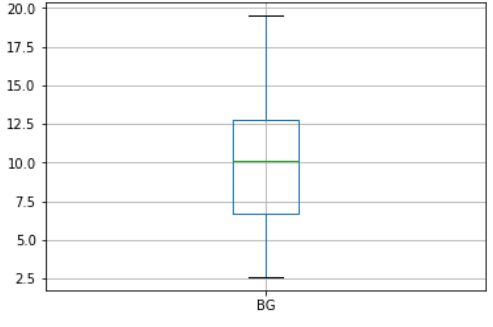
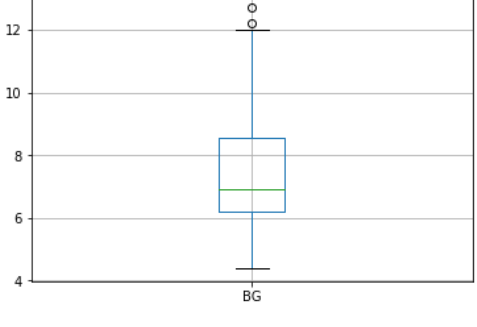
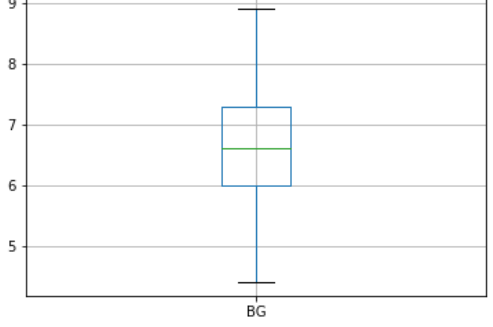


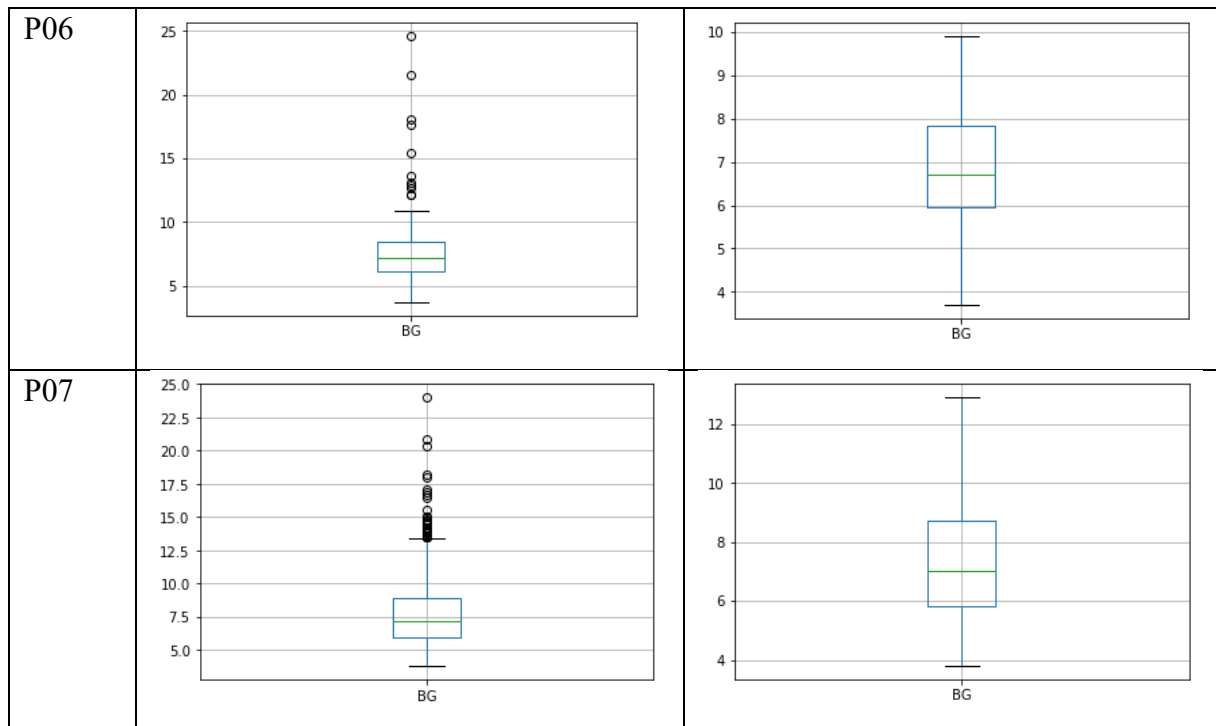
Figure 4.1: Average Blood Glucose Level of Research Subjects by Hour

4.2.5 Treatment of Outliers

To confirm the existence of outliers in patients' blood glucose records, boxplot charts were used. The boxplots that highlight the distribution of each patient's blood glucose records are shown in Table 4.4. By exploring these charts, it was noted that for each patient only a few records of outliers existed in the dataset. Thus, these observations were removed from the patient's record.

Table 4.4: Boxplot of Patient's Blood Glucose Records

Patient ID	Boxplot for Outlier Identification	Boxplot after Outlier Removal
P01	 <p>Boxplot for P01 showing a significant outlier at approximately 22.5. The median is around 6.5, with a box from 4.5 to 10.0. Whiskers extend from 2.5 to 16.0. The outlier is at 22.5.</p>	 <p>Boxplot for P01 after outlier removal. The median is around 6.5, with a box from 4.0 to 10.0. Whiskers extend from 2.5 to 16.0.</p>
P02	 <p>Boxplot for P02 showing multiple outliers between 20 and 30. The median is around 8.5, with a box from 6.0 to 12.0. Whiskers extend from 2.5 to 20.0. Outliers are at approximately 21, 22, 23, 24, 25, 26, 27, 28, 29, and 30.</p>	 <p>Boxplot for P02 after outlier removal. The median is around 8.5, with a box from 6.0 to 11.0. Whiskers extend from 2.5 to 20.0.</p>
P03	 <p>Boxplot for P03 showing two outliers at approximately 24 and 25. The median is around 10.0, with a box from 7.0 to 13.0. Whiskers extend from 2.5 to 22.0. Outliers are at 24 and 25.</p>	 <p>Boxplot for P03 after outlier removal. The median is around 10.0, with a box from 7.0 to 12.5. Whiskers extend from 2.5 to 20.0.</p>
P04	 <p>Boxplot for P04 showing three outliers at approximately 11, 12, and 13. The median is around 6.5, with a box from 6.0 to 8.5. Whiskers extend from 4.0 to 12.0. Outliers are at 11, 12, and 13.</p>	 <p>Boxplot for P04 after outlier removal. The median is around 6.5, with a box from 6.0 to 7.5. Whiskers extend from 4.0 to 9.0.</p>



4.2.6 Data Standardization

During the exploratory data analysis, it can be noted from the visual outputs as shown in Table 4.3 that, there are no indication of missing values in the dataset for all the patients. This is due to the nature of the raw data collected from the patients. Data collected from the research subjects were in the form of records on random time and days with no specific requirement of timing. As the timing of blood glucose records were not standardized, no missing value seem to exist in the raw dataset.

However, as this study intends to develop a prediction model that would be able to predict in 30 and 60 minutes prediction horizon, the time variable on the dataset has to be standardized according to the prediction horizon. For this purpose, two new datasets were developed based on the original dataset. In the first dataset, the time variable is defined to be in 30 minutes interval. Time is defined starting from 00:00 to 23:30 and the existing records based on random timing is rounded to the nearest 30 minutes. This creates 48 observation points for each day. The example of the dataset after the standardization of timing to 30 minutes interval is shown in Figure 4.1.

Index	Date	Time	BG
12	2018-02-14	06:00:00	nan
13	2018-02-14	06:30:00	nan
14	2018-02-14	07:00:00	nan
15	2018-02-14	07:30:00	nan
16	2018-02-14	08:00:00	nan
17	2018-02-14	08:30:00	nan
18	2018-02-14	09:00:00	nan
19	2018-02-14	09:30:00	nan
20	2018-02-14	10:00:00	5.3
21	2018-02-14	10:30:00	nan
22	2018-02-14	11:00:00	nan
23	2018-02-14	11:30:00	nan
24	2018-02-14	12:00:00	nan

Figure 4.2: Example of the dataset after the standardization of timing to 30 minutes interval

Similarly, for the second dataset, the time variable is defined to be in 60 minutes interval. Time is defined starting from 00:00 to 23:00 and the existing records based on random timing is rounded to the nearest 60 minutes. This creates 24 observation points for each day. The example of the dataset after the standardization of timing to 60 minutes interval is shown in Figure 4.2.

Index	Date	Time	BG
0	2018-02-14	00:00:00	nan
1	2018-02-14	01:00:00	nan
2	2018-02-14	02:00:00	nan
3	2018-02-14	03:00:00	nan
4	2018-02-14	04:00:00	nan
5	2018-02-14	05:00:00	nan
6	2018-02-14	06:00:00	nan
7	2018-02-14	07:00:00	nan
8	2018-02-14	08:00:00	nan
9	2018-02-14	09:00:00	nan
10	2018-02-14	10:00:00	5.3
11	2018-02-14	11:00:00	nan
12	2018-02-14	12:00:00	nan
13	2018-02-14	13:00:00	nan

Figure 4.3: Example of the dataset after the standardization of timing to 60 minutes interval

4.2.7 Treatment of Missing Values

The data standardization process that was carried out to standardize the time of blood glucose records according to prediction horizon creates a lot of missing values in the dataset. As patients do not self-record their blood glucose levels for every 30 or 60 minutes unlike in CGM use, the new datasets developed contain many missing values and it requires careful treatment prior to inputting this information into the prediction model. Two strategies were adopted to deal with the issue of missing values, namely, group mean imputation and masking missing values.

Firstly, using the group means imputation strategy, the missing values are replaced with the group mean of all known values of that attribute. In this case, if a blood glucose record is found to be missing at the time interval of 10:00, the dataset will be searched for all other blood glucose records during that exact time on different days. These values were grouped and their mean value was calculated. Imputation is then performed using this mean value for that particular time attribute. This strategy is adopted with the assumption that a particular patient's blood glucose record at a particular time for different days should be similar. Most of the missing values in the dataset were able to be treated by using this strategy.

However, it did not solve all the missing value issues. The total number of remaining missing values after the group means imputation is summarized in Table 4.5. A substantial amount of missing values are still found after the group mean imputation step, except for only one research subject with the Patient ID of P07. This was mainly due to a lack of consistent records by patients at every hour even on different days. Some of the timing is just not feasible for self-recording of blood glucose as patients could be sleeping or working during that hour. This is anticipated in this study, hence why, an LSTM model was proposed to overcome the issue of missing values and inconsistent data. Thus, for the remaining missing values in the dataset, the missing values were imputed with a zero value. As the zero value was not part of the normal range of blood glucose levels for all patients, over time and training, the LSTM model is capable of recognizing these zero values as missing values and treat them accordingly (Che et al., 2018). With this, the problem can be modelled as-is and the LSTM model is encouraged to learn the pattern of the missing values.

Table 4.5: Summary of Missing Values Before and After Group Mean Imputation

Patient ID	30 Minutes Dataset		60 Minutes Dataset	
	Missing values before group mean imputation	Missing values after group mean imputation	Missing values before group mean imputation	Missing values after group mean imputation
P01	1541	408	732	136
P02	8582	1710	4031	570
P03	2088	240	940	0
P04	4129	1672	2018	352
P06	4691	1900	2293	700
P07	7080	0	3153	0

4.3 Data Preparation

As the dataset that is used in this study is predominantly based on a time factor, and the blood glucose records which are the main data input in this research, is a series of data points indexed based on time order, the data is considered as a time-series data. Thus, prior to inputting the time-series data into a deep learning model, a few necessary actions were taken to transform the time-series data to fit the requirements of a deep learning model.

4.3.1 Transform Timeseries to Supervised Learning

Supervised learning models such as the LSTM assumes that the dataset is divided into input (x) and output (y) components. In time-series data, this can be achieved by using the observation from the last time step ($t-1$) as the input and the observation at the current time step (t) as the output. This was performed using the support of the Pandas library in Python to push all values in a series down by one place. This becomes the input variables and the existing time-series will be considered as the output variable. These series are then concatenated to form a data frame to be used as a supervised learning problem.

4.3.2 Transform Timeseries to Scale

Scaling input and output variables is a critical step in using neural network models. Small values in the range of 0 to 1 are usually preferred. A target variable with a large spread of values, in turn, may result in large error gradient values causing weight values to change dramatically, making the learning process unstable. Unscaled input variables can result in a slow or unstable learning process, whereas unscaled target variables on regression problems can result in exploding gradients causing the learning process to fail. Thus, min-max scaling with the help of the Pandas library in Python was performed in this study. The data that is prepared to be inputted into the LSTM model was scaled to a fixed range of 0 to 1. During the testing and validation process, the scale of data was then inverted to return the values back to the original scale so that the results can be interpreted and a comparable error score can be calculated.

4.4 Data Partitioning

The data that has gone through the pre-processing steps and prepared for the model as discussed is then partitioned into training and testing datasets. 80 percent of data from each research subject was allocated as training data and 20 percent of the data was allocated to be used for testing purposes. The final size of the datasets used in this study for each research subject is summarized in Table 4.6.

Table 4.6: Summary of total observations in training and testing datasets

Patient ID	30 Minutes Dataset			60 Minutes Dataset		
	Total	Training Dataset	Testing Dataset	Total	Training Dataset	Testing Dataset
P01	1631	1305	326	822	658	164
P02	9105	7284	1821	4554	3644	910
P03	2304	1844	460	1154	924	230
P04	4223	3378	845	2112	1690	422
P06	4801	3841	960	2403	1923	480
P07	7880	6304	1576	3949	3160	789

4.5 Model Development

In light of the advantages of LSTM in addressing the specific needs of this study, an LSTM model was developed using the Keras library in Python. One of the key advantages of LSTM is that it is able to learn and remember long sequences of data and does not rely on a pre-specified window lagged observation as input. In Keras library, this is referred to as being 'Stateful' and this state was defined as 'True' when defining the LSTM layer. Besides that, the LSTM layer expects the input data to be in a matrix with three dimensions, namely, samples, time steps and features. Thus, the input data was reshaped to meet the requirement using the 'batch_input_shape' function. This function allows specifying the number of observations, time steps and features to be used in the model.

The number of neurons was set to be 2 initially and during hyperparameter tuning step, a different number of neurons such as 1, 2, 4, 8, and 16 were tested to evaluate the better performance of the model. The trend of performance when increasing and decreasing the neurons were monitored. In the final model, the number of neurons was set to be 4 as it gave the best performance comparatively.

Another key parameter to consider in LSTM layer is the learning rate of the model. The learning rate is usually between 0 and 1 value and it functions to control the change in the model in accordance with the estimated error each time the model weights are updated. Learning rate often depends heavily on the type of data and the model used. As personalised models using different datasets are developed in this study, it can be quite challenging to determine the right type of learning rate in accordance with the data for individual models. Thus, an adaptive learning rate approach was chosen to be implemented in this study. The adaptive learning rate method provides a heuristic approach without requiring expensive work in tuning hyperparameters for the learning rate schedule manually. The adaptive learning rate method is also known to provide good performance with sparse data. Thus, the adaptive learning method with the support of Adam optimization algorithm provided by the Keras library was implemented in this model.

The loss function for model evaluation was also required to be set for the LSTM network. As this model is used to solve a regression problem and expected to predict a continuous number, the 'mean_squared_error' was defined as the loss function. This functions as the base to calculate the RMSE value which is used to evaluate this model.

After defining all the key parameters as mentioned, the training data was fit into the model. Initially, the number of training iterations known as epochs was set to be 10. And during the hyperparameter tuning step, a different number of epochs such as 10, 50, 100 and 500 was tested to evaluate the better performance of the model. The trend of performance when increasing and decreasing the epochs were monitored. In the final model, the number of epochs was set to be 50 as it gave the best performance comparatively. The model was fit with all of the training data and was trained to predict each new time step one at a time from the test data. The predict function on the model was set to be 1, for it to make one timestep forecast on the test data. Once the model is fully trained, it then predicts the blood glucose values for all the observations in the test data. The expected value and predicted value is compared and RMSE is then calculated for the model. A line plot comparing the expected test data values and the predicted values is created at the end to provide a clearer understanding of the model performance.

As this study intends to develop personalised prediction models, six different LSTM models were developed in this phase. All six individual models shared the same parameters and settings. Different datasets prepared according to individual patients were supplied to each of the models and the model performance results were recorded individually for each model. Each patient model was trained independently using two types of datasets. First using the 30 minutes dataset and secondly using the 60 minutes dataset. At the end of the cycle, all six model's performance results were compiled, evaluated and analysed to distinguish its overall performance.

4.6 Summary

This chapter highlighted the processes involved in each stage of the implementation of this study. The dataset that is used for the prediction model and the pre-processing techniques that were applied were discussed in detail. The pre-processed data was then prepared to fit the requirements of the prediction model. The steps involved in developing the LSTM model, the parameters that were defined for the model and the evaluation method applied to test the performance of the model was also described in this chapter. In summary, this chapter described the adopted approaches and techniques at each stage of the model implementation. The results that were obtained from this model implementation will be presented in the following chapter.

CHAPTER 5

RESULTS AND ANALYSIS

5.1 Introduction

This chapter presents the findings and performance results of models developed in this study. The performance of the developed personalised blood glucose models and the results achieved is critically analysed. Each model developed is evaluated individually and a comparative analysis is performed at the end. The reasons and causes behind the certain performance behaviour of the models are discussed in detail. The performance of the model achieved in this study will then be compared with previous related work in the field. A detailed comparative analysis is performed and the findings are presented in this chapter.

5.2 Personalised Model's Prediction Result and Analysis

As described in the previous chapter, as this study aims to develop personalised models, six different LSTM models were developed in this study. The RMSE results obtained by each of the models are discussed in subsequent sections. Each of the six models are identified according to the individual patient ID, i.e. P01, P02, P03 and etc. For each patient, the model was trained independently using two types of datasets, i.e. the 30 minutes and 60 minutes datasets. The final RMSE results for each model are evaluated based on 30 minutes and 60 minutes prediction horizon.

5.2.1 Results of Model P01

The prediction results obtained for the model P01 is given in Table 5.1. In addition, the prediction values mapped against expected values are clearly illustrated in Figure 5.1. The 60 minutes prediction delivered a better result compared to the 30 minutes prediction. This is anticipated as the 60 minutes prediction is performed for a lesser number of observations compared to the 30 minutes prediction. It can be noted from the prediction plot that both predictions were able to follow the trend of true values and prediction matches quite accurately to the true values. However, the model did not perform well when predicting a higher range of blood glucose values.

Table 5.1: RMSE Result for Prediction Model P01

	30 Minutes Prediction	60 Minutes Prediction
RMSE	3.04	2.60

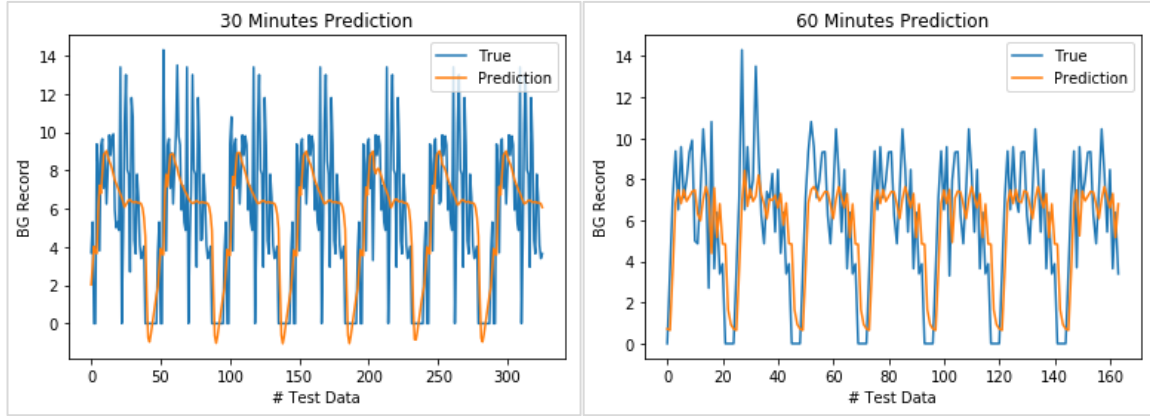


Figure 5.1: Prediction Result Plot for P01

5.2.2 Results of Model P02

The prediction results obtained for the model P02 is shown in Table 5.2. In addition, the prediction values mapped against expected values are clearly illustrated in Figure 5.2. It can be noted from the prediction plot that both predictions were able to follow the trend of true values and prediction matches quite accurately to the true values. However, the plots show that the model did not perform well when predicting the extreme high and low blood glucose values.

Table 5.2: RMSE Result for Prediction Model P02

	30 Minutes Prediction	60 Minutes Prediction
RMSE	3.32	2.71

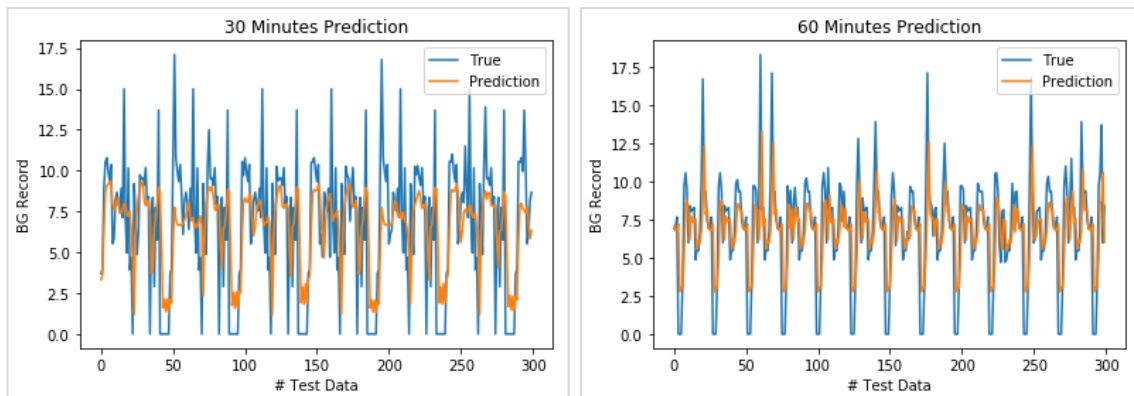


Figure 5.2: Prediction Result Plot for P02

5.2.3 Results of Model P03

The prediction results obtained for the model P03 is shown in Table 5.3. In addition, the prediction values mapped against expected values are clearly illustrated in Figure 5.3. This model recorded the highest error rate compared to others. For 30 minutes prediction, it can be noted that the model performed well for lower blood glucose level values, below 9. And in the 60 minutes prediction, the model performed best for blood glucose levels in the range of 6 to 10.

Table 5.3: RMSE Result for Prediction Model P03

	30 Minutes Prediction	60 Minutes Prediction
RMSE	3.70	3.02

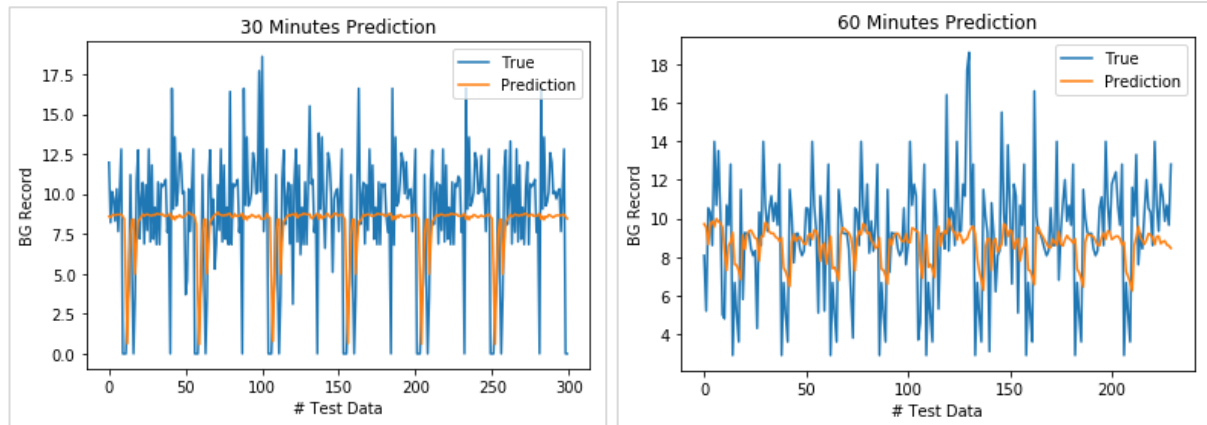


Figure 5.3: Prediction Result Plot for P03

5.2.4 Results of Model P04

The prediction results obtained for the model P04 is shown in Table 5.4. The prediction values mapped against expected values are clearly illustrated in Figure 5.4. For 30 minutes prediction, it can be noted that the model performed best for blood glucose levels in the range of 3 to 6. And in the 60 minutes prediction, the model performed best for blood glucose levels in the range of 5 to 8. However, the plots show that the model did not perform well when predicting the extreme high and low blood glucose values.

Table 5.4: RMSE Result for Prediction Model P04

	30 Minutes Prediction	60 Minutes Prediction
RMSE	3.62	3.08

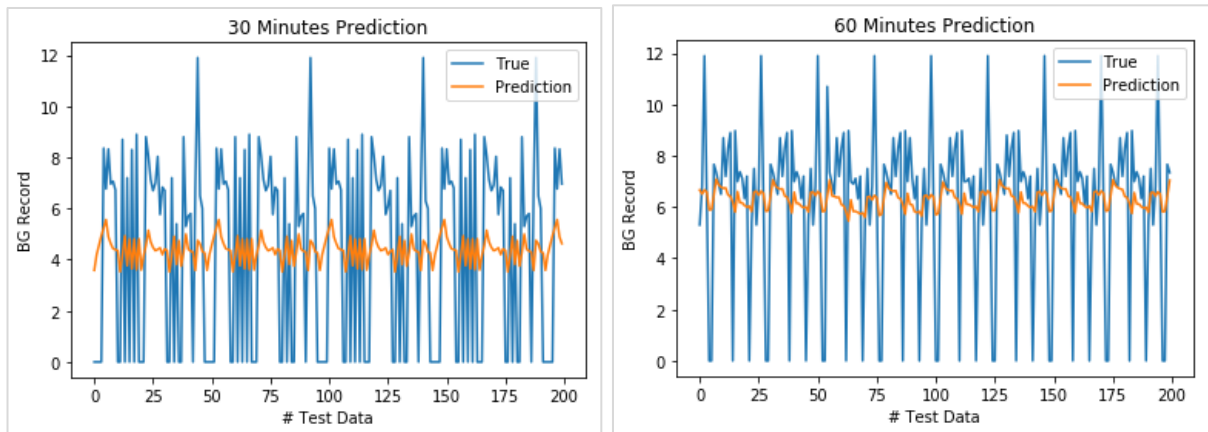


Figure 5.4: Prediction Result Plot for P04

5.2.5 Results of Model P06

The prediction results obtained for the model P06 is shown in Table 5.5. The prediction values mapped against expected values are clearly illustrated in Figure 5.5. It can be noted from the prediction plot that both predictions were able to follow the trend of true values and prediction matches quite accurately to the true values. However, the plots show that the model did not perform well when predicting the extreme high and low blood glucose values. But for blood glucose levels in the range of 2 to 6, it performed extremely well. Both the 30 minutes and 60 minutes prediction models showed similar performance although the 60 minutes prediction model reported a higher RMSE value.

Table 5.5: RMSE Result for Prediction Model P06

	30 Minutes Prediction	60 Minutes Prediction
RMSE	2.75	3.07

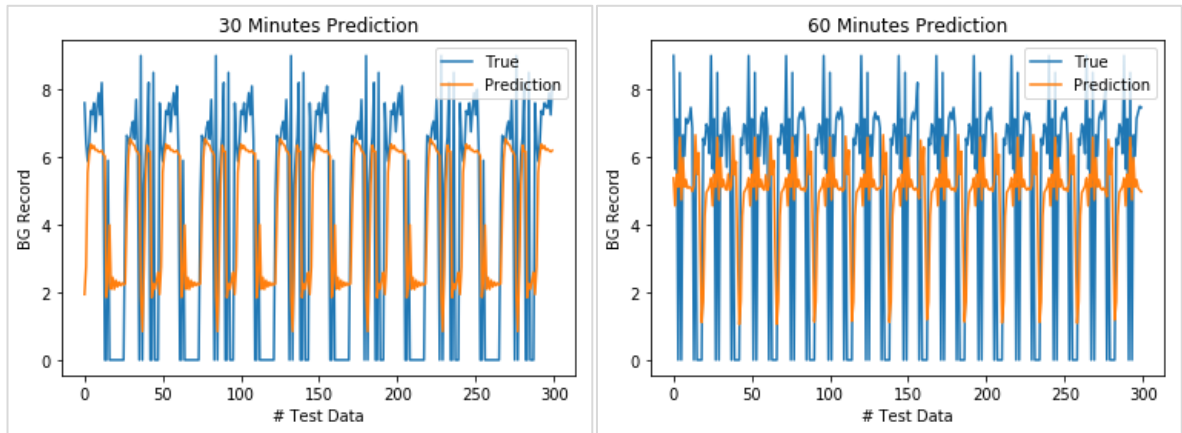


Figure 5.5: Prediction Result Plot for P06

5.2.6 Results of Model P07

The prediction results obtained for the model P07 is shown in Table 5.6. And a snippet of prediction values compared with the expected values are shown in Figure 5.6. It can be noted from this result that this model delivered the best performance with a very low RMSE value. The predicted values are very close to the expected values. The prediction values mapped against expected values are illustrated in Figure 5.7. It can be noted from the prediction plots that both predictions were able to follow the trend of true values and provided the most constant prediction values, averaging in a lower RMSE.

Table 5.6: RMSE Result for Prediction Model P07

	30 Minutes Prediction	60 Minutes Prediction
RMSE	0.76	0.91

```
Record=734, Predicted=7.49, Expected=7.48
Record=735, Predicted=7.49, Expected=7.23
Record=736, Predicted=7.48, Expected=7.25
Record=737, Predicted=7.47, Expected=7.46
Record=738, Predicted=7.47, Expected=7.49
Record=739, Predicted=7.47, Expected=7.51
Record=740, Predicted=7.48, Expected=7.30
Record=741, Predicted=7.47, Expected=7.88
Record=742, Predicted=7.48, Expected=7.09
Record=743, Predicted=7.47, Expected=7.44
```

Figure 5.6: Snippet of Prediction Result for P07

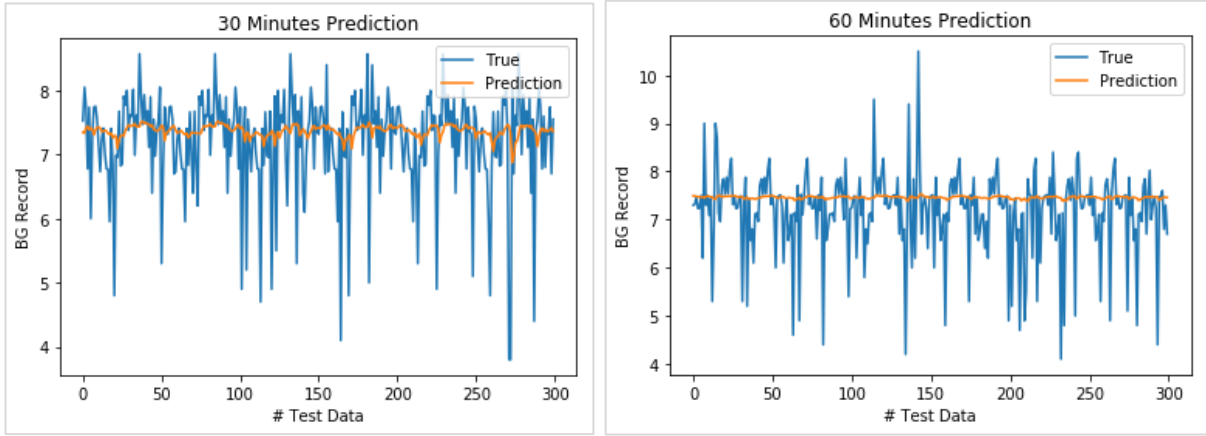


Figure 5.7: Prediction Result Plot for P07

5.3 Overall Prediction Result and Analysis

The individual model's results were aggregated and averaged over all patients. Table 5.7 shows the LSTM model performance averaged over all patients for 30 minutes and 60 minutes prediction horizon. Overall RMSE of 2.87 was achieved for 30 minutes prediction, meanwhile, the 60 minutes prediction achieved 2.57 RMSE. This indicates a significantly good prediction model performance.

Table 5.7: Overall RMSE Result

Patient ID/ RMSE	30 Minutes Prediction	60 Minutes Prediction
P01	3.04	2.60
P02	3.32	2.71
P03	3.70	3.02
P04	3.62	3.08
P06	2.75	3.07
P07	0.76	0.91
Average RMSE	2.87	2.57

Among the six models, model P07 achieved outstanding performance with a very low RMSE of 0.76 and 0.91 respectively. One of the possible reasons for this model's excellent performance is that this model was supplied with the highest number of input data as shown in Figure 5.8. Patient P07 had the highest records of blood glucose levels compared to the rest of the research subjects. However, when analysed more closely, Patient P02 also had a relatively high number of inputs compared to the rest of the research subjects. But the performance of

model P02 was not better than the rest. Thus, it can be said that having a high input data does not guarantee a good prediction performance.

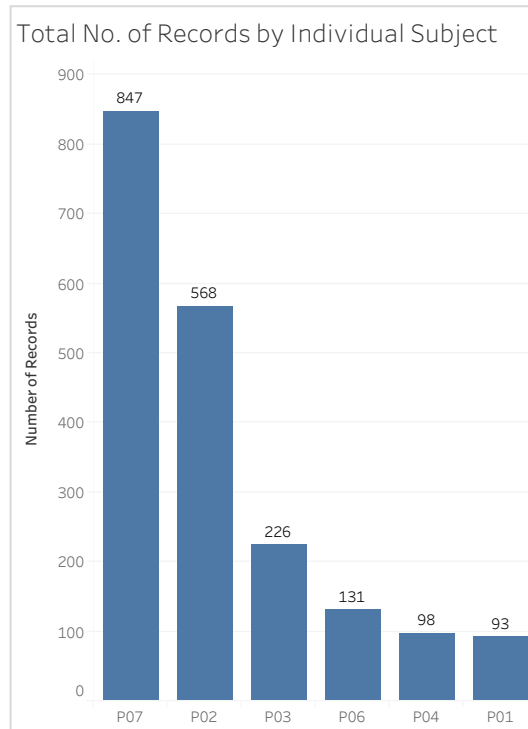


Figure 5.8: Total Number of BG Records by Individual Research Subjects

A more in-depth analysis was performed to understand the reason behind the good and poor performance of the prediction models. Figure 5.9 illustrates the frequency of recording blood glucose levels throughout the day. This chart provides a better view to understanding the relevancy of the issue. It can be noted from these charts that only Patient P07 had a high frequency of records for every hour throughout the day. This shows that the data provided by Patient P07 had a more holistic view of the patient's blood glucose levels throughout the day. Thus, the LSTM model was able to learn the pattern of the patient better and make more accurate predictions.

Meanwhile, for Patient P02 who also had a relatively high number of blood glucose records; it can be noted from the chart that, the patient did produce records throughout every hour of the day. Some particular hour has an extremely high number of records, while some hours does not have any records at all. Thus, learning the pattern of this particular patient could be more difficult for the prediction model. A similar trend is also noticed for Patient P04 who did not have records regularly throughout the day. It must also be noted that Patient P04 had very little observations in total when compared to Patient P02. However, the model performance results

do not differ much for the Patient P02 compared to Patient P04. Thus, this shows that the amount of observations inputted into the model is not a key condition for the model to perform better. However, having more regular monitoring and records throughout the day highly helps the LSTM model to understand the patient's behaviour and pattern better to provide more reliable prediction results. This assumption can be confirmed by the result of Patient P01. Patient P01 had the least amount of records compared to all six patients. However, the patient had more regular records or at least one record for every hour of the day. This really helped in the prediction process, as the model's performance result was relatively good among the other five models.



Figure 5.9: Number of BG Records by Hour for Individual Research Subject

It can be concluded from these results and findings that for an effective blood glucose level prediction, a high number of observations is not important. Reliable predictions can be made if the model is able to understand and learn the patient's blood glucose level trend throughout the day. The major plus point of this model is that the observations inputted into the model, does not have to be collected on the same day but could be performed for different hours on different days over any period of time. The prediction performance can be continuously improved as more hourly observations are supplied to the model.

5.4 Result Comparison with Previous Related Works

The results obtained from this study were compared with previous related works in the field to prove its performance capability. As highlighted in Table 5.8, two research studies in the past also developed blood glucose level prediction models using non-CGM data. However, their models used self-recorded blood glucose values in addition to other attributes such as insulin, meal, physical activity and etc. In the study by Li & Fernando (2016), the authors developed one universal model and compiled datasets of all patients using a clustering technique. Meanwhile, the authors Nguyen & Rokicki (2018), developed personalised models according to individual patients and averaged the error results, similar to this study. Thus, a direct comparison can be made with these studies to verify the performance of the models in this study.

The authors Nguyen & Rokicki (2018) have developed and tested the performance of several prediction techniques such as ARIMA, AVG, LSTM, Random Forest and etc. The lowest average RMSE of 7.57 was achieved for a Random Forest model enhanced with stability filters, in a 60 minutes prediction horizon. This proves that the LSTM model developed in this study is much better in terms of performance, as it was able to achieve a very significantly lower RMSE value compared to the previous study. Although the authors have tested the prediction performance using several prediction techniques, none of the results came close to what is achieved in this study. Furthermore, the previous studies have used multiple-input for their models, increasing the complexity of the model. In this study, only one single input attribute, namely, the self-recorded blood glucose values were used. This confirms that the prediction model developed in this study is less complex, can be easily implemented and provides more reliable prediction accuracy.

Table 5.8: Comparison of Previous Related Works with This Study

Literature	Model Type	Inputs	Prediction Technique	RMSE (mmol/L)	
				Prediction Horizon	
				30 mins	60mins
Li & Fernando (2016)	Universal Model	Blood glucose record, insulin, meal, exercise, sleep	SVM	3.82	
			Decision Tree	2.28	
			Random Forest	2.20	
			Pooled Panel Data	2.19	

			Pre-clustered (2)	2.04	
			Pre-clustered (3)	1.87	
			Pre-clustered (5)	1.84	
			Pre-clustered (9)	1.73	
			Pre-clustered (43)	1.53	
Nguyen & Rokicki (2018)	Personalised Model	Blood glucose record, carbohydrate, insulin, physical activity	Simple Baselines		25.71
			AVG		12.96
			Context AVG		12.53
			ARIMA		13.88
			LSTM (10 iter)		10.41
			LSTM (100 iter)		19.24
			RF		12.05
			Extra Trees (ET)		12.15
			RF + Sanity Filter		8.80
			ET + Sanity Filter		9.01
			RF + Sanity + Stability Filter		8.71
			RF + Stability Filter		7.57
This study (2019)	Personalised Model	Blood glucose record	LSTM	2.87	2.57

5.5 Summary

This chapter presented the findings and performance results of models developed in this study. The performance of the developed personalised blood glucose models and the results achieved for each model were critically analysed. A comparative analysis was performed at the end and the findings from the analysis were discussed in detail. The reasons and causes behind the certain performance behaviour of the models were also identified. Finally, the performance of the model achieved in this study was compared with previous related works in the field. A detailed comparative analysis was performed among them and the findings were presented in this chapter.

CHAPTER 6

DISCUSSION AND CONCLUSIONS

6.1 Introduction

This chapter summarizes the overall study. Each phase of this study, namely the data pre-processing, model implementation and evaluation are briefly discussed and conclusions are formed based on the results achieved in this study. The findings from this study and its contribution to achieving the objectives of this study are analysed. The importance of this study and its contribution to the research field and society in general is discussed in detail. Future recommendations are also provided in this chapter to further enhance the contribution of this study and to ensure this study becomes useful for the general population.

6.2 Discussion and Conclusions

This study was conducted using data collected from six diabetic patients who have manually self-recorded their blood glucose level data on random time and day at their convenience. The amount of data collected varied among each research subject. The key attributes selected from this data and inputted into the model are the date, time and the blood glucose record attributes. Each dataset was pre-processed extensively to prepare the data to be in a fit condition prior to inputting them into the prediction model. Tasks such as data transformation, data standardization, treatment of outliers, treatment of missing values and etc were carried out during this the pre-processing stage. One of the major issues identified in this data was the huge amount of missing values, as it was not a continuously logged data. Group means imputation was adopted to address the issue partially. And the rest of the missing values were made to be handled by the chosen prediction model for this study, the LSTM model. Exploratory data analysis was also performed to identify issues in data, gain a better understanding of data and to generate some insights that were useful during the analysis of model results.

The fully processed and prepared dataset was then split into 80 percent of training and 20 percent of test observations. The training dataset was used in the training of the LSTM model. LSTM model was specifically selected for this study for its capabilities in capturing time-dependencies in the data and finding their relationship to the model output. Hence, six individual LSTM models were developed in this study and each of this model was trained using one research subject's data. During the development phase, hyperparameter tuning was

performed to test and improve the model's performance capabilities. The testing data was then supplied to the finalized model and used to predict the blood glucose levels for each of the research subjects in 30 and 60 minutes prediction horizons.

The results obtained from these models were analysed individually and collectively. Individual model results were interesting and produced many useful insights for this study. One model in particular, of Patient P07, achieved outstanding results with very low RMSE values of 0.76 and 0.91 for 30 minutes and 60 minutes prediction respectively. Other patients' models achieved RMSE values in the range of 2.60 to 3.70. The performance results were aggregated and on average this study achieved RMSE value of 2.87 and 2.57 for 30 minutes and 60 minutes prediction respectively. An in-depth analysis of this result showed that the LSTM prediction model was able to perform well when data supplied into the model consisted of blood glucose records for every hour of the day compared to having a high frequency of records on certain times of the day. The key advantage of this model is that records for every hour of the day does not have to be a continuous record. Instead, it could be records collected randomly on any day throughout any period of time. This is a key finding, as it mainly solves the challenges as discussed in the problem statement of this study. Even the patients who do not use CGM for self-monitoring of blood glucose levels would be able to predict their blood glucose level with the support of this model.

Furthermore, it was also confirmed through a comparative analysis of each model that, the amount of data supplied did not have a significant contribution to the model performance. The LSTM model was able to provide reliable prediction even with small amount of data as long as the data consisted of inputs for different times of the day for the model to understand and learn the patient's blood glucose level behaviour. Thus, it can also be said that the LSTM model successfully handled the long-term temporal dependencies in the time series data. And they were also able to utilize the patterns of missing data to achieve better prediction results.

Finally, the results achieved from this study were compared with previous related research works in the field. It was established that the performance of models developed in this study is better than the benchmarked models. A very significantly lower RMSE value was achieved in this study. Furthermore, besides providing a more reliable prediction, the prediction model developed in this study, was found to be less complex and can be easily implemented as it only requires one type of input data which is the self-monitored blood glucose values. With proper

implementation, this model could be very successful as it is a very convenient option for diabetic patients who do use CGM devices for monitoring blood glucose levels.

6.3 Importance and Contributions of the Study

This study on blood glucose level prediction yielded two key contributions. Firstly, prediction of blood glucose level using only non-CGM data was implemented in this study and this has not been attempted and addressed in past studies. Although prediction of blood glucose levels has been explored by many researchers in the past, majority of the works relied on CGM data as the main input for their models as its more easily available and provides a very clear pattern of patients behaviour. The two previous works that explored the use of non-CGM data also relied on many other inputs such as meal information, physical activity, insulin consumption, sleep and etc which is inconvenient for patients to self-monitor and record this information constantly. However, even with all these additional information supplied to the model, the prediction outcome was not better than this study. This study has confirmed that reliable prediction can be produced with only one input, delivering a less complex and easily implemented model for the society in general.

Secondly, this study successfully addressed the issue of missing values and long term temporal dependencies in data using the LSTM model. It was found that a patient could record their blood glucose levels at random times of the day for any period of time. And as long as every time points within a day can be covered even with just one record, a reliable prediction can be produced by the model. The model is proven that it is not data-hungry, though more data supplied can contribute to better accuracy. Thus, with this contribution, even patients who do not use CGM for self-monitoring of blood glucose levels would be able to predict their blood glucose level effectively.

In summary, with proper implementation in future, this prediction model can serve as a cost-effective and a very convenient solution for the majority of diabetics patients around the world who do not use a CGM device for monitoring of blood glucose levels. Having better control of blood glucose levels can prevent calamities and greatly help in ensuring the patient's safety.

6.4 Future Recommendations

It is recommended that a mobile application for blood glucose level prediction is developed using the prediction model from this study. The mobile application should provide an interactive interface for users to input their own self-monitored blood glucose level records. The application should be able to input this data into the model and deliver a personalised prediction for patients anytime and anywhere at their convenience.

REFERENCES

- Ben Ali, J., Hamdi, T., Fnaiech, N., Di Costanzo, V., Fnaiech, F. & Ginoux, J.-M. (2018). Continuous blood glucose level prediction of Type 1 Diabetes based on Artificial Neural Network. *Biocybernetics and Biomedical Engineering*. [Online]. 38 (4). p.pp. 828–840. Available from: <https://remote-lib.ui.ac.id:2053/science/article/pii/S020852161830127X>.
- Balakrishnan, N.P., Samavedham, L. & Rangaiah, G.P. (2013). Personalized mechanistic models for exercise, meal and insulin interventions in children and adolescents with type 1 diabetes. *Journal of Theoretical Biology*. 357. p.pp. 62–73.
- Bremer, T. & Gough, D.A. (1999). Is blood glucose predictable from previous values? A solicitation for data. *Diabetes*. 48 (3). p.pp. 445–451.
- Bunescu, R., Struble, N., Marling, C., Shubrook, J. & Schwartz, F. (2013). Blood Glucose Level Prediction Using Physiological Models and Support Vector Regression. *2013 12th International Conference on Machine Learning and Applications*. [Online]. p.pp. 135–140. Available from: <http://ieeexplore.ieee.org/document/6784600/>.
- Che, Z., Purushotham, S., Cho, K., Sontag, D. & Liu, Y. (2018). Recurrent Neural Networks for Multivariate Time Series with Missing Values. *Scientific Reports*. [Online]. 8 (1). p.pp. 1–12. Available from: <http://dx.doi.org/10.1038/s41598-018-24271-9>.
- Contreras, I., Bertachi, A., Biagi, L., Oviedo, S. & Vehí, J. (2018). Using grammatical evolution to generate short-term blood glucose prediction models. *CEUR Workshop Proceedings*. 2148. p.pp. 91–96.
- Contreras, I., Oviedo, S., Vettoretti, M., Visentin, R. & Vehí, J. (2017). Personalized blood glucose prediction: A hybrid approach using grammatical evolution and physiological models. *PLoS ONE*. 12 (11). p.pp. 1–16.
- Diabetes, D.O.F. (2012). Diagnosis and classification of diabetes mellitus. *Diabetes Care*. 35 (SUPPL. 1).
- Fiorini, S., Martini, C., Malpassi, D., Cordera, R., Maggi, D., Verri, A. & Barla, A. (2017). Data-driven strategies for robust forecast of continuous glucose monitoring time-series. *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*. (July 2017). p.pp. 1680–1683.
- Gamboa, J.C.B. (2017). *Deep Learning for Time-Series Analysis*. [Online]. Available from: <http://arxiv.org/abs/1701.01887>.
- Gani, A., Gribok, A. V., Lu, Y., Ward, W.K., Vigersky, R.A. & Reifman, J. (2011). Universal Models For Predicting Glucose Concentration In Humans. *WO Patent WO/2010/*. 14 (1).

p.pp. 157–165.

- Georga, E.I., Protopappas, V.C. & Fotiadis, D.I. (2011). Glucose Prediction in Type 1 and Type 2 Diabetic Patients Using Data Driven Techniques. *Knowledge-Oriented Applications in Data Mining*. [Online]. (May 2014). Available from: <http://www.intechopen.com/books/knowledge-oriented-applications-in-data-mining/glucose-prediction-in-type-1-and-type-2-diabetic-patients-using-data-driven-techniques>.
- Hamdi, T., Ben Ali, J., Di Costanzo, V., Fnaiech, F., Moreau, E. & Ginoux, J.M. (2018). Accurate prediction of continuous blood glucose based on support vector regression and differential evolution algorithm. *Biocybernetics and Biomedical Engineering*. [Online]. 38 (2). p.pp. 362–372. Available from: <https://doi.org/10.1016/j.bbe.2018.02.005>.
- Hidalgo, J.I., Colmenar, J.M., Kronberger, G., Winkler, S.M., Garnica, O. & Lanchares, J. (2017). Data Based Prediction of Blood Glucose Concentrations Using Evolutionary Methods. *Journal of Medical Systems*. 41 (9).
- Hochreiter, S. & Jürgen Schmidhuber, J. (1997). LONG SHORT-TERM MEMORY. *Neural Computation*. [Online]. 9 (8). p.pp. 1735–1780. Available from: <http://www7.informatik.tu-muenchen.de/~hochreit%0Ahttp://www.idsia.ch/~juergen>.
- IDF (2017). *IDF Diabetes Atlas*. 8th Editio. [Online]. Available from: www.diabetesatlas.org.
- Jebb, A.T., Parrigon, S. & Woo, S.E. (2017). Exploratory data analysis as a foundation of inductive research. *Human Resource Management Review*. [Online]. 27 (2). p.pp. 265–276. Available from: <http://dx.doi.org/10.1016/j.hrmr.2016.08.003>.
- Lehmann, E.D. & Deutsch, T. (1998). Compartmental models for glycaemic prediction and decision-support in clinical diabetes care: Promise and reality. *Computer Methods and Programs in Biomedicine*. 56 (2). p.pp. 193–204.
- Li, J. & Fernando, C. (2016). Smartphone-based personalized blood glucose prediction. *ICT Express*. [Online]. 2 (4). p.pp. 150–154. Available from: <http://dx.doi.org/10.1016/j.icte.2016.10.001>.
- Mougiakakou, S., Prountzou, K. & Nikita, K. (2005). A real time simulation model of glucose-insulin metabolism for type 1 diabetes patients. *Conference proceedings : ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Conference*. 1 (3). p.pp. 298–301.
- Nguyen, T.N. & Rokicki, M. (2018). *On the Predictability of non-CGM Diabetes Data for Personalized Recommendation. (arXiv:1808.07380v1 [cs.CY])*. [Online]. p.pp. 3–8. Available from: <http://arxiv.org/abs/1808.07380>.
- Oviedo, S., Vehí, J., Calm, R. & Armengol, J. (2017). A review of personalized blood glucose

- prediction strategies for T1DM patients. *International Journal for Numerical Methods in Biomedical Engineering*. 33 (6).
- Plis, K., Bunesco, R., Marling, C., Shubbrook, J. & Schwartz, F. (2014). A Machine Learning Approach to Predicting Blood Glucose Levels for Diabetes Management. *Modern Artificial Intelligence for Health Analytics*. p.pp. 35–39.
- Rollins, D.K., Bhandari, N., Kleinedler, J., Kotz, K., Strohbehn, A., Boland, L., Murphy, M., Andre, D., Vyas, N., Welk, G. & Franke, W.E. (2010). Free-living inferential modeling of blood glucose level using only noninvasive inputs. *Journal of Process Control*. [Online]. 20 (1). p.pp. 95–107. Available from: <http://dx.doi.org/10.1016/j.jprocont.2009.09.008>.
- World Health Organization (WHO) (2016). Diabetes country profiles (Malaysia). *World Health Organization*. [Online]. p.p. 2016. Available from: <http://www.who.int/diabetes/country-profiles/en/#M>.
- Zecchin, C., Facchinetti, A., Sparacino, G. & Cobelli, C. (2014). Jump neural network for online short-time prediction of blood glucose from continuous monitoring sensors and meal information. *Computer Methods and Programs in Biomedicine*. [Online]. 113 (1). p.pp. 144–152. Available from: <http://dx.doi.org/10.1016/j.cmpb.2013.09.016>.

APPENDIX A

ETHICAL APPROVAL OF RESEARCH PROJECT

APPENDIX B

TURNITIN SIMILARITY REPORT

APPENDIX C

LOG SHEETS FOR SUPERVISORY SESSION