# Main Points and Structure (around 2000 words)

*Dissertation Title for Reference - Blood Glucose Levels Analysis and Predictions Using Data Mining Techniques*

* Data mining
  + What is it?
  + What are its uses?
  + Different types of data mining techniques
* Data mining in the medical field especially diabetes
  + Mining techniques most used in medical data mining *(paper 2)*
  + Short overview of diabetes (what it is, its symptoms and effects …)
  + Briefly explain CGMs (and their use for diabetic people) and how data can be mined and used to make predictions
* Prediction Algorithms using Diabetes Data
  + Classification Algorithms *(paper 1)*
  + Results obtained by some papers using specific algorithms
  + Best algorithms chosen for scenario

# Literature Review

## Section 1 – Introduction to Data Mining

Data mining is the process of analysing pieces of data and processing it into useful information. It is the process of exploring hidden knowledge from large amounts of data in search of consistent patterns and meaningful relationships between variables.

Data mining is part of the Knowledge Discovery Process (KDP). Before data mining algorithms can be applied, data from varies sources gets integrated into a single data store called target data, it is them pre-processed and transformed into standard format. At this point data mining algorithms can be used to process the data and produce patterns or rules. The output can now be interpreted into useful knowledge or information.

Over the past years several data mining techniques have been developed the most commonly used ones being Association, Classification, Clustering, Prediction and Sequential Patterns.

Many industries are using these tools to extract recurring tendencies in order to predict behaviours and future trends, allowing businesses to make intelligent and productive decisions. In the medical field, data mining can be used to make predictions on patients based on their records to be able to predict conditions and symptoms related to some conditions in advance therefore, thanks to an early diagnosis, precautions can be taken to prevent future complications.

## Section 2 - Data Mining in the Medical field specifically Diabetes

The Medical field has enormous amounts of data and because of this, medical data mining has grown in its popularity over the past years. While classification is the technique generally used in medical data mining, other techniques that have proven useful in finding patterns in medical data are clustering, association and outlier.

In the health care sector data mining is especially useful as it uses medical data for analysis to offer improved health care at better costs. Data mining in health care plays a significant role in prediction and diagnosis of several health problems such as heart disease, diabetes, cancer and skin disease among many others. This is important since early diagnoses of diseases and medical conditions are vital in order to prevent future complications.

### *Diabetes*

***Note:*** *This section explains Diabetes, it’s symptoms, characterises and possible side effects*

Medical data mining in regards to diabetes has become quite a researched subject as it is a chronic disease that affects millions of people worldwide. In fact, according to WHO (World Health Organization) and the International Diabetes Foundation, it has become one of the leading causes of death worldwide. Statistics show that in 2017 it caused around 4 million deaths and affected about 1 in every 7 births. It is also a leading cause of obesity and it has a significant impact on the quality of life of people suffering from diabetes and their families, especially when complications arise.

Diabetes, otherwise known as Diabetes Mellitus occurs when the pancreas is unable to produce enough insulin, or when the body cells cannot make proper use of the insulin produced, due to reduced sensitivity. Without the ability to produce or effectively use insulin, glucose (sugars found in carbohydrates) levels in the blood are elevated (this is known as hyperglycaemia). Characteristically a person is diagnosed with diabetes because of high levels of sugar in their blood. A person is considered diabetic when blood glucose (BG) levels are generally higher than 11mmol/L or higher then 7mmol/L when fasting. Symptoms associated with diabetes include unusual thirst, weight change, lack of energy, blurred vision and frequent or recurring infections as the body takes longer to heal. Diabetes can lead to many complications like heart disease, kidney failure, blindness and amputation. However, this can be prevented with an early diagnosis.

There are three main types of diabetes Type 1, Type 2 and Gestational. Type 1 is an autoimmune disease occurring at a young age of below 20 years, here the pancreatic cells that produce insulin have been destroyed. Type 2 occurs when various organs of the body become insulin resistant, and this increases the demand for insulin. At this point, the pancreas doesn’t make the required amount of insulin. This usually occurs in age groups above 40. Gestational diabetes occurs during pregnancy, as the pancreas does not produce enough insulin.

### *Diabetes Management and Technologies*

***Note:*** *This section explains diabetes management and some technologies that help diabetics with this   
 process*

As soon as a person is given the diagnosis, it is very important to manage and keep track of their diabetes. This means that in many cases diets need to be changed, exercising becomes more important and glucose levels need to be checked daily, at least 3-4 times. While changing your diet, adjusting to medication and including some exercising routine might be though at first, the most tedious part is having to check the BG. Unfortunately, this is also very important as it shows the patients progress and reactions to the lifestyle changes as well as any changes that might happen that effectively can result is a change of medication or treatment.

Traditionally, BG is checked by using a small needle to prick a finger and extract a small amount of blood. Then using a Blood Glucose Meter, the blood gets absorbed by testing strip attached to the meter, which in return produces the BG reading from the blood sample. It is important that these results are written down somewhere as doctors can determine progress based on those readings.

A less invasive technique of monitoring BG is Continuous glucose monitoring (CGM). This is not technically BG monitoring as the sensors with a CGM machine are placed into the patient’s body but not into the bloodstream. The sensors measure the glucose in the interstitial fluid found around the body’s cells. This sensor is placed under the skin through a needle and a transmitter is placed on the skin where the sensor was inserted. As the name implies this system checks the glucose levels continuously as the sensor is always there. With most systems the results can be viewed on a mobile application and are stored in a database. This might still sound invasive, however it is less so compared to finger pricks as some sensors can last up to 13 days before they start failing. Complete non-invasive systems are currently being researched or tested and some are even expected to be readily available within the next two years.

Whichever method mentioned above is chosen, keeping record of glucose readings is very important.

### *Data Mining Diabetic Data*

***Note:*** *This section explores medical data mining using data from diabetic patients and what predictions can   
 be made from it.*

As explained in [4], data mining in health care is being used to offer improved healthcare at lower costs. It also plays a significant role in prediction and diagnosis of various health problems including diabetes.

In many papers the most common analysis made using diabetic data is predicting weather or not a person is diabetic based on multiple variables. From there statistic, such as what is the most common age group for people with diabetes or which gender is most likely to have diabetes, are calculated as can be seen in papers [4], [6] and [7], where the authors used Classification Techniques.

As mentioned previously, for people who have diabetes keeping record of their glucose levels is very important. Many people do this especially if they use a CGM since it keeps record of the data automatically. Ideally along with the glycaemic index and the time it was taken other variables such as food consumption, physical activity, stress levels and sleeping hours are also recorded. With this amount of data there are infinitely more possibilities for prediction models and analysis that can go beyond diagnosing diabetes and be able to help the diagnosed individual to control their condition. This can be seen in [11] where researchers sate that ‘dietary intake is a central determinant of BG levels. Different food effects individual bodies in a different way, therefor if two people eat the exact same meal and check their glucose levels two hours later, the readings could be very different. The study presented in [11] uses such records to try and predict what they call the Postprandial (post meal) Glycaemic Response (PPGR) based on individual’s records of Glucose levels and the Food consumed in order to create a personalised nutritional plan.

Papers [11] – [18] all take advantage of CGM technologies and apply algorithms to using the data from the CGMs in order to predict short-term BG. Since prevention is better than cure, most models also use meal information and aim to predict cases of hyperglycaemia, which can lead to long-term complications, and hypoglycaemia which can cause a coma.

If these algorithms can be refined and show a high level of accuracy they could be utilized in semi closed-loop devices such as CGMs for guiding therapy in diabetes patients. This can lead to better glycaemic control and hence reducing the risk of complications [13].

## Section 3 - Prediction Algorithms used for Diabetic Data

***Note:*** *This section explores different Algorithms that are used to make predictions using diabetic data as well   
 as the results obtains from previous studies.*

The most common technique used in medical data mining is Classification. There are many classification techniques, however the ones that are generally used include Decision trees, Bayesian classifier, Random Forest, Random tree, classification by back-propagation and rule based classifiers. Classification is performed in two steps; Model construction and Model Usage. In the first step the prediction model is built using appropriate algorithm. Next, the prediction model is applied to actual data and prediction is done accordingly [4].

When it comes to prediction and diagnoses of diabetes the most common techniques used are decisions trees, k-nearest neighbour, SVMs and Naïve Bayes, however, when it comes to short-term BG prediction, Neural Networks seem to be quite popular and effective.

In [11] decision trees are used for predicting the personalised post meal glycaemic response. In this research the trees are inferred sequentially, with each tree trained on the residual of all previous trees and make a small contribution to the overall prediction. The features of each tree are selected by an inference procedure from 137 features representing meal content (e.g. energy, macronutrients, micronutrients); daily activity (e.g., meals, exercises, sleep times); blood parameters (e.g. HbA1c%, HDL cholesterol); CGM-derived features; questionnaires; and microbiome features. For this scenario decision trees worked as expected as with the increase of carbohydrate content in meals, their algorithm predicted a higher post meal glycaemic response which is what realistically happens.

Paper [13] uses Artificial Neural Networks (ANN) for predictions. The ANN models performed well when predicting at normal (>3.9 and <10 mmol/L) and hyperglycaemic ranges (≥10 mmol/L); however, glucose concentrations in areas of hypoglycaemia were commonly overestimated. Although it is stated that this might be due to the minimal occurrences of hypoglycaemic events within the training data. The results also indicate that an increase in predictive window leads to a decrease in predictive accuracy of the ANN model. It is hypothesized that the underestimation of hyperglycaemic extremes is due to the extension of the predictive window and the associated inability of the neural network to determine oscillations and trends in glycemia as well as the occurrence of other relevant variables that where not taken into consideration such as emotional states, insulin dosages, and meals, which may occur within the predicted time window and can affect the ANN weights.

Another study that uses ANNs to predict BG levels can be found in paper [14] where they applied simulated data and real time-series data to a new ANN approach. The performance of their new prediction algorithm was tested on 5 virtual patients generated in silico via a Type 1 Diabetes simulator and on one real patient. Their results have proven to be better than 2 other published studies [17] and [18], that use the same data.

## References

1. Kalyani M Raval, “Data Mining Techniques”, *International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 10, October 2012*
2. IDF (International Diabetes Federation) Diabetes Atlas - 8th Edition 2017
3. K. Saravananathan and T. Velmurugan, “Analyzing Diabetic Data using Classification Algorithms in Data Mining”, *Indian Journal of Science and Technology, Vol 9(43), DOI: 10.17485/ijst/2016/v9i43/93874, November 2016*
4. S. R. Priyanka Shetty, Sujata Joshi, “A Tool for Diabetes Prediction and Monitoring Using Data Mining Technique”, *International Journal of Information Technology and Computer Science(IJITCS), Vol.8, No.11, pp.26-32, 2016. DOI: 10.5815/ijitcs.2016.11.04*
5. Joseph L. Breault, Colin R. Goodall and Peter J. Fos, “Data mining a diabetic data warehouse”, *Artificial Intelligence in Medicine 26 (2002) 37–54*
6. Aiswarya Iyer, S. Jeyalatha and Ronak Sumbaly, “Diagnosis of Diabetes Using Classification Mining Techniques”, *International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.5, No.1, January 2015*
7. Pardha Repalli, “Prediction of Diabetes Using Data Mining Approach”
8. Saman Hina, Anita Shaikh and Sohail Abul Sattar, “Analyzing Diabetes Datasets using Data Mining”, *Journal of Basic & Applied Sciences, 2017, 13, pgs. 466-471*
9. Eleni Georga, Vasilios Protopappas, Alejandra Guillen, Giuseppe Fico, Diego, Ardigo, Maria Teresa Arredondo, Themis P. Exarchos, Demosthenes Polyzos and Dimitrios I. Fotiadis, “Data Mining for Blood Glucose Prediction and Knowledge Discovery in Diabetic Patients: The METABO Diabetes Modeling and Management System”, *31st Annual International Conference of the IEEE EMBS Minneapolis, Minnesota, USA, September 2-6, 2009*
10. Michael L. Raymer, Travis E. Doom, Leslie A. Kuhn, and William F. Punch, “Knowledge Discovery in Medical and Biological Datasets Using a Hybrid Bayes Classifier/Evolutionary Algorithm”, *IEEE Transactions on Systems, Man, And Cybernetics—Part B: Cybernetics, Vol. 33, No. 5, October 2003*
11. Zeevi et al, “Personalized Nutrition by Prediction of Glycemic Responses”, *Cell 163, pgs. 1079–1094, November 19, 2015*
12. Riccardo Bellazzi, Ph.D. and Ameen Abu-Hanna, Ph.D., “Data Mining Technologies for Blood Glucose and Diabetes Management”, *Journal of Diabetes Science and Technology Volume 3, Issue 3, May 2009*
13. Scott M. Pappada, B.S., Brent D. Cameron, Ph.D., Paul M. Rosman, D.O., F.A.C.P., F.A.C.E., “Development of a Neural Network for Prediction of Glucose concentration in Type 1 Diabetes Patients”, *Journal of Diabetes Science and Technology Volume 2, Issue 5, September 2008*
14. C. Zecchin, A. Facchinetti, G. Sparacino, G. De Nicolao, and C. Cobelli, “A New Neural Network Approach for Short-Term Glucose Prediction Using Continuous Glucose Monitoring Time-Series and Meal Information”, *33rd Annual International Conference of the IEEE EMBS, Boston, Massachusetts USA, August 30 - September 3, 2011*
15. C. Gireesh1, V. Punna Rao, “Blood Glucose Prediction Algorithms for Hypoglycemic and/or Hyperglycemic Alerts”, *IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 5, No 3, September 2012*
16. Francesca Zanderigo Ph.D., Giovanni Sparacino Ph.D., Boris Kovatchev Ph.D. and Claudio Cobelli Ph.D., “Glucose Prediction Algorithms from Continuous Monitoring Data: Assessment of Accuracy via Continuous Glucose Error-Grid Analysis”, *Journal of Diabetes Science and Technology Volume 1, Issue 5, September 2007*
17. Giovanni Sparacino, Francesca Zanderigo, Stefano Corazza, Alberto Maran, Andrea Facchinetti, and Claudio Cobelli, “Glucose Concentration can be Predicted Ahead in Time From Continuous Glucose Monitoring Sensor Time-Series”, *IEEE Transactions on Biomedical Engineering, Vol. 54, No. 5, May 2007*
18. C. Perez-Gandia, A. Facchinetti, G. Sparacino, C. Cobelli, E. Gomez, M. Rigla, A. de Leiva, and M. Hernando, “Artificial neural network algorithm for on-line glucose prediction from continuous glucose monitoring”, *Diabetes Technology & Therapeutics, Volume 12, Number 1, pages 81–88, 2010*