Data-Driven Predictions of Glycemic Extremes Using Machine Learning

Adeoluwa Oyinlola (**4214719**) Glory Ndumbuchi Kekeocha (**4234927**)
Hamid Hussain (**4216526**) Chuku Lydia Chinasa (**4234877**)
Shayan Mudassa (**4237995**)

***Abstract*—Diabetes is a formidable and chronic ailment characterized by elevated blood sugar levels, presents severe complications such as Hyperglycemia and hypoglycemia if left untreated and undetected. These complications pose pervasive global risks, significantly impacting patient outcomes and safety. Continuous glucose monitoring systems have emerged as breakthrough advancements, enabling effective blood glucose level monitoring. However, the challenge remains in reducing or preventing diabetes-related complications. This paper proposes a machine learning approach to predict the likelihood of hyperglycemia or hypoglycemia occurrences based on time series data, anticipating a patient's predisposition to these conditions in the next 2 hours. Here, the system integrates input data from glucose sensors (CGM), carbohydrate intake (CHO), and insulin intake. Machine learning models explored in this study include Long Short-Term Memory (LSTM), Random Forest Classifier (RF), and Support Vector Machines, with LSTM and RF emerging as the best-performing model with an average accuracy of 94.7% and 96.3% respectively. This research aims to highlight the application of ML algorithms in predicting diabetes and aiding in the timely implementation of preventive measures and lifestyle changes to mitigate these health challenges.**

Index Terms— Machine Learning, Hypoglycaemia, Hyperglycaemia, Blood glucose Level, Diabetes.

I. Abbreviations and Acronyms

The following section provides a concise guide to the acronyms and abbreviations pertinent to this discourse. Familiarity with these terms will enhance clarity and facilitate a comprehensive understanding of the subsequent discussions.

- DT: Decision Tree
- SVM: Support Vector Machine
- HBGI: High Blood Glucose Index
- LBGI: Low Blood Glucose Index
- RNN: Recurrent Neural Network
- BG: Blood Glucose
- CGM: Continuous Glucose Monitoring
- CHO: Carbohydrate
- R.F: Random Forest
- LSTM: Long Short-Term Memory
- SMBG: Self-Monitoring of Blood Glucose
- DM: Diabetes Mellitus

II. Introduction

Diabetes collectively refers to a class of metabolic diseases defined by elevated blood glucose levels. The incidence of diabetes has been rising in developed as well as developing nations in recent years. In 2015, 415 million people world-wide were believed to have diabetes [1]. Inadequate management of diabetes can result in severe complications impacting the kidneys, eyes, peripheral and autonomic nervous systems [2]. *Fig. 1* shows in percentages the global estimates of diabetes and projections for 2045 [3].

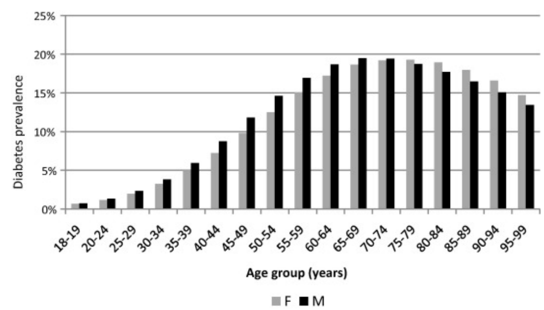

Fig. 1: Global estimates of diabetes prevalence (%) and projections for 2045

Common severe complications linked to diabetes include Hyperglycemia and Hypoglycemia. Hyperglycemia is the medical term used for elevated blood glucose (sugar) levels. It occurs from the body's inability to efficiently regulate blood sugar levels, resulting in high concentrations of glucose in the bloodstream. Hyperglycemia can cause lower limb occlusions, nerve injury, and further consequences such as decay, necrosis, and localized or total foot gangrene, which occasionally require amputation [4]. On the other hand, Hypoglycemia is brought on by extremely low blood sugar (glucose) which affects the brain which depends only on glucose for survival [5]. In those with diabetes mellitus, hypoglycemia can cause symptoms that vary from minor, like palpitations and anxiety, to severe, such seizures, coma, and possibly even death [4]. Managing the occurrence of hypoglycemia and hyperglycemia in people with diabetes poses a formidable challenge, and requires continuous blood glucose monitoring [6].

Regular monitoring multiple times a day in order to reduce the possible harm caused by diabetic complications is essential. The use of prediction models empowers individual with diabetes to properly manage the way they control their blood glucose levels by anticipating future values [7]. This paper introduces a methodology employing machine learning techniques to precisely predict the likelihood of hypoglycemia or hyperglycemia occurring within the subsequent two hours. The aim of this study is to construct a model that analyzes past data, including glucose levels, carbohydrate intake, and insulin intake, to predict whether a subject with type I diabetes will encounter hyperglycemia or hypoglycemic events in the coming two hours. The analysis is based on a simulated dataset featuring information from 22 subjects diagnosed with type I diabetes.

In the subsequent sections of this study, we have presented a review of state-of-the-art literatures concentrating on works pertinent to our study. The next section explores the methodology, providing an intricate overview of the steps taken during data pre-processing, the machine learning models selected for predictions, and the testing and validation strategies employed. The last two sections include a discussion of results illustrating the outcomes and performance metrics of the implemented machine learning pipeline across various models and a critical assessment of the viability of our proposed solution.

III. Background

A. Type I and Type II Diabetes

Diabetes Mellitus (DM) encompasses a cluster of metabolic disorders marked by persistent hyperglycemia due to inadequate insulin secretion, impaired insulin action, or a combination of both factors [8]. Type 1 diabetes stems from an autoimmune response targeting proteins within the islet cells of the pancreas [9]. In this condition, the body's immune system erroneously attacks and destroys these vital cells responsible for insulin production, leading to a deficiency in insulin. *Fig. 3* shows the difference between Type 1 diabetes and Type 2 diabetes [10].

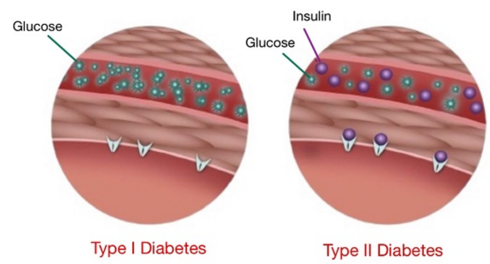

Fig. 2: Type 1 and Type 2 diabetes

On the other hand, Type 2 diabetes is a multi-faceted ailment influenced by a blend of genetic predispositions, insulin insufficiency, and insulin resistance [11]. Genetic factors play a pivotal role in the susceptibility to Type 2 diabetes, impacting the body's ability to produce adequate insulin and respond effectively to its actions. Lifestyle elements such as obesity, overeating, stress, and the natural aging process can exacerbate insulin resistance, further exacerbating the condition.

B. Hypoglycaemia and Hyperglycaemia

While determining precise rates remains challenging, the frequency of hypoglycemia differs significantly between type 2 and type 1 diabetes, with a notably lower incidence in the former [12]. Individuals managing type 1 diabetes routinely grapple with the reality of hypoglycemic episodes as they endeavor to enhance or sustain glycemic control. The recurrent experience of asymptomatic hypoglycemia underscores the inherent challenges faced [13].

Hypoglycemia is not merely an inconvenience but a potential emergency, signaling the central nervous system's struggle to meet its energy demands. [14]. The gravity of the situation extends further, with untreated hypoglycemia carrying the potential for lasting neurologic damage or, in severe cases, fatalities. *Fig. 4* shows the the difference between an hypoglycemic blood and normal blood. *Fig. 4* shows three different levels of blood glucose [15].

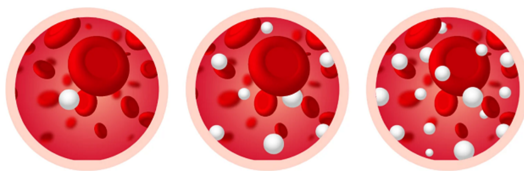

Fig. 3: Diagram showing various Blood Glucose levels

Contrary to hypoglycemia with its neuroglycopenic effects, exploring the ramifications of hyperglycemia on cognitive-motor performance faces challenges [16] due to the lack of a well-established physiological mechanism explaining how elevated blood glucose levels impact brain function negatively [17]. The intricate pathways through which hyperglycemia influences cognitive-motor performance remain elusive, despite its recognized association with cognitive impairments, posing a continual challenge in diabetes research [18].

IV. Literature Review

According to a paper published in 2023, the study focused on developing predictive models for blood glucose variations in type 2 diabetes mellitus (T2DM) patients by utilizing sixteen machine learning algorithms such as Logistic Regression, Decision Tree, Random Forest, and XGBoost for training. However, the ultimate outcomes were achieved through ensemble learning and modified random forest inputting [19]. This approach highlighted the importance of considering fasting blood glucose (FBG) and glycosylated haemoglobin (HbA1c) as key indicators, with an additional focus on factors like medication compliance, follow-up outcomes, dietary habits, BMI, and waist circumference.

Another study aimed to improve the safety of Type 1 Diabetes Mellitus (T1DM) patients under sensor-augmented pump (SAP) therapy by designing and evaluating an early warning system (EWS) for hypoglycemic and hyperglycemic episodes. A warning algorithm and data-driven online adaptive prediction models were combined by the EWS. Three modelling strategies were evaluated: recurrent neural network (RNN) models, auto-regressive with an output correction module (cARX) models, and autoregressive (ARX) models. After processing the model outputs, the warning algorithm sends out notifications for impending occurrences. 23 T1DM patients receiving SAP treatment participated in the assessment; and the hybrid cARX/RNN-based EWS had a superior performance. It attained an accuracy level of 100.0% in predicting hypoglycemic and hyperglycemic events with a daily false alarm rate of 0.8 and a median detection time of 16.7 minutes in advance. When cARX and RNN models were combined, they performed better than when they were used separately, resulting in accurate, timely event prediction with few false alarms [20]. Using advanced warning algorithms and real-time adaptive, data-driven glucose prediction models, this integrated EWS provides T1DM patients with enhanced safety and comfort.

Despite the availability of applications like The Diabetes: M app and the Nightscout platform, which aid in diabetes management and real-time access to glucose levels, they lack predictive functions and warning capabilities. The Diabetes: M app serves as a diabetic diary, while the Nightscout platform, despite supporting advanced glucose measurement methods, lacks warnings for exceeding normoglycemia levels. In response to these limitations, the authors of this article employed classification algorithms, including K-nearest neighbors (KNN), Naïve Bayes Classifier, and logistic regression, to assess the likelihood of hyper or hypoglycemia based on patient-collected data. The studies, conducted over a 3-month period with individuals having type 1 diabetes, incorporated parameters such as measurement time, pre-meal blood glucose levels, insulin dosage, meal carbohydrate content, and pre-meal physical activity [21]. To ensure real-time warnings, the algorithms were tested on smartphones using the Android Profiler tool. Performance tests revealed the significant impact of the training dataset size on operational time, with the KNN algorithm standing out as the most effective solution due to its non-parametric nature and adaptability to the unique characteristics of each diabetic individual. Despite real-world data not always meeting theoretical assumptions, the KNN algorithm demonstrated superior accuracy in both three-class and four-class studies. Table 1 shows the operational efficiency of the different diabetes samples on the Android Profiler tool.

	10 Samples	100 Samples	500 samples
CPU	13.9%	27.7%	28.8%
RAM	67.3MB	86.7MB	83.5MB
Operation	0.05s	0.32s	5.09s

TABLE I: Operational efficiency of Android Profiler tool

V. Methodology

This section demonstrates the various steps that have been taken from data preparation to modeling. We shall be doing a systematic and technical analysis of the methods that have been applied, presentation of relevant framework employed to solve the problem.

A. Data Understanding

Training the machine learning model and classifiers is a rather small step in building an efficient and intelligent model that can accurately predict the possibility of an hypoglycemic or hyperglycemic event in the next 2 hours, the real work lies in understanding and processing the data [22]. An exhaustive research to understand the intricacies of the data was done before any feature engineering or training were carried out. Taking time to understand what we are dealing with before thinking about modeling is a crucial step that has paved way for flexibility in the model and to building a robust model [23]. The dataset comprises 22 synthetic subjects diagnosed with type I diabetes. Their meal plans, insulin administration, and Continuous Glucose Monitoring (CGM) data were simulated using the FDA-approved UVA-Padova simulator. Sampling is conducted at 5-minute intervals, resulting in 288 samples per day [24]. The data cuts across 3 age groups: child, adolescent and adult. The task involves predicting hyperglycemic or hypoglycemic events 24 samples ahead. Table 2, shows the distribution of files in each class of data.

Each data in the dataset is a '.csv' file with time series columns: BG, CGM, CHO, insulin, LBGI, HBGI and Risk.

Class	No. of files	Time series data
Child	5	6
Adolescent	7	6
Adult	10	6

TABLE II: Distribution of data in each class

B. Data Preprocessing

Data preprocessing stage in machine learning refers to the various techniques and methods employed to make the dataset fit for training [25]. In data mining and machine learning in general, it is a common belief that 80% of once time should be dedicated to preprocessing and the remaining 20% for training [26]. As a result of the multiple samples related to one class of subject, the time series files were segmented into three different folders where each folder consists of files of the same class which were named: child, adolescent and adult. Each file was named according to the folder and concatenated with a 3 digit number which goes in an increasing order. For instance, a file in the adolescent dataset would be "adolescent_005". Since a separate classifier would be built for each class of data, we load the folder and preprocess each separately. The files in each folder were concatenated into one single file. After this, the time column was converted into the pandas datetime format after which time features (hours, minutes and second were extracted from the time column. To conclude the first section of data preprocessing, the 'Time' column was dropped as it will not be used for modeling. The next section of data preprocessing is Feature Engineering.

1) Feature Engineering: Feature engineering in machine learning and data analysis involves the necessary steps and techniques adopted to transform the original data in its raw form to create new features to improve the efficiency of the model in a specified task [27]. The main purpose of feature engineering is to increase the predictive ability of the model by providing it with more and relevant information [28]. A number of feature engineering techniques were carried on the original dataset.

1) Lag features: For CGM, CHO, and insulin which represent past values while capturing temporal patterns.
2) Rolling mean: For CGM, CHO, and insulin smooths time-series data by calculating the averages over a specified window(3), revealing underlying trends and patterns.
3) Interaction features: This is done by the product of CGM and insulin. It captures synergistic effects and relationships that may enhance predictive modeling.
4) Future CHO intake features: This is done with window size of 2 hours or 24 samples. It provides information about future carbohydrate intake, providing the model with knowledge of upcoming dietary choices for improved prediction.
5) Combined feature for whole past data and memory past: This integrates information from a subset of past observations of CGM, CHO and insulin, offering a holistic representation of historical patterns.

VI. Class of Machine Learning

The choice of machine learning model to use in this predictive analysis is a delicate one and as a result of this, adequate care was taken before selection was made. Recurrent Neural Networks (RNN) have been developed to handle ordinal or temporal dataset. A type of RNN, therefore, was used. Long Short Term Memory was the first choice. LSTM is a variant of RNN that can learn long term dependencies. *Fig. 5* shows a simple LSTM architecture. The LSTM model uses four layers to overcome the

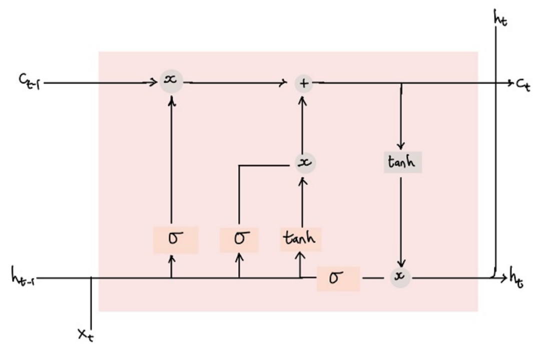

Fig. 4: LSTM with four cell states

vanishing gradient problem: forget gate (f), input gate (i), output gate (o), and value generated by tanh (g). The forget gate, f is the first section in the architecture. It decides the useful information.

$$f = \sigma(W_f h_{t-1} - U_f x_t) \tag{1}$$

An output close to 1 means the information is kept and when it is 0, the information can be forgotten. The second model used for comparative analysis was RF. Random forests consist of an ensemble of tree predictors called decision tress, where each tree relies on the values of a randomly sampled vector, chosen independently and following the same distribution across all trees within the forest [29]. It is important to note that while RF are excellent classifiers in machine learning, they are vastly prone to over-fitting. The last classifier that was used for training is Support Vector Machines. It is a type of supervised machine learning algorithm used for classification and regression problems. It works by finding the optimal hyperplane that separates classes in a feature space. It offers good accuracy, uses less memory because they use a subset of training points in the decision phase.

VII. Modeling

The first step before training the LSTM model was to specify the conditions for hypoglycaemia and hyperglycaemia. A new column named 'Target' is added to the dataframe and initialized with 1, representing the 'normal' condition. Conditions are then defined to identify instances of hypoglycemia and hyperglycemia based on the CGM values shifted by a specific time period (2 hours or 24 samples) backward. Hypoglycemia is identified when the CGM value is less than 70, and hyperglycemia is identified when the CGM value exceeds 180. The 'Target' column values are updated based on the conditions: set to 0 for hypoglycemia and 2 for hyperglycemia. To ensure no future CGM data is used for prediction, the CGM column is shifted by one time step backward, ensuring that the target values are assigned to the correct time points and no future CGM data is used for prediction. As a result of the shifting, the first row of the CGM column will contain a NaN value and this row was dropped. The 'Target' column was dropped from the combined dataframe and used as the output data or y. The input features were normalized using MinMax scaler to scale between 0 and 1 because of some large and extreme values while 'Target' column was one-hot encoded for easy computation. A window size of 10 was used for training the LSTM model and the dataset was splitted into 80-20 training and test set. *Fig. 5* is a flowchart of the proposed model used.

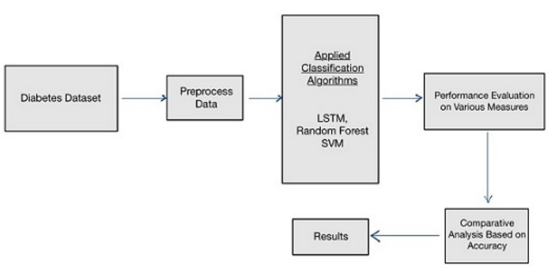

Fig. 5: Proposed Model diagram

The RF classifier and SVM follow the same preprocessing as the one used for the LSTM model. The results and evaluation are discussed extensively in the next section. *Fig. 6* shows the LSTM model architecture.

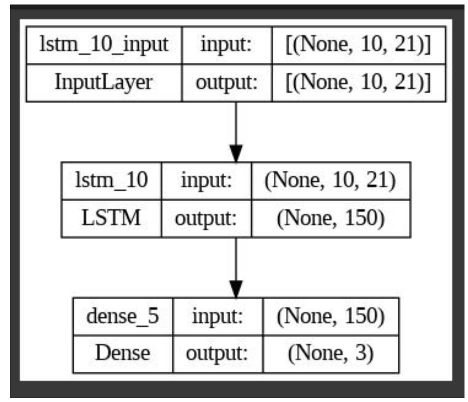

Fig. 6: LSTM model plot

VIII. Results and Evaluation

In this section, we present the evaluation results of the three models used for training. The results show the accuracy, precision, recall and f1-score of each model and for each class. Hypoglycaemia is 0, Normal blood sugar is 1, while hyperglycaemia is 2.

Mathematical formulars as metrics for evaluation,

$$Accuracy = \frac{TP + TN}{TP + FN + TN + FP} \tag{2}$$

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

$$f1 = 2 * \frac{precision * recall}{precision + recall} \tag{5}$$

Table 2, 3 and 4 show the accuracy, precision, recall and f1-score of the LSTM model, RF classifier and SVM respectively.

	Precision(%)	Recall (%)	f1-score (%)
0	95	96	95
1	96	95	95
2	93	95	94

TABLE III: LSTM for child data: Accuracy= 95.1%

	Precision(%)	Recall (%)	f1-score (%)
0	97	97	97
1	96	96	96
2	91	93	92

TABLE IV: RF for child data: Accuracy=96.3%

	Precision(%)	Recall (%)	f1-score (%)
0	74	77	76
1	73	82	77
2	83	4	8

TABLE V: SVM for child data: Accuracy=73.5%

Table 5, 6 and 7 show the classification report for the adolescent dataset using LSTM, RF and SVM respectively.

	Precision(%)	Recall (%)	f1-score (%)
0	87	85	87
1	97	97	97
2	90	89	90

TABLE VI: LSTM for Adolescent data:Accuracy=95.1%

	Precision(%)	Recall (%)	f1-score (%)
0	93	84	89
1	97	99	98
2	95	91	93

TABLE VII: RF for Adolescent data: Accuracy=96.4%

	Precision(%)	Recall (%)	f1-score (%)
0	0	0	0
1	79	100	88
2	96	10	17

TABLE VIII: SVM for Adolescent data: Accuracy=79.1%

	Precision(%)	Recall (%)	f1-score (%)
0	90	84	87
1	95	97	96
2	91	84	87

TABLE IX: LSTM forAdult data: Accuracy=93.9%

	Precision(%)	Recall (%)	f1-score (%)
0	96	90	93
1	97	98	98
2	93	89	91

TABLE X: RF for Adult data: Accuracy=96.3%

	Precision(%)	Recall (%)	f1-score (%)
0	50	6	11
1	77	99	87
2	73	3	6

TABLE XI: SVM for Adult data: Accuracy=77.0%

Our investigation delved into interpreting results from ensemble machine learning models tailored for predicting diabetes complications—specifically, hypoglycemia and hyperglycemia. Despite the simplicity and reliance on limited data, our experiments underscore the significant potential of data classification algorithms in predicting glycemic developments. The Random Forest algorithm emerged as the most accurate predictor, achieving impressive results of 96.3%, 96.4%, and 96.3% across child, adolescent, and adult datasets respectively. However, a notable limitation lies in its uncertain real-world performance due to susceptibility to overfitting.

A. Discussion

A thorough evaluation of the results obtained from the three models explains that the choice of metric to be used for evaluation solely depends on the nature of the problem and the data. Accuracy serves as a general metric to assess a model but an impulsive decision to decide the performance of a model based on its high accuracy result could be very wrong and might lead to death in some dire situations. This presumption of often leads to type 1 or type 2 errors. In ML, these means that a null hypothesis was rejected being a false positive or a null hypotheses was not rejected being a false negative. Considering the sensitivity of the problem in building an efficient classification model in predicting hypoglycaemia or hyperglycemia in 2 hours for a patient, the precision and recall of the model takes precedence over the accuracy .

References

[1] K. Ogurtsova, J. da Rocha Fernandes, Y. Huang, U. Linnenkamp, L. Guariguata, N. H. Cho, D. Cavan, J. Shaw, and L. Makaroff, "Idf diabetes atlas: Global estimates for the prevalence of diabetes for 2015 and 2040," *Diabetes research and clinical practice*, vol. 128, pp. 40–50, 2017.

[2] D. Control, C. T. of Diabetes Interventions, and C. D. S. R. Group, "Intensive diabetes treatment and cardiovascular disease in patients with type 1 diabetes," *New England Journal of Medicine*, vol. 353, no. 25, pp. 2643–2653, 2005.

[3] N. Cho, J. Shaw, S. Karuranga, Y. Huang, J. da Rocha Fernandes, A. Ohlrogge, and B. Malanda, "Idf diabetes atlas: Global estimates of diabetes prevalence for 2017 and projections for 2045," *Diabetes Research and Clinical Practice*, vol. 138, pp. 271–281, 2018.

[4] K. Liu, L. Li, Y. Ma, J. Jiang, Z. Liu, Z. Ye, S. Liu, C. Pu, C. Chen, Y. Wan, *et al.*, "Machine learning models for blood glucose level prediction in patients with diabetes mellitus: Systematic review and network meta-analysis," *JMIR Medical Informatics*, vol. 11, no. 1, p. e47833, 2023.

[5] G. E. Lienhard, J. W. Slot, D. E. James, and M. M. Mueckler, "How cells absorb glucose," *Scientific American*, vol. 266, no. 1, pp. 86–91, 1992.

[6] V. Agrawal, P. Singh, and S. Sneha, "Hyperglycemia prediction using machine learning: A probabilistic approach," in *Advances in Computing and Data Sciences: Third International Conference, ICACDS 2019, Ghaziabad, India, April 12–13, 2019, Revised Selected Papers, Part II 3*, pp. 304–312, Springer, 2019.

[7] M. A. Atkinson, G. S. Eisenbarth, and A. W. Michels, "Type 1 diabetes," *The Lancet*, vol. 383, no. 9911, pp. 69–82, 2014.

[8] J. Ozougwu, K. Obimba, C. Belonwu, and C. Unakalamba, "The pathogenesis and pathophysiology of type 1 and type 2 diabetes mellitus," *J Physiol Pathophysiol*, vol. 4, no. 4, pp. 46–57, 2013.

[9] J. M. Lawrence, J. Divers, S. Isom, S. Saydah, G. Imperatore, C. Pihoker, S. M. Marcovina, E. J. Mayer-Davis, R. F. Hamman, L. Dolan, *et al.*, "Trends in prevalence of type 1 and type 2 diabetes in children and adolescents in the us, 2001-2017," *Jama*, vol. 326, no. 8, pp. 717–727, 2021.

[10] E. Health, "The difference between type 1 and type 2 diabetes," 2017. Accessed: January, 2023.

[11] D. M. Lewis, "A systematic review of exocrine pancreatic insufficiency prevalence and treatment in type 1 and type 2 diabetes," *Diabetes Technology & Therapeutics*, vol. 25, no. 9, pp. 659–672, 2023.

[12] P. E. Cryer, J. N. Fisher, and H. Shamoon, "Hypoglycemia," *Diabetes care*, vol. 17, no. 7, pp. 734–755, 1994.

[13] J. Morales and D. Schneider, "Hypoglycemia," *The American journal of medicine*, vol. 127, no. 10, pp. S17–S24, 2014.

[14] J.-M. Guettier and P. Gorden, "Hypoglycemia," *Endocrinology and Metabolism Clinics*, vol. 35, no. 4, pp. 753–766, 2006.

[15] E. Society, "Hypoglycemia," 2023. Accessed: January, 2023.

[16] M. Mouri and M. Badireddy, "Hyperglycemia," in *StatPearls [Internet]*, StatPearls Publishing, 2023.

[17] D. J. Cox, B. P. Kovatchev, L. A. Gonder-Frederick, K. H. Summers, A. Mccall, K. J. Grimm, and W. L. Clarke, "Relationships between hyperglycemia and cognitive performance among adults with type 1 and type 2 diabetes," *Diabetes care*, vol. 28, no. 1, pp. 71–77, 2005.

[18] S. I. Taylor, Z. S. Yazdi, A. L. Beitelshees, *et al.*, "Pharmacological treatment of hyperglycemia in type 2 diabetes," *The Journal of Clinical Investigation*, vol. 131, no. 2, 2021.

[19] X. Tao, M. Jiang, Y. Liu, Q. Hu, B. Zhu, J. Hu, W. Guo, X. Wu, Y. Xiong, X. Shi, *et al.*, "Predicting three-month fasting blood glucose and glycated hemoglobin changes in patients with type 2 diabetes mellitus based on multiple machine learning algorithms," *Scientific Reports*, vol. 13, no. 1, p. 16437, 2023.

[20] E. Daskalaki, K. Nørgaard, T. Züger, A. Prountzou, P. Diem, and S. Mougiakakou, "An early warning system for hypoglycemic/hyperglycemic events based on fusion of adaptive prediction models," *Journal of diabetes science and technology*, vol. 7, no. 3, pp. 689–698, 2013.

[21] M. Omieljanowicz, T. Ignatowicz, and A. Omieljanowicz, "Predicting of hypo-and hyperglycemia in patients with type 1 diabetes with limited data," in *International Symposium on Applied Computing for Software and Smart Systems*, pp. 105–116, Springer, 2023.

[22] F. Hohman, K. Wongsuphasawat, M. B. Kery, and K. Patel, "Understanding and visualizing data iteration in machine learning," in *Proceedings of the 2020 CHI conference on human factors in computing systems*, pp. 1–13, 2020.

[23] E. Sullivan, "Understanding from machine learning models," *The British Journal for the Philosophy of Science*, 2022.

[24] M. Gadaleta, A. Facchinetti, E. Grisan, and M. Rossi, "Prediction of adverse glycemic events from continuous glucose monitoring signal," *IEEE journal of biomedical and health informatics*, vol. 23, no. 2, pp. 650–659, 2018.

[25] C. V. G. Zelaya, "Towards explaining the effects of data preprocessing on machine learning," in *2019 IEEE 35th international conference on data engineering (ICDE)*, pp. 2086–2090, IEEE, 2019.

[26] S. García, J. Luengo, and F. Herrera, *Data preprocessing in data mining*, vol. 72. Springer, 2015.

[27] G. Dong and H. Liu, *Feature engineering for machine learning and data analytics*. CRC press, 2018.

[28] P. Duboue, *The art of feature engineering: essentials for machine learning*. Cambridge University Press, 2020.

[29] L. Breiman, "Random forests," *Machine learning*, vol. 45, pp. 5–32, 2001.