Exploring Machine Learning Methods to Predict Hypoglycemic States in Diabetes Type I Patients

Estimating Time to Onset: Leveraging Glucose, Activity, and Insulin Data

Masterarbeit

Zur Erlangung des Grades
Master of Science (M. Sc.)
im Studiengang
Medical Data Science

eingereicht von
Beyza Cinar
Abgabedatum: 21.03.2024

Erstprüferin: Jennifer Daniel Onwuchekwa

Zweitprüferin: Prof. Dr. Maria Maleshkova

Universität Siegen
Lebenswissenschaftliche Fakultät
Lehrstuhl Medizinische Informatik mit Schwerpunkt mobile Gesundheitsinformationssysteme

Wintersemester 2023/24

Plagiatserklärung

Ich versichere, dass ich die schriftliche Ausarbeitung selbständig angefertigt und keine anderen als die angegebenen Hilfsmittel benutzt habe. Alle Stellen, die dem Wortlaut oder dem Sinn nach (inkl. Übersetzungen) anderen Werken entnommen sind, habe ich in jedem einzelnen Fall unter genauer Angabe der Quelle (einschließlich des World Wide Web sowie anderer elektronischer Datensammlungen) deutlich als Entlehnung kenntlich gemacht. Dies gilt auch für angefügte Zeichnungen, bildliche Darstellungen, Skizzen und dergleichen. ***Insbesondere versichere ich, dass ich alle wörtlichen und sinngemäßen Übernahmen aus anderen Werken sowie die Verwendung KI-basierter Textgeneratoren als solche kenntlich gemacht habe.*** *Ich nehme zur Kenntnis, dass die nachgewiesene Unterlassung der Herkunftsangabe als versuchte Täuschung gewertet wird.*

I certify that I have written this paper independently and that I have not used any other resources than those indicated. All passages taken from other works in terms of wording or meaning (including translations) have been clearly marked as borrowed in each individual case, with a precise indication of the source (including the World Wide Web and other electronic data collections). This also applies to attached drawings, pictorial representations, sketches, and the like. ***In particular, I assure that I have marked all verbatim and analogous copies from other works as well as the use of AI-based text generators as such.*** *I acknowledge that the proven omission of the indication of origin will be considered as attempted deception. (Translated with www.DeepL.com/Translator)*

Ort, Datum	Name, Unterschrift

Abstract

Hypoglycemia is a serious condition associated with increased mortality in patients with type 1 diabetes, which is an incurable autoimmune disease. Hypoglycemia is defined by blood glucose levels below 70 mg/dL. The causes can include excessive insulin injections, skipping meals, or increased physical activity. It can occur suddenly and may be asymptomatic, impeding timely preventive measures. Thus, innovative technologies, such as machine learning, can help to predict the state before it occurs. Prediction models are mainly classified as short- and long-term prediction horizons (PHs) of up to 2 hours and up to 24 hours, respectively. Most research conducted in the field of diabetes forecasts blood glucose values. Still, the obtained accuracy may not be sufficient to prevent hypoglycemia due to the possible time lag of CGM devices. Moreover, most studies focus on one PH only. This thesis included short- and long-term PHs in the same classification model to consider multiple use cases and to enable better decision support. The predicted times are 5-15 min, 15-30 min, 30 min-1 h, 1-2 h, 2-4 h, 4-8 h, 8-12 h, 12-24 h before hypoglycemia. The input features are prior glucose measurements, the administered basal and bolus insulin dosages, and acceleration data. First, a correlation analysis between the input features and the classes is conducted. Thereafter, RNN and CNN are explored to classify the onset of hypoglycemia based on the proposed nine classes. Furthermore, training with six classes classifying up to 4 hours before the onset is compared. Finally, subject-specific models are tested. The population-based correlation analysis reveals a very weak association between basal insulin and glucose, and between basal insulin and acceleration data. An individual correlation analysis showed stronger relationships, but the scores varied significantly among the subjects. For the classification model with nine classes, the best results are obtained with a LSTM model. Subject-specific models improve the performance. However, only classes 0-2 could be well classified with recalls of 98%, 72%, and 50%, respectively. A population-based model with only six classes obtains better results with recalls of 99%, 73%, and 56% for classes 0, 1, and 2, respectively. In conclusion, the proposed system that includes short- and long-term PHs is not feasible with the data or models used. Whereas, a model classifying multiple short-term horizons up to 4 hours before hypoglycemia produces promising results with improved precision, and F1-measure and indicates that at least 60% of events can be predicted which is increased to approximately 70% in subject 563.

Contents

1. Introduction 1

2. Foundamentals 4

2.1. Diabetes Mellitus . . . 4

2.1.1. Type 1 Diabetes . . . 5

2.1.2. Complications of Type 1 Diabetes . . . 7

2.1.3. Therapy and Technology . . . 8

2.2. Machine Learning . . . 15

2.2.1. Neural Networks . . . 15

2.2.2. 1DCNN . . . 16

2.2.3. ResNet . . . 18

2.2.4. LSTM . . . 19

3. State of the Art 22

3.1. Short-Term Prediction Horizons . . . 22

3.2. Long-Term Prediction Horizons . . . 26

3.3. Variables and Features Impacting the Performance . . . 32

3.4. Identified Research Gaps . . . 33

4. Methodology 34

4.1. Dataset: OhioT1DM . . . 35

4.1.1. Pre-processing . . . 36

4.1.2. Correlation Analysis . . . 40

4.2. Model Architectures . . . 42

4.3. Metrics used for the Model Evaluation . . . 49

4.4. Expected Limitations . . . 51

5. Results 52

5.1. Preliminary Data Analysis . . . 52

5.2. Results of the Correlation Analysis . . . 60

5.2.1. Population-based Correlation . . . 60

5.2.2. Individual-based Correlation . . . 68

5.3. Deep Learning models . . . 72

5.3.1. Population-based Models using 9 Classes . . . 72

5.3.2. Subject-specific and Population-based Models using 9 Classes . . . 79

5.3.3. Population-based Models using 6 Classes 89
5.3.4. Subject-specific and Population-based Models using 6 Classes . . 95

6. Discussion **103**

7. Conclusion **109**

8. Future Work **111**

A. Appendix **124**

List of Figures

1. Pathology of type 1 diabetes . . . 6
2. Hormone-pumps . . . 12
3. Solution to reduce exercise-induced hypoglycemia . . . 14
4. Basic architecture of a Feedforward Network . . . 16
5. 1DCNN model for sequential data . . . 17
6. ResNet architecture . . . 19
7. Architecture of an LSTM model and its cells . . . 20
8. Pipeline of this thesis . . . 34
9. Architecture of applied RNN models . . . 44
10. Confusion-matrices of RNN models . . . 45
11. Architecture of applied CNN models . . . 46
12. Confusion-matrices of CNN models . . . 48
13. Architecture of applied hybrid model . . . 49
14. Architecture of LOOCV . . . 49
15. Parameters in the last 48 hours before hypoglycemia (1) . . . 56
16. Parameters in the last 48 hours before hypoglycemia (2) . . . 57
17. Parameters in the last 48 hours before hypoglycemia (3) . . . 58
18. Parameters in the last 48 hours before hypoglycemia (4) . . . 59
19. Pairwise-plots (1) . . . 63
20. Pairwise-plots (2) . . . 64
21. Pairwise-plots (3) . . . 65
22. Pairwise-plots (4) . . . 66
23. Pairwise-plots (5) . . . 67
24. Population-based confusion-matrices across 9 classes . . . 79
25. Subject-specific confusion-matrices across 9 classes . . . 87
26. Population-based and subject-specific confusion-matrices across 9 classes 88
27. Population-based confusion-matrices across 6 classes . . . 94
28. Subject-specific confusion-matrices across 6 classes . . . 101
29. Population-based and subject-specific confusion-matrices across 6 classes 102

List of Tables

1. State of the Art: Short-term PHs . . . 23
2. State of the Art: Long-term PHs . . . 27
3. Number of samples in the OhioT1DM dataset: 2018 . . . 36
4. Number of samples in the OhioT1DM dataset: 2020 . . . 36
5. Assignment of the classes . . . 39
6. Rule of thump for correlation interpretation . . . 40
7. Comparison of LSTM and BiLSTM models . . . 44
8. Comparison of ResNet and 1DCNN models . . . 47
9. Samples per class with raw data . . . 53
10. Samples per class for pre-processed time series data of 8 hours . . . 53
11. Population-based Pearson and Spearman correlation analysis . . . 61
12. Macro average metrics of each model for each subject using 9 classes . . . 76
13. Population-based LSTM results for each subject using 9 classes . . . 77
14. Performance of each class for each model using 9 classes . . . 78
15. Subject-specific macro average metrics of each model for each subject using 9 classes . . . 83
16. Subject-specific LSTM results for each subject using 9 classes . . . 84
17. Population-based LSTM results with less test data for each subject using 9 classes . . . 85
18. Comparison of population-based and subject-specific approaches for each model using 9 classes . . . 86
19. Performance of each class for each model using 6 classes . . . 91
20. Macro average metrics of each model and each subject using 6 classes . . . 92
21. LSTM results for each subject using 6 classes . . . 93
22. Subject-specific macro average metrics of each model and each subject using 6 classes . . . 97
23. Subject-specific LSTM results for each subject using 6 classes . . . 98
24. Population-based LSTM results with less test data for each subject using 6 classes . . . 99
25. Comparison of the population-based and subject-specific approach for each model using 6 classes . . . 100
26. Pearson correlation analysis for each class for each subject (1) . . . 125
27. Pearson correlation analysis for each class for each subject (2) . . . 126

28. Pearson correlation analysis for each class for each subject (3) 127
29. Pearson correlation analysis for each class for each subject (4) 128
30. Population-based ResNet results for each subject using 9 classes 129
31. Population-based hybrid model results for each subject using 9 classes . . 130
32. Subject-specific ResNet results for each subject using 9 classes 131
33. Population-based ResNet results with less test data for each subject using 9 classes 132
34. Subject-specific hybrid model results for each subject using 9 classes . . . 133
35. Population-based hybrid model results with less test data for each subject using 9 classes 134
36. Population-based ResNet results for each subject using 6 classes 135
37. Population-based hybrid model results for each subject using 6 classes . . 136
38. Subject-specific ResNet results for each subject using 6 classes 137
39. Population-based ResNet results with less test data for each subject using 6 classes 138
40. Subject-specific hybrid model results for each subject using 6 classes . . . 139
41. Population-based hybrid model results with less test data for each subject using 6 classes 140

Acronyms

1DCNN One Dimensional Convolutional Neural Networks. 3, 4, 16, 17, 19, 21, 33, 42, 44, 46, 47

AI Artificial Intelligence. 2, 11, 12, 15, 36, 104, 109

BGL Blood Glucose Level. 4, 5, 7–9, 12, 22

BiLSTM Bidirectional Long Short-Term Memory. 20, 42–44

CGM Continuous Glucose Monitoring. , 1, 2, 9, 11, 35, 105

CNN Convolutional Neural Networks. , 3, 16–19, 24–26, 30, 33, 42, 43, 46, 109

LOOCV Leave One Out Cross-Validation. 30, 47

LSTM Long Short-Term Memory. , 3, 4, 16, 20, 21, 24–26, 29, 33, 42–49, 72–75, 79–82, 89, 91, 95, 96, 106–108, 110

macc Magnitude of Acceleration. 60, 62, 68–71, 104

MLP Multilayer Perceptron. 15

NN Neural Networks. 15, 24

PH Prediction Horizon. , 2, 3, 22–34, 104, 106, 110, 111

ResNet Residual Network. 3, 42, 44, 46–49, 72–75, 79–82, 89–91, 95, 96, 107, 108, 110

RNN Recurrent Neural Networks. , 3, 19, 21, 22, 24, 25, 42, 43, 46, 109

SVM Support Vector Machines. 15, 28–31, 33, 111

T1D Type 1 Diabetes. 1–9, 12, 13, 35, 39

1. Introduction

Diabetes mellitus is one of the fastest increasing and most prevalent chronic diseases and is predicted to affect approximately 1.3 billion people worldwide by 2050 [1]. It is a long-term disease in which glucose cannot be metabolized normally, leading to continuously elevated blood glucose levels [2]. Diabetes mellitus is mainly classified into type 1, type 2, and type 3 diabetes. Type 1 Diabetes (T1D) is an incurable autoimmune disease in which insulin production is destroyed, type 2 diabetes can result from an unhealthy lifestyle and diet, and type 3 diabetes which is gestational diabetes occurs during pregnancy [3]. In particular, T1D may be more difficult to manage because its onset can occur in childhood and youth. It has been reported that T1D accounts for 2% of all cases of diabetes [4] and has a global prevalence of 9.5% [5]. Patients with T1D cannot produce sufficient insulin. Therefore, blood glucose levels are not harmonically regulated and external insulin injections are required. The main goal of diabetes therapy is to maintain normal glucose levels and to prevent increased glucose concentrations above 180 mg/dL defined as hyperglycemia, and decreased glucose concentrations below 70 mg/dL, defined as hypoglycemia. As a consequence, hyperglycemia and hypoglycemia are associated with vascular complications and comorbidities. Notably, hypoglycemia is more often seen in patients with T1D. It can be life-threatening and is mainly caused by inadequate insulin dosages. In addition, it could be impacted by low meal intakes [4]. Other direct or long-term causes of hypoglycemia can include extensive activity necessitating increased glucose levels in muscle cells, which may impact insulin sensitivity hours later [6, 4]. These associations can lead to fear and anxiety. As a result, patients with T1D are less active and feel discouraged from participating in sports [6, 7].
Hypoglycemia is a serious condition, but it can be prevented by glucose intake if the condition is detected before it occurs. If the event is asymptomatic and unnoticed, preventive self-actions may be impeded. Nocturnal hypoglycemia, for instance, appears to be mostly asymptomatic and is considered to be one of the main causes of sudden death of T1D. Moreover, children are at higher risk for hypoglycemia, nocturnal hypoglycemia, and exercise-induced hypoglycemia, because of their active and unpredictable lifestyle [8, 7]. Therefore, technological approaches could help to prevent the state by predicting adverse events, thus enabling timely treatment. Currently, diabetes care is improved by Continuous Glucose Monitoring (CGM) devices, which are sensors that measure glucose concentrations in interstitial fluid under the skin. CGM devices can be augmented with machine learning techniques to forecast future glucose values [9]. They also visualize

glucose values in real time and can alert the patients of abnormal patterns. Machine learning is a function approximation problem and learns patterns from the given data to draw inferences. As a result, if adverse events can be predicted before their occurrence, the risk can be reduced by adjusting the intake of food, insulin dosage, or activity type. Artificial Intelligence (AI) and data analysis can support individualized decisions based on the patient's current condition [6]. Moreover, insulin pumps worn on a belt were designed to subcutaneously infuse insulin substitutions. If CGM devices are connected to insulin pumps, insulin dosages can be adjusted and controlled automatically. Such a system is called an artificial pancreas or closed loop system [9, 10]. In this context, previous studies developing systems based on machine learning methods for glucose prediction have concluded that physical activity, data of the given insulin dosage, and information about the meal intake can improve the application's performance for short-term Prediction Horizons (PHs) and horizons above 45 minutes [11, 12]. However, available hybrid insulin pumps are not fully automatic and require manual user interventions of food intake and physical activity. Wearable devices may therefore be used for data estimation. Currently, various Food and Drug Administration-approved devices measuring heart rate, galvanic skin response, and acceleration data are available. Wearables are portable small computers or devices that can be embedded with sensors and wireless technology so that physiological parameters can be continuously estimated and visualized in smart devices. In 2006-2007, the first wearable fitness trackers were introduced [13]. In those, an accelerometer measures the static or dynamic acceleration forces and the rate of change in velocity along multiple axes. The experienced physical movement of a mechanical element within the accelerometer structure is then converted into an electrical signal [14]. A literature review has shown that most diabetes research based on short-term PHs which range from 30-120 minutes focuses on forecasting glucose values using regression. Contrariwise, studies based on long-term prediction horizons which range from 2-24 hours more often classify adverse events with binary classification. However, to the authors' knowledge, no study has integrated short-term and long-term PHs into the same model, which is possible in classification systems, unlike regression models. A model, which identifies the risk from 24 hours to 15 minutes before, can enable short-term preventive actions, and the adaption of daytime activities and insulin dosages. With this in mind, as a difference from other present studies, this thesis includes multiple time horizons into one model to classify the time to the onset of hypoglycemia in patients with T1D. Utilized data includes glucose concentrations, applied basal and bolus insulin dosages, and physical activity. The capability of machine learning methods is investigated for the

proposed task. In this context, deep learning methods were chosen because no feature engineering is required and the performance does not depend on extracted features. This study compares Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) architectures including One Dimensional Convolutional Neural Networks (1DCNN) and Long Short-Term Memory (LSTM) models. In particular, answers to the following research question are investigated:

"**In patients with T1D, how effective are machine learning models in predicting the time until the occurrence of hypoglycemia when utilizing glucose levels, physical activity data, and insulin values as input features?**"

Notably, this thesis will make the following contributions:

1. It investigates the correlation between estimated glucose, basal insulin, bolus insulin, and the magnitude of acceleration for different time intervals before hypoglycemia.
2. It investigates the performance of deep learning models while comparing Residual Network (ResNet), LSTM, and a hybrid model of ResNet and LSTM.
3. It compares between population-based and subject-specific models.
4. Finally, it compares between short-term and long-term prediction horizons.

The objective is to decrease the risk of hypoglycemic events, including insulin-induced and exercise-induced hypoglycemia, by alerting patients with T1D beforehand. Conclusively, this thesis aims to promote short-term self-management as well as the long-term prediction of hypoglycemic events which could improve decision support. The included PHs are 5-15 minutes, 15-30 minutes, 30 minutes-1 hour, 1-2 hours, 2-4 hours, 4-8 hours, 8-12 hours, 12-24 hours, and 24–48 hours before the event. Those interventions could provide better management, better planning of daytime activities and meals, better life quality, and better physiological and psychological health of T1D patients.
This study is further structured as follows. Chapter 2 gives the necessary background information about T1D and its therapy, and about machine learning, in particular about the applied deep learning methods such as 1DCNN and LSTM. Then, chapter 3 explores the state of the art in diabetes research, hypoglycemia estimation, and research gaps. The methodology of this work is presented in chapter 4 by describing the utilized data and its pre-processing steps, the methods for the correlation analysis, and the model architectures. The results are presented in chapter 5 while chapter 6 discusses the findings, and chapter 7 concludes the main contributions of this work. Finally, chapter 8 highlights possible future work.

2. Foundamentals

This chapter gives the necessary background information to understand the motivation and aim of this thesis. Furthermore, the theory of applied technological methods is demonstrated. To begin with, the metabolic disease "Diabetes Mellitus", especially type 1 diabetes with its therapy methods and complications are described. Here, the focus is set on hypoglycemia, the various insulin replacements, and the impact of physical activity. Afterward, machine learning, neural networks, and deep learning are explained while highlighting models like 1DCNN and LSTM since those are utilized in the experiments.

2.1. Diabetes Mellitus

Diabetes mellitus is a chronic disease in which insulin cannot be secreted or used efficiently by the body. Furthermore, the glucose metabolism can be destroyed as to why more glucose fluctuates in the blood. Those dysfunctions and deficiencies lead to elevated and varying Blood Glucose Levels (BGLs) [3]. In particular, major complications of diabetes are hyperglycemia and hypoglycemia. Hyperglycemia is defined by increased BGLs above 180 mg/dL, and chronic hyperglycemia highly impacts life quality because it can lead to organ damage and further dysfunctions. In addition, hyperglycemia can cause microvascular and macrovascular diseases such as retinopathy, nephropathy, neuropathy, or cardiovascular diseases [3, 15]. Contrariwise, hypoglycemia is defined by decreased BGLs below 70 mg/dL and can be life-threatening since it can result in coma or in death in the worst case. Therefore, both conditions are dangerous and should be avoided.
537 million people were living with diabetes in 2021 according to the International Diabetes Federation, and it was predicted that 643 million people could be diagnosed with diabetes by 2030 and 783 million people by 2045 [16]. In 2023, it was reported that the expected cases could rise from 529 million to 1.3 billion by 2050 [1]. Moreover, diabetes-related complications affected 6.7 million deaths worldwide in 2021. For only Europe, it is revealed that 61 million persons had diabetes and 1.1 million people died from diabetes in 2021. In addition, an expected rise to 67 million persons by 2030 and 69 million persons by 2045 is predicted [16]. Following these reports, it can be seen that diabetes mellitus is a global health problem and needs more preventive actions to decrease the rising incidence rate.
Diabetes mellitus is classified into different types based on its etiology. It is mainly differentiated between type 1, type 2, and type 3 diabetes, all of which may require different treatments and management. T1D is an autoimmune disease that mostly has a

genetic etiology as will be explained in subsection 2.1.1.
In Type 2 diabetes, sufficient insulin can be produced by the pancreas but it cannot be used effectively due to insulin resistance. Hence, the BGLs are elevated [3, 15]. In 2010, it was reported that 90-95% of all diabetic patients are diagnosed with type 2 diabetes [3, 15], increasing to 96% in 2023 according to the Institute for Health Metrics and Evaluation [1]. The etiology is based on unhealthy diet and lifestyle including obesity, stress, and physical inactivity. Type 2 diabetes is curable in its early stages since the glucose concentrations can be controlled with an adequate lifestyle [3]. The main complications of type 2 diabetes are hyperglycemia and resulting comorbidities.
Type 3 diabetes is gestational diabetes and develops before or during pregnancy because of the pregnancy hormones which could destroy the produced insulin. Type 3 diabetes is normally said to disappear after delivery but the affected are at increased risk of developing type 2 diabetes [15].

2.1.1. Type 1 Diabetes

Type 1 diabetes is one of the most prevalent chronic diseases in children and often has its onset in youth. In 2010, it was reported that the incidence and prevalence are increasing and around 5-10% were estimated to have T1D [3]. In 2020, Mobasseri et al. presented that the incidence of T1D was 15 per 100.000 people and the worldwide prevalence was 9.5% [5]. In 2021, the Robert Koch Institute reported that 1.5 million people under 20 years were diseased with T1D. In addition, for Germany, a global incidence rate of 3-4% is estimated for the last decades, while the estimated prevalence in children and adolescents was 235.5 per 100.000 persons in 2020 of whom more boys were affected [17].
As indicated earlier, T1D is an autoimmune disorder in which the immune system attacks the pancreatic cells. Therefore, insulin-producing β-cells in the islets of Langerhans are damaged by white blood cells which are called B-cells and T-cells [3, 15, 18, 19]. In this process, B-cells present produced antigens to T-cells, so that falsely identified invaders are eliminated [15]. Markers could be islet cell auto-antibodies, auto-antibodies to insulin, auto-antibodies to GAD identified as GAD65, and auto-antibodies to the tyrosine phosphatases identified as IA-2 and IA-2 [3]. If the β-cells are destroyed, insulin cannot be produced sufficiently, as why external insulin injections are required as a replacement therapy [18, 19]. Additionally, Bolli et al. present that the β-cell destruction could result in a deficiency of amylin secretion. Given this, T1D could be a dual hormone deficiency disease. These deficiencies would usually cause high glucose variations in patients with T1D, and a more difficult glucose control especially after the meal [20].

Amylin is also a hormone secreted by the β-cells. Amylin secretion is activated by insulin secretion when a person starts to eat, and the hormone could impact glucose homeostasis [20]. Besides, patients with T1D can also develop a dysfunction of α-cells which secrete glucagon to prevent hypoglycemia [10]. Figure 1 summarizes the pathology of T1D.

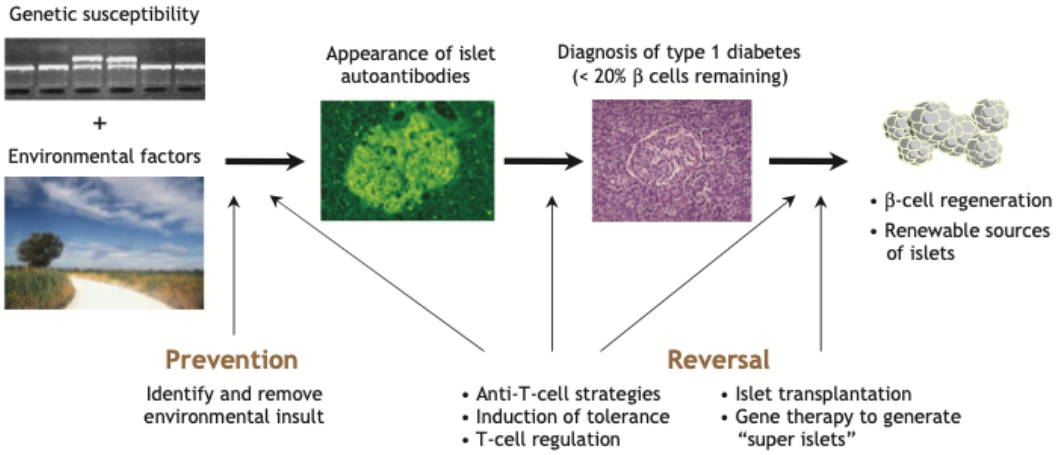


Figure 1: Pathology of type 1 diabetes [21]

T1D is categorized as a polygenic disorder [22], and the etiology could be explained by genetic predispositions, particularly in children and youth-onset type 1 diabetes. Besides, external factors such as viruses, the environment, and diet or stress could trigger β-cell destruction [3, 15, 19]. Twin studies have also revealed that non-genetic factors could contribute to the activation of T1D [22]. The destruction or dysfunction of pancreatic tissues may also be caused by other diseases, such as chronic pancreatitis, trauma, or surgical removal of the pancreas. Moreover, endocrine diseases, such as excessive growth hormone production and Cushing's syndrome, could impact the onset due to significantly increased cortisol production, as stated by Nilam et al. [15]. In addition, Lucier et al. asserted that the metabolic, genetic, and immunogenetic characteristics of T1D could differ among subjects. Age could also influence the disease development and treatment as to why personalized treatment methods and therapies are suggested [19].
Finally, T1D is classified into three stages. The first stage is the early asymptomatic stage, which is defined by normal fasting glucose and normal glucose tolerance [19, 22]. The development of autoimmunity can usually be detected by the presence of at least two circulating pancreatic islet auto-antibodies, despite the absence of symptoms at the early stage [19, 22]. Stage 2 which is called asymptomatic dysglycemia, can be diagnosed if a large number of β-cells have already been destroyed and if multiple auto-antibodies are detected. It can be further differentiated by impaired fasting glucose or impaired

glucose tolerance [19]. Lastly, stage 3 is identified with the onset of hyperglycemia caused by insufficient insulin secretion with clinical symptoms [19, 22]. Those patients could randomly have glucose concentrations above 199 mg/dL, and a fasting glucose concentration above 125 mg/dL [19].

2.1.2. Complications of Type 1 Diabetes

Patients with T1D are at a higher risk for developing critical comorbidities as listed in 2.1 and other autoimmune disorders, such as autoimmune thyroid disease, celiac disease [19], Graves' disease, Hashimoto's thyroiditis, autoimmune hepatitis, or pernicious anemia [3]. Moreover, they could have a relatively higher risk of developing cardiovascular diseases. Especially, women with T1D could be at greater risk [23] Lucier et al. point out that the mortality risk could be 2-5 times higher compared to non diseased [19]. The main issue is that the blood glucose concentration cannot be harmonically regulated without external help resulting in continuously elevated BGLs. Subsequently, if T1D is not treated with insulin replacements, and the patient is often in a hyperglycemic state, severe conditions like diabetic ketoacidosis can occur which is estimated to appear more often in youth [19, 15]. Ketoacidosis is also called a diabetic coma and is a consequence of insulin deficiency, falsely managed insulin therapy, and acute, severe conditions [24].
Hypoglycemia is another serious life-threatening condition of patients with T1D resulting in coma, acute brain damage, arrhythmia, and death in the worst case [20, 25]. In this context, 6-10% of T1D related death cases were estimated to be caused by hypoglycemia, according to reports of Cederblad et al. [25]. Hypoglycemic conditions are mainly a consequence of insulin therapy, particularly with inadequate and extensive insulin dosages, as well as an insulin injection at the wrong time [24, 19]. Furthermore, skipped meals or insufficient meal intake can lead to decreased blood glucose concentrations. Lastly, physical activity in which more glucose is consumed by muscle cells can provoke hypoglycemic events [24]. Moreover, it is mentioned that insulin sensitivity can vary based on the daytime. Here, Haak et al point out that it is increased during the night, after physical fitness, or after improved glucose control [24]. Overall, the main challenge for patients with T1D is to maintain normal glucose values and to prevent serious fluctuations since the treatment of elevated glucose concentration can result in an unwanted decrease.
Hypoglycemia can be classified into 3 levels. Stage 1 is non-severe and characterized by glucose concentrations between 54 mg/dL to less than 70 mg/dL. It is helpful to predict the occurrence of stage 1 since it alerts to take preventive actions and to prevent further decrease. Stage 2 is clinically important and characterized by significantly decreased

glucose values less than 54 mg/dL. Notably, it requires immediate action. Stage 3 is severe hypoglycemia and can cause (long-term) cognitive impairment. Affected then depend on external help [26] as to why the condition is particularly dangerous for children. Severe hypoglycemia often results in asymptomatic hypoglycemia or nocturnal hypoglycemia since the patient is unaware of their glucose drop. In particular, the elderly would be at higher risk for severe hypoglycemia. Besides, nocturnal hypoglycemia is a life-threatening condition since it can result in sudden death. Additionally, impaired glucose awareness is associated with frequent hypoglycemia. [26]. As mentioned already, asymptomatic hypoglycemia is a challenging condition and hard to predict. Mujahid et al. remark that consuming 15-20 grams of fast-acting carbohydrates could prevent further decrease of BGLs if the glucose levels decrease below 70 mg/dL. Nevertheless, the state needs to be predicted before because the glucose requires time to get into the blood which could take 10-15 minutes [27]. Moreover, a snack before sleeping is reported to reduce the risk of nocturnal hypoglycemia, if a possible occurrence is detected before [28]. Technological approaches are still being investigated and improved to alert patients of an incoming event. Advanced methods such as machine learning can be utilized. Correspondingly, the model needs to predict an event at least 15 minutes before so that the consumed glucose can be regulated. In particular, technology could support parents with the management of their children's diabetes. The management of glucose concentrations of children with T1D is more complicated since they can have a more active lifestyle which does not follow a routine provoking adverse events.

If not asymptomatic, symptoms can be "diaphoresis, tachycardia, lightheartedness, confusion, hunger, visual changes, and tremors" according to Lucier et al. [19]. Unawareness is highlighted to appear commonly if the disease is ongoing for a longer time [19]. Other consequences of hypoglycemia can be psychological conditions such as fear and anxiety. Those conditions can lead to reduced insulin dosages or less physical activity further leading to an increased risk for comorbidities [26, 7]. Additionally, hypoglycemia could reduce life quality, cause mood swings, increase stress, and could decrease concentration [26]. Altogether, a severe hypoglycemic episode could be associated with an increased risk of death within 5-10 years [7, 25].

2.1.3. Therapy and Technology

To prevent the complications described in the previous subsection 2.1.2, it is relevant to monitor the blood glucose values frequently, ideally continuously. Thus, consequences of inadequate actions can be detected and preventive actions can be taken timely. The

monitoring and control of other parameters such as blood pressure and cholesterol levels could be relevant as well [15].

CGM

Conventional methods for the assessment of blood glucose concentrations were based on intravenous blood analysis. Then, improved technology enabled more frequent daily self-monitoring with blood analysis by finger-pricks. Nowadays, medically certified CGM devices are available which are based on electrochemical sensors. Currently, they are the most accurate devices for measuring blood glucose levels continuously. With these advances, diabetes management rapidly improved, as well as the life quality of the patients [22]. Furthermore, new technologies were enabled such as insulin pumps, and artificial pancreas systems.

CGM devices are based on a biotechnological approach that uses enzyme reactions, a wireless sensor, a transmitter, and a receiver to measure BGLs in the interstitial fluid. The wireless sensor is inserted into the subcutaneous tissue under the skin and the glucose is measured usually at intervals of 5-15 minutes [19, 29]. The readings are then exchanged with the receiver presenting the outcome [19]. These can be augmented with further trend estimation or prediction models to alert the patient of adverse events. Moreover, it can be coupled with an insulin pump and help to adjust insulin dosages [19]. CGM devices are intended for long-term monitoring and usually last for multiple days. Nevertheless, they have a limited lifetime because they are based on electrochemical reactions. Furthermore, they need to be calibrated at least twice a day with finger prick-based blood analysis [30, 10]. One disadvantage is a reported time lag of 6-12 minutes since the glucose is not directly assessed from the blood but from interstitial flood [30, 10]. Here, the glucose is transported from the vascular to the interstitial space [10], as why the prediction accuracy of machine learning models needs to be as accurate as possible.

To summarize, CGM devices can assist a healthier lifestyle and educate patients since the consequences of actions are better understood [19, 29].

Insulin

Patients with T1D depend on external insulin treatment to decrease their glucose levels. Therapy with insulin started in 1921-1922 and new advances, medications, and technologies were invented over those last 100 years [31, 32, 20]. Nowadays, the physiology and biochemistry of insulin and pancreatic cells are well understood [20] enabling precise replacement therapies. Insulin is a hormone produced by pancreatic cells. The pancreas

regulates its secretion and delivers it into the portal vein in cyclic pulses. The produced insulin binds to its receptors and affects carbohydrate, lipid, and protein metabolism, while each metabolic process could have a different insulin sensitivity [20]. Besides, different insulin sensitivity is observed in glucose production and use. It is known that endogenous glucose production can be stopped with even small increases in insulin, so even a little overdosage can provoke hypoglycemia [20]. Normally, once a person starts a meal or once the glucose concentration in the blood increases, the insulin secretion activates [15]. Conversely, insulin secretion decreases if the plasma glucose concentration decreases [20]. Moreover, insulin enables the glucose to get into the cells and reach organs and tissues. Hence, if insulin is not produced or cannot be used, the glucose cannot leave the blood and is elevated [15].

Insulin replacements are classified into two groups: long-acting basal insulin and rapid-acting bolus insulin. Basal insulin helps to maintain a stable glucose level in the fasting state and during the night. In automatic insulin pump therapies, basal insulin is infused in small dosages during the day, while the insulin is often given before sleep with external self-injections [4]. In contrast, bolus insulin regulates the glucose rise after the meal [4]. Regularly used short-acting insulin, has an onset in 30 minutes to 1 hour and peaks in 2-4 hours with a duration of 5-8 hours. A rapid-acting insulin is reported to have its onset in 12-30 minutes, a peak in 1-3 hours, and a duration of action of 3-6 hours [19]. Moreover, ultra-rapid-acting insulin has a quicker onset and shorter duration of action. Basal insulin is often only given once or twice a day and can normally last for 20-24 hours [19, 31]. Also, some basal replacements are mentioned to last for more than 24-42 hours [19]. Intermediate insulin which more often leads to hypoglycemia as reported in [19], has its onset in 1-2 hours, peak of action at 2-8 hours, duration of 12-24 hours, and is usually given before breakfast or sleeping.

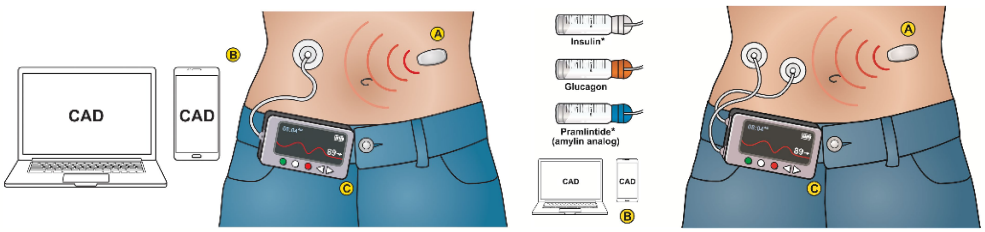
For the calculation of the appropriate initial daily insulin dosage, the person's weight in kilograms is multiplied by 0.2 to 0.6 units. The basal needs would be generally 0.4 to 0.5 of the daily needed dosage, while the rest consists of rapid-acting insulin injected before or after the meal. The initial formula is personalized and changed based on external factors. Moreover, the insulin needs vary over the lifespan of the patient. In this context, in the early stage, when first diagnosed, less insulin may be more appropriate while in puberty increased dosages may be required [19].

Multiple daily insulin injections using basal and bolus replacements, continuous subcutaneous insulin infusion through an insulin pump, or the use of automated insulin delivery systems are available nowadays [19]. The main goal is to keep the glucose concentration in

a normal range without high fluctuations [32]. Normally produced insulin autonomously adjusts and activates the insulin secretion to maintain a target range of 72-180 mg/dL. They have a short duration and action of approximately half time under 5 minutes as reported by Home and Mehta [31]. However, insulin replacements still cannot reach the activation time and duration of real insulin. Therefore, new drugs and new technological methods are investigated.

Insulin pumps generally consist of an insulin tank, a pump, and a controller. The dosage is externally controlled and predefined [4]. Those can be based on open-loop systems requiring user input such as meal intake, or they can be based on closed-loop systems, coupled with a CGM device, which is then called an artificial pancreas. Currently, hybrid models are more in use requiring user input and control [4]. Lucier et al. point out that more often rapid-acting insulin is used in insulin pumps which are usually delivered every 5 minutes. The basal rates can be programmed and corrected, and insulin delivery can be stopped if a hypoglycemic event is sensed. In addition, advanced systems can automatically correct bolus dosages [19]. The artificial pancreas is intended to imitate the function of a pancreas for glucose control. It is designed as a fully closed-loop system. It is highlighted that the ideal insulin imitation would be one replacing prandial and basal insulin needs [32]. Those systems are also augmented with further algorithms and AI so that future predictions and patient behavior can be included in the insulin dosage calculation. Automated insulin pumps are said to reduce the occurrence of especially nocturnal hypoglycemia. They would also lead to better management of target values. Furthermore, they would lessen the anxiety of patients and reduce the duration of events [10, 19]. However, limitations include the subcutaneous time lag in the diffusion of glucose and insulin from blood, the time lag of the estimated glucose by the CGM device, and the delay and variation in insulin absorption and action, particularly of subcutaneous rapid-acting insulin analogs [10, 32, 4]. Therefore, it is asserted that stopping insulin with sensed decreasing glucose values would possibly not prevent hypoglycemia [4]. Irregular behavior such as meals, or illness can also cause unpredicted glucose fluctuations [10]. Another challenge is the timely prediction of exercise-induced hypoglycemia to adjust the administered insulin dosage.

Those limitations led to the research of dual hormone insulin pumps [19] which usually deliver glucagon and insulin to reduce hypoglycemic events. Figure 2 visualizes the difference between a single hormone and multiple hormone pump as an artificial pancreas system. As can be seen, the products are designed as minimally as possible and should not be very obstructive for the user which increases the complexity. According to Infante



(a) (b)

Figure 2: Hormone-pumps [10]
(a) Single hormone insulin pump (b) Dual hormone pump

et al., a study has reported that dual hormone closed loop systems can reduce the time spent in a hypoglycemia. However, the system was only tested with controlled in-clinic exercises. It was associated with more glucose values in hyperglycemic ranges than with a single hormone pump [10]. Given this, dual hormone systems would reduce the time in hypoglycemic events, but reportedly do not eliminate the risk and contrariwise, could lead to hyperglycemia. Dual hormone systems could be beneficial for a subgroup of patients with T1D such as athletes, subjects with a recent history of severe hypoglycemia, or subjects suffering from hypoglycemia unawareness [10].
On the whole, physical activity is still a challenge for single hormone pumps and dual hormone pumps currently are not flawless. Thus, AI can be utilized to prevent exercise-induced hypoglycemia to decrease the risk of severe events. Furthermore, Home and Mehta point out that 24-hour euglycemia is currently not easily managed and suggest the need for better control and prediction algorithms. In particular, those should focus on intra-person variations of BGLs as well as the insulin sensitivity influenced by physical exercise, meal intake, and other physiological or hormone changes such as stress [31].

Exercise

Exercise-induced hypoglycemia is still a major problem, which is mostly caused by the administered insulin, since it is not automatically stopped or adjusted, and still has an action duration. However, physical activity is recommended for improved life quality and fitness. It can help to control BGLs and could increase insulin sensitivity possibly leading to decreased insulin dosage requirements [15, 7]. Furthermore, physical activity is positively associated with a decreased risk of vascular complications and comorbidities [18]. Nevertheless, the management of glucose values can be more difficult during physical activity, resulting in an avoidance of sports due to increased fear and anxiety of

hypoglycemia. Additionally, the fear could cause increased carbohydrate consumption mitigating the advantages of exercising. The main problem is that exercise-induced hypoglycemia might be asymptomatic and not realized [7]. Kelly et al. outline that for patients with T1D an optimal glucose concentration during exercise is between 108-250 mg/dL [33]. Children are said to have an increased risk of exercise-induced hypoglycemia. Cockroft et al. report that in a study 30% of participants had a hypoglycemic event after the exercise and even 30 minutes of activity could increase the risk for hypoglycemic events by 30%. It is highlighted that almost 11.8% of adults and 6.2% of children had at least one severe exercise-induced hypoglycemia per year in 2013 [7].

The challenge is that insulin replacements are not directly secreted into the portal vein as why the insulin levels cannot be rapidly and dynamically decreased. It is mentioned that it could even increase since in patients with T1D, exercise could cause increased subcutaneous blood flow [34]. Here, it is differentiated between resistance exercise and high-intensity exercise. During the first category, glucose is remarked to decrease due to the increased glucose consumption of muscles. During the latter category, the glucose levels may increase since the endogenous glucose release is increased and the rise in muscle insulin sensitivity may be weakened [34]. Therefore, the pump system should notice the start of an exercise session, its type, and its intensity. If the exercise is known, insulin pumps can modify the insulin dosage calculation. Also, if insulin is self-injected, exercise should be ideally planned before. The dosage would depend on the time after the meal. It is suggested to start reducing the administered insulin by 50% if the type and duration are not known. Furthermore, the basal insulin could be reduced by 80% from 40 minutes before until the end of the exercise session [7]. Studies further report the importance of the timing of physical activity. A fasting state could help to prevent hypoglycemia. A session of less than 45 minutes possibly could even result in less risk of hypoglycemia over the next 24 hours after the exercise [7]. Morning exercise could also decrease the risk of hypoglycemic events compared to afternoon exercise sessions [35]. Lastly, it is reported that exercise should be avoided if a hypoglycemic event is experienced in the last 24 hours since it increases the risk of another event. [7].

To sum up this section, Paldus et al. finally suggest the use of physiological signals as additional inputs for algorithms. Lactate, ketones, accelerometry, heart rate, galvanic skin response, skin temperature, and blood volume pulse could inform about the onset, offset, and intensity of physical activity as can be seen in their proposal in figure 3. Lactate is asserted to correlate and increase with exercise intensity. In addition, ketone would inform about the exercise type. Moreover, machine learning can be utilized for

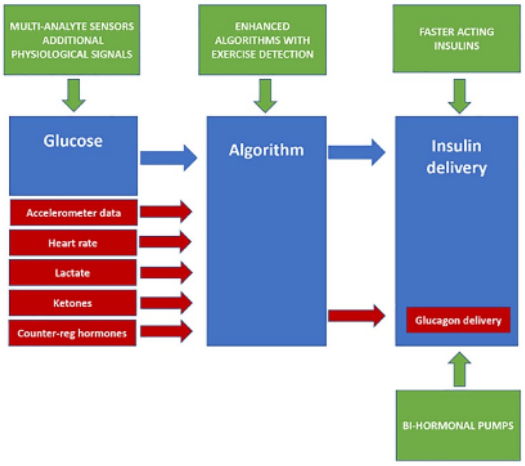


Figure 3: Solution to reduce exercise-induced hypoglycemia [34]

personalized models based on the behavior, profile, and patterns of the user [34]. Lastly, looking at the effect of exercise and the action and duration times of insulin replacement therapies, it is noticed that exercise-induced hypoglycemia can occur 24 hours after the activity and may not be predictable immediately without additional information. Also, the accuracy could be improved if the type of activity is known.

2.2. Machine Learning

Machine Learning is a sub-field of AI in which patterns are learned from the given input data. The models are classified into unsupervised learning, supervised learning, and reinforcement learning [36]. Unsupervised models most often solve clustering problems and are utilized for data pre-processing or dimension reduction. Here, the input dataset is not annotated with a label but is grouped based on the patterns in the data [37]. In healthcare, unsupervised models can be used to find subgroups in patients, or in the case of diabetes, the time series can be grouped based on similar patterns. Here, only the number of wanted groups and the input features are given to the model. Contrariwise, in supervised learning, each sample of the input data is labeled and the model finds a function describing and separating the instances based on their distinctive patterns [37]. Common problems are classification and regression tasks, in which the outcome is either the annotated class or a numerical value such as the glucose value, respectively [36]. In the case of diabetes research, usually, the glucose value is predicted with regression models, and the onset of an event or the onset of a condition is predicted with a classification model. Common machine learning models in diabetes research include Support Vector Machiness (SVMs), support-vector-regressions (SVR), K-Nearest-Neighbor (KNN), and Decision Trees. Those require feature engineering, hence the input features need to be manually computed and selected. In SVMs, a hyperplane is computed to separate the data.

2.2.1. Neural Networks

For advanced problems and larger datasets, Neural Networkss (NNs) are approached, which are function approximation problems and extract features autonomously. A basic NN consists of three layers being the input layer, a hidden layer, and the output layer. Here, the connections between the neurons and layers imitate the connections of neurons in the brain. Besides, a deep learning model is a neural network with multiple hidden layers and deeper connections. Deep learning is often utilized for larger datasets such as image classification and natural language processing or for sequential data like time series [36]. A Multilayer Perceptron (MLP) is the simplest deep learning architecture in which the output neurons of the previous layer are forwarded and connected to the next layer. Each layer computes its own output and the connections of each layer are controlled by a set of weights [38]. The initial weights are mostly selected randomly and updated in the iterative learning process. Here, the error between the predicted and the true output is

estimated through the loss function and the weights are adjusted in the back-propagation process to minimize the error. The iteration stops either with a predefined epoch size or if a predefined condition is met [38, 39]. Fawaz et al. provide the following equation of a general non-linearity transformation which computes the output of each layer:

$$A_{l_i} = f(\sigma_{l_i} * X + b) \tag{1}$$

Accordingly, X represents the data, σ_{l_i} contains the set of weights, b is the bias, and A_{l_i} is the activation function of the neurons for the layer l_i [38]. In the final output layer, an activation function decides if the output should be classification or regression. A softmax layer has the size of the set of all classes and outputs a probability distribution over the classes. Whereas, a sigmoid function is used for regression and outputs a possible number [38]. The basic feed-forward model is visualized in figure 4.

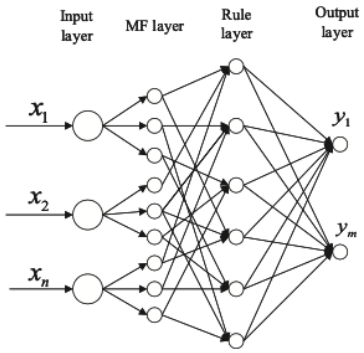


Figure 4: Basic architecture of a Feedforward Network [40]

Most commonly used deep learning approaches for time series classification or regression include 1DCNN and LSTM as can be seen in chapter 3. Therefore, both models are used and compared in this thesis. A typical time series could be described as an input feature vector of $X_t \epsilon \mathbb{R}^{F_0}$ with the length F_0 and a defined time step t which is greater than 0 but not greater than the measured time T for each layer l as T_l. Furthermore, each sequence is either annotated with a class or with an actual true value for a regression output [41].

2.2.2. 1DCNN

CNN are advantageous for autonomously extracting features and reducing the dimension. Thus, feature engineering is not necessary beforehand and the outcome does not depend on the quality of the chosen features. For one-dimensional data, 1DCNN are utilized, which only perform one-dimensional convolutions [39].

CNN usually consists of three primary layers which are the convolutional layers, the pooling layers, and the fully connected layers [29, 30]. Figure 5 shows the architecture and process of a 1DCNN model if utilized for sequential data. First, the convolutional layers

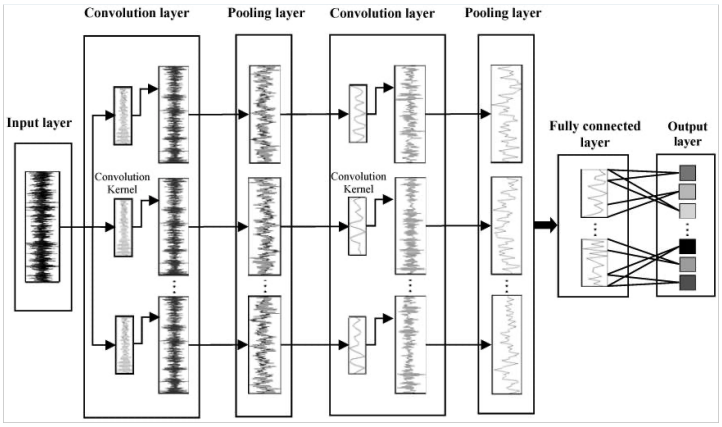


Figure 5: 1DCNN model for sequential data [39]

extract local features and generate a one-dimensional (1D) feature map. The weights are then shared with local connections on the same input feature map [42, 39]. Each kernel is assigned to extract different features from the input feature map. Then, a set of 1D filters is applied over each layer L that analyzes the patterns of the input sequence. The filters for each layer are parameterized by the tensor $W^{(l)}\epsilon\mathbb{R}^{F_l\times dF_{l-1}}$ and biases $b^{(l)}\epsilon R^{F_l}$ [41]. Huang et al. describe a 1D convolutional layer by the following equation in which L represents the layer, $f()$ the non-linear activation function, k the kernels, j the number of kernels, M the channel number of the input x_i^{l-1}, $*$ the convolution operator, and $l\epsilon 1, ..., L$ the layer index [39]:

$$x_j^l = f(\sum_{i=1}^{M} x_i^{l--1} * k_{ij}^l + b_j^l) \tag{2}$$

Here, based on the input size and input structure, a different filter size and number of deepness of layers may be more appropriate. If the data is characteristic enough, too many layers would not increase the performance and could lead to overfitting [42]. Furthermore, the pooling layers are discriminative classifiers and reduce the dimension of feature maps while preserving the most relevant information and enabling a better generalization [42, 38]. However, an inappropriate pooling size could lead to worse performance since important information could be lost [42]. The pooling layer can be local or global. Local poolings reduce the data over a sliding window of the time series. Whereas, global poolings reduce the time series over the whole time dimension to a single

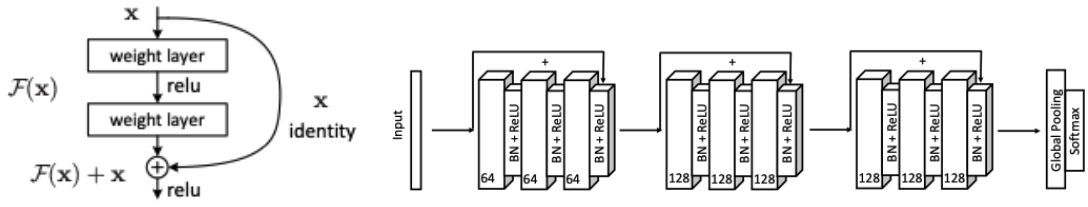
value [38]. Lastly, normalization layers can be approached to enable a fast converge [38]. After the convolutions, fully connected layers can be applied and lastly, the final output layer is added which classifies and outputs the results [39].
The training of the network is summarized as follows. First, after building the architecture, the weights and bias are initialized. Then, a learning rate η is selected and the output for each layer is calculated. In the next step, the weights and bias are updated by back-propagation which simply is based on the gradient descent method described by the following equation presented by Zhao et al. [43]:

$$p = p - \eta \frac{\partial E}{\partial p} \tag{3}$$

CNN are initially applied for images but can be also used to extract patterns in natural language, in time series classification, and for prediction tasks with reported high accuracy. CNN are practical in time series since it can extract local features as well as temporal patterns [44, 38]. Furthermore, CNN are reported to be robust against noisy data and outliers as why they would work well with data acquired by sensors [30]. Additionally, Hwong et al. point out that CNN have the potential to predict blood glucose levels precisely and can be used for diabetes management. They can learn non-linear relationships between the input and the output, which would be crucial in predicting glucose levels [30]. Nevertheless, the disadvantages, that this thesis will possibly experience, are the need for large datasets, large amounts of labeled data, and the overfitting of the trained data which is a challenge for generalized and population-based models [29, 30]. Lastly, compared to most of the machine learning methods, CNN and deep learning, in general, are difficult to interpret and explain [29].

2.2.3. ResNet

Residual Networks (ResNet) are one of the state-of-the-art deep learning models usually applied for image classification. It is also shown to perform well on univariate and multivariate time series data [38]. The architecture is based on stacked CNN layers which are the residual blocks. Then, a shortcut connection is added to each block to overcome the vanishing gradient problem [46]. Figure 6 (a) visualizes the architecture of one residual block in which x denotes the input and the identity mapping of the shortcut connection, $F(x)$ is a function for the stacked layers, and finally $F(x) + x$ is the output [45]. Furthermore, an example network for time series data is given in figure 6



(a) (b)

Figure 6: ResNet architecture
(a) ResNet block [45] (b) 3 block ResNet for time series data [46]

(b). However, the same drawback as in 1DCNN and CNN can be seen. It is reported that small datasets of time series data can easily result in overfitting [46].

2.2.4. LSTM

In feed-forward neural networks and CNN time series, each time stamp of the time series has its own weight as to why the temporal information is lost [38]. Hence, a network architecture with backward connections like RNN may be more appropriate for considering contextual information [30]. Here, the input contains the information at the previous time-steps which is a great advantage for sequential data [36]. Karim et al. mention that RNN maintain a hidden vector h which is updated at time step t for the final prediction. They define RNN in the following equation in which tanh (hyperbolic tangent function) is the activation function, W is the weight matrix, I is the projection matrix, and y_t is the outcome [41]:

$$h_t = tanh(Wh_{t-1} + Ix_t), \tag{4}$$

$$y_t = softmax(Wh_{t-1}).$$

Here, a softmax function is applied for a classification model but it can be exchanged with a sigmoid activation. Furthermore, deeper architectures can be created if h is inputted to another RNN as defined in the following equation by Karim et al. [41]:

$$h_t^l = \sigma(Wh_{t-1}^l + Ih_t^{l-1}) \tag{5}$$

However, simple RNN have a gradient vanishing and exploding problem during back-propagation which hinders the learning of optimized results. Therefore, more advanced

models are investigated. A state-of-the-art recurrent architecture is the LSTM model which can save long-term information and forget a part of their hidden state by integrating gating functions into their state dynamics [36, 30, 47]. Thus, a memory vector m is included which controls the state updates and outputs. According to Karim et al., the computation at time step t would be the following equations, in which the sigmoid function is represented by σ, the $\odot$ represents element-wise multiplication, W^u, W^f, W^o, and W^c are the recurrent weight matrices and I^u, I^f, I^o, and I^c represent projection matrices [41]:

$$
\begin{aligned}
g^u &= \sigma(W^u h_{t-1} + I^u x_t) \\
g^f &= \sigma(W^{tf} h_{t-1} + I^f x_t) \\
g^o &= \sigma(W^o h_{t-1} + I^o x_t) \\
g^c &= tanh(W^c h_{t-1} + I^c x_t) \\
m_t &= g^f \odot m_{t-1} + g^u \odot g^c \\
h_t &= tanh(g^o \odot m_t)
\end{aligned}
\tag{6}
$$

x_t is the input while h_t is the output for each time step. Furthermore, g^u, g^f, and g^o are the input, forget, and output gates, while g^c is a vector with new possible values for the cell state [47]. Figure 7 visualizes the architecture of an LSTM model. Here, m_t is replaced with $c(t)$, and g^o, g^f, g^u, g^c with $o(t)$, $f(t)$, $i(t)$, c_in respectively, while the final output h_t by $h(t)$. Additionally, Aiello et al. highlight that "*during temporal unfolding, both h_t and m_t are passed to the temporal replica of the next cell in the fold*" [47]. A variation of conventional forward LSTM models are Bidirectional Long Short-Term

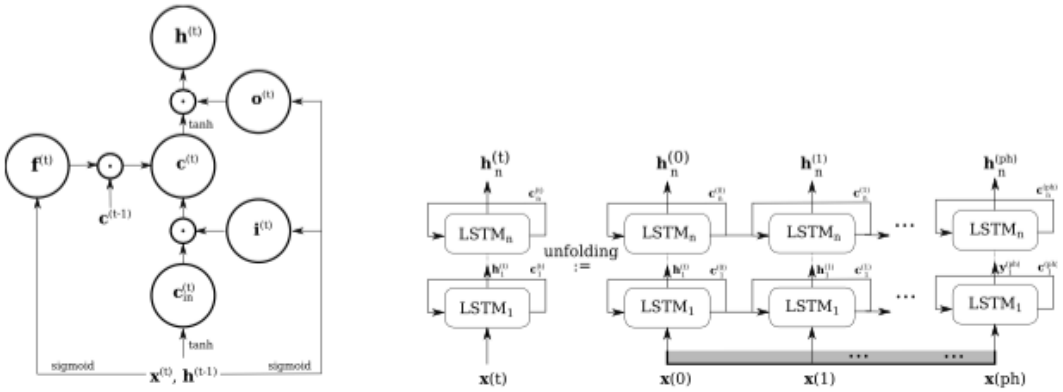


Figure 7: Architecture of an LSTM model and its cells [47]

Memory (BiLSTM) models, consisting of two LSTM layers. Here, the first layer is applied

to train the input data in its original direction from the past to the future, while the other layer takes the input data in its revered direction, thus training backward [48]. RNN models are reported to be superior in natural language processing and speech recognition [36]. Moreover, they are mostly utilized for regression and prediction tasks of time series data in the financial or healthcare domain. Hwong et al. further point out the potential in glucose time sequence data since the long-term temporal dependencies are considered [30]. But, RNN and LSTM are mostly applied for forecasting and not for classification regarding diabetes research [38] as also seen in chapter 3. Disadvantages include overfitting with inappropriate regularization, the high computational cost of the training process, and the sensibility to noise [30, 38].

As can be seen, deep learning models have common advantages such as automatic feature extraction and learning from large and complex datasets. Especially, in diabetes research they could be superior for multivariate time series. However, compared to conventional machine learning models, deep learning models are more difficult to interpret [30]. When comparing 1DCNN and LSTM, it can be said that 1DCNN is computationally more efficient as it can reduce the features and neuron connections between layers enormously. Also, it can better generalize [40, 39, 36], and is robust against noise, but it cannot capture temporal context that well. In the state of art models both architectures are combined to extract features with 1DCNN and compute temporal dependencies with LSTM [36, 30].

3. State of the Art

As it was demonstrated, diabetes care is improved with technological advances and the inclusion of artificial intelligence, enabling better self-control. In the last decades, much research has been conducted in the diabetes field. The literature review shows that studies can be classified into glucose forecasting, insulin forecasting, classification of the onset of diabetes, and lastly, the classification of the onset of adverse events. In particular, diabetes research has progressed with methods predicting blood glucose values. Here, Bremer and Gough are reported to be pioneers considering past glucose values in 1999 [49, 50]. In the same year, Tresp et al. were among the first researchers to apply RNN and a linear error model on time series to imitate the blood glucose metabolism [51].
In the following, recent studies contributing to the field of diabetes and hypoglycemia research will be presented. This thesis classifies the onset of hypoglycemia, therefore only glucose forecasting and hypoglycemia prediction models are reviewed. Related work was searched in the "ScienceDirect" database, in the "IEEE Xplore" dataset and on "Google Scholar" with the following search terms: ("hypoglycemia" AND "classification" AND "glucose" AND "insulin" AND ("exercise" OR "activity") AND "diabetes" AND ("machine learning" OR "deep learning")), ("hypoglycemia" AND "classification" AND "glucose" AND "insulin" AND "exercise"). Additionally, the keyword ("long-term") was added. The results were filtered to studies being published since 2020 since the interest lay in the most advanced methods. Some related studies published from 2013 onward were also added after being highlighted in other reviews. The final set consisted of 19 studies of which 10 focus on forecasting glucose values, seven focus on the classification of events, and three studies do prediction and classification within the same work.
The selected studies are classified into short- and long-term PHs and for each PHs, the best methods are highlighted in subsection 3.1 and subsection 3.2, respectively. Furthermore, the applied input sequence length and input features are compared in subsection 3.3. Lastly, identified shortcomings are summarized in subsection 3.4.

3.1. Short-Term Prediction Horizons

Short-term PHs are defined as a horizon from 30-120 minutes supporting timely preventive action of a carbohydrate intake. Table 1 presents all studies covering short-term PHs which are ordered after their publication date. In total nine studies focus on regression while two studies predict and classify. Here, it can be seen that most works only predict up to 60 minutes of BGL, which are six in total [54, 55, 56, 59, 61, 60]. Whereas Swain et

Table 1: State of the Art: Short-term PHs

Study	**Aim**	**Model**	**Input**	**Performance of PHs**			
				30	**45**	**60**	**≤120**
Georga et al., 2013 [52]	prediction (RMSE)	SVR	15	6.03		7.14	7.62
Zarkogianni et al., 2015 [53]	prediction (RMSE)	SOM	10	11.42		19.58	31.00
Munoz Organero et al., 2020 [54]	prediction (RMSE)	LSTM	40/ 9	6.42		11.35	
Seo et al., 2021 [55]	prediction (RMSE)	CNN	29	17.8	23.2	28.1	
Nemat et al., 2022 [56]	prediction (RMSE)	CCMBA + LSTM	6/6	20.09/ 19.439		32.901/ 34.791	
Jaloli et al., 2022 [57]	prediction (RMSE)	CNN + LSTM	168/ 59	9.28/ 9.81		16.51/ 18.32	23.45/ 25.12
Phadke et al., 2022 [58]	prediction (MAPE)	ANN	12	0.037		0.069	0.148
Zhu et al., 2022 [59]	prediction (RMSE)	FCNN	12/12/ 25	18.64/ 20.23/ 20.25		31.07/ 35.4/ 34.03	
Zhu et al., 2022 [59]	classification (SE, PRE)	FCNN	12	84.09/ 65.69		68.58/ 60.64	
Zhu et al., 2022 [60]	prediction (RMSE)	LR	12	20.92	28.99	35.28	
Zhu et al., 2022 [60]	classification (SE, PRE)	LR	12	0.76, 0.66	0.72, 0.58	0.70, 0.56	
Chen et al., 2023 [61]	prediction (RMSE,MAPE)	CNN	10	10.51, 0.029		15.98, 0.032	
Swain et al., 2023 [62]	prediction (MAPE)	p-LSTM	5	0.0628	0.0846		

Abbreviation: PH = Prediction Horizon

al. have a maximum PH of 45 minutes [62], Jaloli et al. predicted up to 90 minutes [57] and lastly, three studies forecast 120 minutes of glucose values [52, 53, 58]. Among the

PHs, different algorithms are approached which are Support Vector Regression (SVR), Linear Regression (LR), NN. Furthermore, deep learning models are applied, specifically LSTM and CNN, but also Artificial Neural Networks (ANN), Fast-adaptive and Confident Neural Networks (FCNN), Self Organizing Maps (SOM), and hybrid learning methods of CNN and LSTM. A trend for utilizing more deep learning and hybrid models is noticed with advancing time.

The best performance for forecasting glucose values with a PH of 30 minutes, 60 minutes, and 120 minutes is obtained with an SVR trained with 10-fold cross-validation by Georga et al. obtaining a Root Mean Squared Error (RMSE) of 6.03 mg/dL, 7.14 mg/dL, and 7.62 mg/dL, respectively. Their input data consists of 15 patients from the METABO dataset of whom glucose, insulin, carbohydrate, and physical activity were collected for 10 days. Additionally, the accuracy of correctly predicted hypoglycemia is presented with 87%, 83%, and 85% for 30, 60, and 120 minutes, respectively. These results indicate that more than 80% of hypoglycemic events could be predicted accurately even 2 hours before [52]. Munoz et al. obtained a similar RMSE value for a PH of 30 minutes with 6.42 mg/dL testing in nine real patients from the D1NAMO dataset, but their model is not that stable for a longer PH of 60 minutes, as it increases to 11.35 mg/dL. The input features consist of glucose, insulin data, and meal intake collected for 4 days. Contrariwise, Munoz et al. applied a hybrid model integrating mathematical models to a LSTM RNN network to simulate the metabolic process of glucose. They trained with a hold-out validation using 70% of data for training and 30% for testing. Furthermore, they initially trained on 40 virtual patients achieving much better results of 3.45 mg/dL and 4.72 mg/dL, for 30 and 60 minutes, respectively. But, those results cannot be taken as a baseline for comparison [54].

The third best results which are clinically acceptable, are noticed in Jaloli et al. achieving RMSE values of 9.28 mg/dL, 16.51 mg/dL, and 23.45 mg/dL for 30, 60, and 120 minutes, respectively. They approached a hybrid LSTM and CNN model with the input data of 168 participants from the Replace-BG dataset of whom glucose, insulin, and carbohydrate data were collected in free-living conditions. Their dataset is the largest utilized dataset among all studies. Thus, the results represent more patients and could be used for a population-based model. It is also induced that the proposed method is indeed suitable for analyzing time series patterns to forecast glucose values. Moreover, the DIAdvisor dataset of 59 subjects of whom data was collected in hospital settings, was used for better validation. Accordingly, for the test dataset similar outcomes were obtained with RMSE values of 9.81 mg/dL, 18.32 mg/dL, and 25.12 mg/dL for 30, 60, and 120 minutes,

respectively. Consequently, good stability and generalization are presented, and the model does not overfit much to the trained data, even if both datasets were collected under different settings [57]. Compared with the results of Georga et al., the RMSE value for 120 minutes of Jaloli et al. is increased by 15.83 mg/dL. This variation can either depend on the dataset, input features, or the utilized machine learning model. Hence, it cannot be highlighted if conventional machine learning models or deep learning models are better for a short-term PH of up to 60 minutes. Nevertheless, within their study, the hybrid LSTM CNN model outperformed LSTM, autoregressive model and exogenous input (ARX), and SVR models [57]. The last relevant results are reported by Chen et al. utilizing a hybrid model of CNN and Transformers which are usually used for natural language processing. They achieved an RMSE value of 10.51 mg/dL and 15.58 mg/dL, for 30 and 60 minutes, respectively which is similar to the performance of Jaloli et al. but improved for the PH of 60 minutes. However, only data from 10 virtual patients is used [61], while the model of Munoz et al. is not outperformed.

Coming now to the applied classification models, only two studies are presented using FCNN and LR, both by Zhu et al. [59, 60]. The regression outcomes are not that clinically acceptable if compared with the other studies, but the results for hypoglycemia classification indicate that more than 70% of hypoglycemic events could be predicted even one hour before. Zhu et al. utilized FCNN based on an attention-based RNN in their first work. The performance is validated with the OhioT1DM, ARISES, and ABC4D datasets, only using glucose measurements as input. They trained the model with a hold-out validation with 80% as training data and 20% as validation data. For 30 and 60 minutes, hypoglycemia is classified with sensitivity and precision of 84.09% and 65.69%, and of 68.58% and 60.64%, respectively for the 12 patients of the OhioT1DM dataset [59]. In the other study, linear regression is applied having worse results in glucose forecasting. In contrast, more input features are used including insulin, carbohydrates, and parameters from a wristband like heart rate, galvanic skin response, and physical activity. For the classification of hypoglycemia, achieved sensitivity and precision are 76.08% and 65.65%, for a 30-minute PH, and 70.30% and 56.20%, for a 60-minute PH, respectively [60]. Both studies have similar performance with a higher sensitivity than precision. Hence, most of the hypoglycemic events are foreseen by the system. Nevertheless, the precision needs improvements, as it is not that efficient to have false positive alarms often. This can be caused due to the imbalance of the dataset and less available hypoglycemic events compared to hyperglycemia and euglycemia. Comparing the outcomes, it cannot be seen if the included exercise data in the LR model has any impact since it is asserted that the

FCNN, in general, performs better than a basic LR algorithm.
Conclusively, Georga et al. achieved the best performance despite publishing their work in 2013 with an SVR [52], followed by Munoz et al. using LSTM and transfer-learning [54]. Additionally, the worst results are obtained by Nemat et al., and Swain et al. using only six and five patients of the OhioT1DM database, respectively and both utilizing deep learning models indicating that deep learning should be used for larger datasets [56, 62]. Lastly, studies often use multiple datasets for better validation and generalization, and more input data seems to have better performance.

3.2. Long-Term Prediction Horizons

Chapter 2.1.3 showed that insulin replacements can be active for up to 42 hours, thereby affecting glucose metabolism. Furthermore, exercise can influence glucose metabolism for up to 24 hours, which is why longer PHs can enable better day-to-day management, and better choice of meals and insulin dosages, as discussed in 2.1.3. Here, the long-term PH ranges from 2-24 hours. Looking at table 2, it is noticed that applied models mainly include conventional machine learning algorithms such as Decision Trees, Random Forest, Support Vectors, and LR. Neural networks and deep learning models such as CNN and ANN are utilized, but rarely in comparison. It is noticed that none of the presented studies approaches an LSTM network for longer PHs probably due to their long computation time. Among presented works, seven studies do classification [63, 64, 65, 28, 6, 67, 11]. Phadke et al. forecast glucose values with a regression model and is the continuation of the study in table 1, since they have predicted from 30 minutes up to 24 hours [58]. Lastly, Tyler et al. focus on regression and classification [66]. The works cannot be compared that easily since not every study approaches the same classification task. Oviedo et al., and Vehi et al. for instance, classify between level 1 (below 70 mg/dL) and level 2 (below 54 mg/dL) hypoglycemia [63, 65] while Bertachi et al., Vehi et al., and Parcerisas et al. approach nocturnal hypoglycemia classification [64, 65, 28], and Tyler et al., and Piersanti et al. approach exercise-induced hypoglycemia [66, 6]. Finally, Vehi et al., Alvardo et al., and Felizardo et al. classify the risk of the occurrence of hypoglycemia in the next 24 hours [65, 67, 68].
Phadke et al. forecast glucose values and cannot achieve better results compared to already presented studies but have stable results among the PHs. They are the only researchers considering short- and long-term PHs up to 24 hours for glucose forecasting. The achieved Mean absolute percentage error (MAPE) values are 0.037, 0.069, 0.149, 0.215, and 0.134 for 30, 60, and 120 minutes, and 3 and 24 hours, respectively. Considering

Table 2: State of the Art: Long-term PHs

Study	**Aim**	**Model**	**Input**	**Performance of PHs**			
				≤2h	**≤4h**	**6h**	**24h**
Oviedo et al., 2019 [63]	classification (SP, SE) (level 1, level 2)	SVC	10				0.79, 0.71/ 0.81, 0.77
Bertachi et al., 2020 [64]	classification (SE, SP)	SVM	10			0.78, 0.82	
Vehi et al., 2019 [65]	classification (SE, SP)	GE, SVM, ANN, NCD	10/6/ 100	0.48, 0.93	0.69, 0.80/ 0.75, 0.81	0.44, 0.86	0.99, 0.92
Tyler et al., 2022 [66]	classification (SE, SP)	MARS	20	0.73, 86	0.56, 96		
Tyler et al., 2022 [66]	prediction (RMSE)	MARS	20	18.7	23.0		
Phadke et al., 2022 [58]	prediction (MAPE)	ANN	12	0.148	0.215		0.134
Pakerisas et al., 2022 [28]	classification (F1,SE,SP)	SVM	10			0.76, 0.74, 0.76	
Piersanti et al., 2023 [6]	classification (PRE, SE, SP, F1)	DT	50				0.87, 0.76, 0.87
Alvardo et al., 2023 [67]	classification (accuracy)	CNN	4				0.79
Felizardo et al., 2023 [68]	classification (SE, SP)	RF + SkNN	54				0.45, 0.89

Abbreviation: PH = Prediction Horizon

the increase in PHs and the change in performance, the prediction of 24 hours seems reasonable, is even better than with 2 hours, and significantly better than with 3 hours. They applied an ANN model while utilizing glucose, insulin, carbohydrate, physical activity, and self-reported data from 12 patients of the OhioT1DM dataset [58].
Moving on to the results of classification systems, a sensitivity and F1-measure of at

least 70% is achieved in most of the studies [63, 66, 28, 6], indicating that at least 70% of all hypoglycemic events could be predicted. However, some studies can only achieve a sensitivity up to 40% or 60% [65, 66, 68] which still needs improvement.

In the case of severity classification, Oviedo et al. used a Support Vector Classifier (SVC) for a private dataset with 10 patients of whom glucose, insulin, and carbohydrate data were collected under free-living conditions to estimate the risk of hypoglycemia within the next 24 hours. The model was trained individually on each patient with a hold-out validation using 80% for training and 20% for testing. A median specificity of 79% and a sensitivity of 71% for level 1 hypoglycemia was obtained. The sensitivity and specificity for level 2 hypoglycemia were 81% and 77%, respectively [63]. Vehi et al, have considered two different PHs of 1 hour and 4 hours for predicting level 1 and level 2 hypoglycemia. For the 1-hour classification, a model based on grammatical evolution (GE) was approached while using glucose and insulin data, carbohydrate information, and the circadian rhythm of patients. The dataset consists of 100 virtual patients. They present a model based on a problem-specific free-context grammar and a fitness function consisting of a glucose-specific mean squared error. The model is trained with a hold-out validation using 66.6% of the data for training and 33.3% for testing. A population accuracy of 86.1%, sensitivity of 48.5%, and specificity of 93% is achieved [65]. Furthermore, postprandial hypoglycemia 4 hours after the meal was classified with SVMs. The input data consisted of 10 patients of whom glucose, insulin, and carbohydrate were collected. They trained a population model with a hold-out validation using 80% of the data for training and 20% for testing. Finally, sensitivity and specificity of 69% and 80% are achieved for level 1 and 75% and 81% for level 2 hypoglycemia, respectively [65]. Despite defining a longer PH, Oviedo et al. achieved a better sensitivity and almost the same specificity. Thus, it can be asserted that support vectors can predict at least 69% of hypoglycemic events. Also, the severity of the event 4 hours and even 24 hours before can be predicted allowing the person to plan their day and meal accordingly.

Tyler et al. used a PH of 4 hours as well. They classify the risk for hypoglycemia induced by aerobic exercise 40 minutes after the start of the exercise and 4 hours after the exercise. The utilized dataset consists of 20 patients' glucose and physical activity data. Furthermore, the prior history of exercise information is used for training a Multivariate adaptive Regressionssplines (MARS) model. The models are trained with a hold-out validation. The patients performed 8 identically designed clinical exercises and had the same schedule while having measured their data. A PH of 40 minutes achieves a sensitivity, specificity, and accuracy of 73%, 95%, and 88% for individualized models,

and 73%, 86%, and 81% for a population model, respectively. Both models have similar results and indicate that at least 80% of hypoglycemic cases could be known before starting the exercise. The performance for 4 hours after the exercise was 56%, 96%, and 84% for the personalized models for sensitivity, specificity, and accuracy, respectively. Moreover, for a 16-fold cross-validation-based population model a sensitivity of 79%, a specificity of 61%, and an accuracy of 69% was achieved. Comparing both models, the overall performance of personalized models is better but the population model has increased sensitivity by 23%. In total, 79% of all hypoglycemic events could be predicted which can be caused within the next 4 hours by exercise [66]. Therefore, it is asserted that personalized models perform better and a high sensitivity needs more training data so that all the variations between and within persons can be learned. It is noticed that Tyler et al.'s results are much better in sensitivity compared to Vehi et al.'s. Furthermore, Tyler et al. have predicted glucose values by the same model and the same PHs and achieved an RMSE value of 18.7 mg/dL for 30 minutes which does not outperform previous studies. However, an RMSE value of 23.0 mg/dL is reported for 4 hours which surprisingly is better than the 2-hour prediction of Jaloli et al. indicating that a MARS model can be utilized for long-term prediction and classification [66].

Turning now to nocturnal hypoglycemia, a PH of 6 hours is used by all studies, and mostly an input sequence length of 6 hours is chosen. All presented studies utilize the glucose and physical activity data of 10 patients as input. Bertachi et al. applying a SVM achieved the best performance with a population model. They obtained a sensitivity of 78.75%, a specificity of 82.15%, and an accuracy of 80.77%. The performance of SVMs is compared with NLP and outperforms it. As limitations, the less available data and the exhaustive feature selection for SVMs are reported [64]. Vehi et al. approached ANN and overcame one of the limitations by not doing manual feature engineering. Reported population outcomes are 80.1%, 44.0%, and 85.9% for accuracy, sensitivity, and specificity, respectively [65]. Vehi et al. have better specificity and similar accuracy but the sensitivity is much lower with a difference of 34.75%, hence 56% of hypoglycemic events are missed. Lastly, Parcerisas et al. used SVMs as well and compared them to different machine learning methods including LSTM. Likewise, SVMs performed best. The model was trained with a hold-out validation with 80% of data used for training and 20% for testing. A sensitivity and specificity of F1-measure of 74% and 77% are achieved for the population model, respectively. The median specificity for individualized models decreased to 68% while the sensitivity performed the same [28]. They cannot outperform Bertachi et al. but obtain better sensitivity than Vehi et al., as to why it is induced

that SVMs may be better than simple ANN, NLP, and LSTM models for the long-term classification of hypoglycemia with a PH of 6 hours. Neural networks could possibly not learn patterns well with less data leading to more missed cases and decreased sensitivity or overfitting. With the model of Bertachi et al. 78% of all hypoglycemic events could be predicted before going to sleep.

The last field of interest was the risk assessment of hypoglycemia within the next 24 hours which was the focus of four studies. In this context, Vehi et al. clustered different glucose profiles of patients while using a data mining model based on a normalized compression distance (NCD). They used 100 virtual patients while utilizing glucose and insulin data. The final performance of the classifier was 92% and 99% for specificity and sensitivity, respectively [65]. Contrariwise, Piersanti et al. used a decision-tree-based model to estimate the long-term prediction risk of exercise-induced hypoglycemia. They used a dataset of 50 children and teenagers from the Diabetes Research in Children Network (DirecNet) multi-center study group and are the first group to study children in this literature review. Hypoglycemia was defined with a threshold of 60 mg/dL and participants experiencing hypoglycemia within the next 24 hours after exercising were labeled as hypoglycemia. A model trained with Leave One Out Cross-Validation (LOOCV) achieved results of 85.5%, 87.2%, 86.9%, 87.2%, 76.1%, and 86.9% for AUC, classification accuracy, precision, sensitivity, specificity, and F1-measure, respectively [6]. These could be very good results from a clinical perspective, as almost 80% of all hypoglycemic events occurring due to exercise could be predicted one day before and the control of glucose values is much more difficult with children. Furthermore, this group has used more input data than most of the studies presented. Alvardo et al. used only 4 patients of whom only data of glucose was measured as why the method cannot validate a general appliance. However, the method itself is very different compared to usual time series classification methods and even with fewer participants, good results are obtained. They utilized a transformer function to generate an image of the time series sequence and apply a CNN model. The images were labeled as hypoglycemia if a hypoglycemic event had occurred within the day by a threshold of 70 mg/dL. The model was trained with hold-out validation in which 75% of the images were used for training, 15% for validation, and 10% for testing. The average classification accuracy of the validation process was 80%, while an accuracy of 78.78% was achieved for the test data. An accuracy of 88% is reported to be the best performance, while an accuracy of 73% is said to be the worst performance of all cases [67]. Nevertheless, the sensitivity and precision are not computed which disables a comparison. Finally, Felizardo et al.

applied an ensemble model with 2 classifiers which are Random Forest and Subspace k-Nearest neighbor (RF and SkNN). They utilized the University of California Irvine diabetes dataset of 54 patients whose data of glucose, carbohydrate, exercise, and therapy inputs as contextual information was collected. The model was trained with leave one patient out validation and it was classified between 3 classes, being no risk, risk for hypoglycemia, and hypoglycemic event. Hypoglycemia was defined by a threshold of 75 mg/dL to not miss any events. For the reported results, only the produced outcomes for 23 subjects who had an accuracy higher than 60% and false alarms less than 30% were considered. Thus, an average accuracy, sensibility, specificity, and false alarm rate of 75.3%, 45.4%, 89.4%, and 13.5%, are achieved respectively. Overall, 76.2% of all events were predicted. It is pointed out that only for 53% of patients more than 70% of the events could be predicted. For 52% of those patients, a false alarm rate of less than 15% was achieved indicating high variations between patients. Consequently, the model cannot be applied for each individual with the same performance [68].

Thus, Oviedo et al. and Bertachi et al. achieved the best population models in which more than 70% of the events could be predicted 24 hours before. Both of those studies utilized conventional machine learning methods being SVM and DTs. Additionally, for the classification of nocturnal and postprandial hypoglycemia, a SVM performed best for sensitivity. Nevertheless, there were not many studies utilizing deep learning models for a better comparison of classification and regression applied for long-term PHs. Alvardo et al. achieved good accuracy but did not present more metrics. The main drawback in neural networks is that available datasets do not have sufficient training samples which leads to overfitting, and decreased learning performance and sensitivity outcomes.

Altogether, comparing short- and long-term PHs in diabetes research, glucose forecasting is more often utilized for shorter PHs up to 120 minutes. In contrast, classification is more often applied for longer horizons up to 24 hours.

3.3. Variables and Features Impacting the Performance

It is noticed that the input sequence length is highly relevant for the final accuracy of the model and depends on the targeted PH. Not every study informs about their input sequence length since also not every model works with time series data. For short-term prediction and classification, it can be seen that Munoz et al. use a sequence length of 9 hours for a maximum PH of 1 hour [54], Jaloli et al. use 3 times the PH [57], and Netmat et al. use 90 minutes of prior history data to predict a maximum of 1 hours of glucose values [56]. Lastly, Vehi et al. use 2 hours prior measurements for the classification of 1 hour. Now turning to long-term prediction and classification, Vehi et al. utilize 1 hour of glucose data before the meal to predict postprandial hypoglycemia 4 hours after the meal [65]. Bertachi et al., Parcerisas et al., and Vehi et al. extract 6 hours of history data before sleeping to classify nocturnal hypoglycemia within 6 hours after sleeping [64, 28, 65]. And finally, all studies having a PH of 24 hours use 24 hours of prior data [63, 65, 11, 67]. Hence, it can be asserted that an input length at least as long as the PH should be utilized. Moreover, double as the input seems reasonable for short-term prediction and classification and it can be asserted that longer sequences lead to better results ranging from 3 to 9 hours. For long-term input sequence lengths, at least the time of PH seems to be reasonable.

Concerning the input data used, almost all studies reported that the information regarding insulin dosage increased the performance compared to using only glucose. In addition, physical activity has been reported to further improve performance and stabilize the model [53, 52, 60, 28]. Zarkogianni et al. point out that including exercise information could mainly influence the performance of hypoglycemia identification [53]. Georga et al. assert that glucose may be appropriate for short-term PHs but more input variables should improve the performance for longer PHs [52]. Nocturnal hypoglycemia may be related to daily physical activity and multiple daily insulin dosages [64, 28]. Lastly, Phadke et al. say that uni-variate data cannot achieve the same results as multi-variate data for a PH greater than 45 min, and highlight that more features than only glucose should be utilized for better performance [58]. Conclusively, often information about glucose, insulin, and carbohydrates is used as input data, followed by exercise, and lastly, three studies only use glucose [55, 59, 67]. Heart rate and galvanic skin response are rarely used.

3.4. Identified Research Gaps

As could be identified there are some limitations and shortcomings in diabetes research. The first major challenge is the insufficient size of datasets, which could lead deep learning models to overfit and disable generalizable pattern learning. Currently, most available multi-variate datasets have only data collected from 10 to 15 patients. Thus, a population-based model might not achieve very good performance. However, personalized models could not be feasible due to small datasets per subject. Especially, studies focusing on hypoglycemia research face this shortcoming, since hypoglycemic events do not occur as often as hyperglycemia and the dataset becomes imbalanced.

The second limitation is the possible time lag of CGM devices, as also presented in chapter 2.1.3. Hence, regression models need to be very accurate, but presented studies achieve an approximate mean RMSE of 10-15 mg/dL, while the best RMSE values are between 6-9 mg/dL for 30 minutes. With this in mind, short-term hypoglycemia prevention might be challenging and not that precise. In contrast, a classification model only analyzes the patterns in data and predicts the risk for an adverse event as to why the numerical differences are less relevant. Furthermore, classification models can be more robust against noise and missing data, being shortcomings of wearable data. Lastly, classification methods can work better with imbalanced data and simplify the problem by focusing on the event only [69].

The third observed shortcoming is that most studies focus on short-term PHs of 30-60 minutes. These can enable preventive self-actions but cannot be used for the adjustment of insulin dosages, suitable meals, and exercise duration listed as the main causes of hypoglycemia. Furthermore, most long-term PHs range only up to 6 or 24 hours and are based on binary classification or at most three classes. Particularly, only one PH is used in one model, and the larger the PH, the worse the results get. Subsequently, none of the presented studies have integrated multiple PHs in one model, which should be possible using a classification system.

Lastly, it can be seen that presented works prefer machine learning over deep learning, as why 1DCNN, LSTM, and hybrid models of CNN and LSTM are not often applied for long-term classification tasks while those could learn long-term relationships better and may be suitable for longer input sequence lengths. But, it can be seen that with presented studies SVMs obtain the best performance while neural networks need an improvement for sensitivity as this metric is often worse than specificity.

4. Methodology

Following the presented shortcomings, this thesis approaches a different perspective by classifying the time to the onset of hypoglycemia while including multiple prediction horizons. Thus, if a person is alerted to have hypoglycemia in the next 4 hours and preventive actions do not reduce the risk, the model would predict the risk 1-2 hours or 30-15 minutes before hypoglycemia until no risk is assessed anymore. If the risk is foreseen 24 hours before, it could help to manage meals and sports activities. Whereas, a PH of 5-12 hours before the onset of the event can help to adjust insulin dosages. Therefore, multiple PHs can enable a better decision of daytime activities, can be useful for artificial insulin pumps as severe and adverse events could be predicted before giving the dosage, and can allow short-term prevention by a glucose intake.
This chapter presents the used dataset, the applied pre-processing steps, and the distribution of classes in subsection 4.1. Then, subsection 4.1.2 introduces the method for the correlation between the glucose concentration, the basal insulin dosage, the bolus insulin dosage, and the magnitude of acceleration values. Subsection 4.2 explores the applied models[1] and subsection 4.3 explains the used metrics. Finally, the main expected limitations of this work are listed in subsection 4.4. The process flow chart visualizing the pipeline of this thesis can be seen in figure 8.

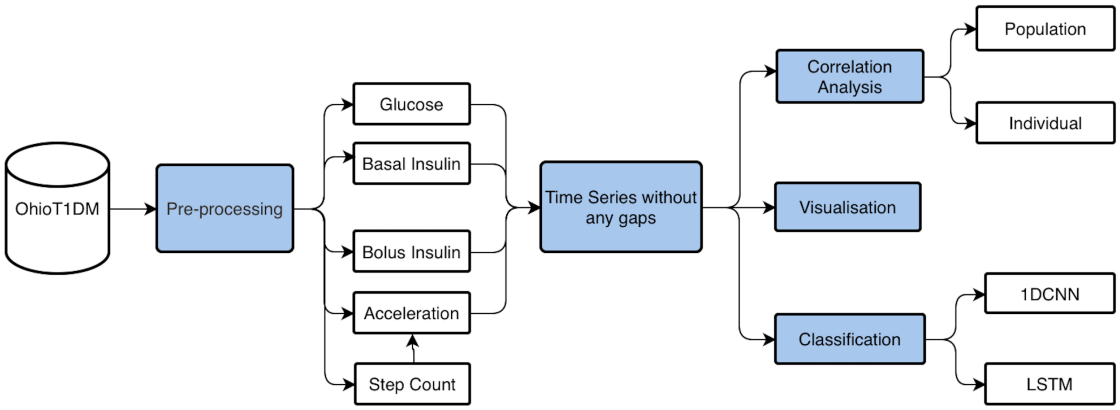


Figure 8: Pipeline of this thesis

[1]The code can be found in https://github.com/Mirai22/Hypoglycemia_Detection_MA

4.1. Dataset: OhioT1DM

Looking at previously utilized databases collecting data in free-living conditions in chapter 3, this thesis has decided to use the most often chosen OhioT1DM dataset because features such as glucose, insulin dosage, and exercise were collected. It contains data from only 12 patients, but as not that many public datasets with more input data are available, the state-of-the-art standards are met, unless own data is collected. The D1NAMO dataset contains only data of nine patients with T1D and insulin information is not collected, and virtual simulators do not have exercise data. Furthermore, as could be seen, if models trained on virtual data are tested on real data, the same accuracy range is not achieved.

In this context, the OhioT1DM dataset contains data of six patients of whom data was collected in 2018, and another six patients of whom data was later collected in 2020. The patients' age ranges between 20 and 80 years while mostly 40-60 are representative. From all of the 12 patients, seven subjects are males and five subjects are females. Accordingly, each subject had data recorded for 8 weeks. A CGM device estimated the substantial glucose values every 5 minutes. Additionally, the subjects have worn an insulin pump and a fitness tracker or health wristband to record physiological parameters and vital signs. The dataset utilized the Medtronic Enlite CGM sensor for all patients, three patients have worn the Medtronic 630G pump and nine patients have worn the Medtronic 530G pump, hence there is no great variation of insulin pumps allowing a more uniform data representation. The pumps have recorded basal and bolus insulin. The basal rate is reported to be infused continuously until a new basal rate is set. Moving over to the wearables, the 2018 cohort was given the Basis Peak fitness bands collecting data in 5-minute intervals, while the 2020 cohort was given the Empatica Embrace collecting data in 1-minute intervals. Both wristbands measured the galvanic skin response, the skin temperature, and the subject's sleep time. But as a difference, the Basis Peak band estimated the heart rate, the air temperature, and the step count while the Empatica Embrace band collected the magnitude of acceleration. Additionally, self-inputted data is available, such as the meal intake as carbohydrate data and the meal time, glucose values by finger pricks, self-reported times, duration and intensity of exercise, sleep, work, stress, and illness. The data was stored in XML files. Lastly, it is reported that the time and the month information is shifted, as to why the date information cannot be used as a feature [70]. Tables 3 and 4 summarize the number of total instances as well as the number of hypoglycemic data points that are less or equal to 70 mg/dL for the 2018 cohort and 2020 cohort, respectively. Here, the intra-person variations can be

seen and it is foreseen that the model can be influenced by a bias since subjects 540, 567, and 575 experienced the most time in hypoglycemic states while subjects 544, 584, and 588 experienced significantly less. Lastly, the total time of missing glucose data is summarized per patient.

Table 3: Number of samples in the OhioT1DM dataset: 2018

Subject ID	559	563	570	575	588	591
Number of Hypoglycemic events	518	329	227	1173	136	570
Total time of missing glucose values in hours/days	137/ 5.7	91/ 3.8	63.6/ 2.6	113.8/ 4.7	46/ 1.9	166/ 6.9
Number of Samples	12792	14365	13500	13283	15295	13037

Table 4: Number of samples in the OhioT1DM dataset: 2020

Subject ID	540	544	552	567	584	596
Number of Hypoglycemic events	986	188	408	925	137	300
Total time of missing glucose values in hours/days	111/ 4.6	205.7/ 8.6	300/ 12.5	263/ 11	119/ 5	251/ 10.5
Number of Samples	14843	13339	11444	13247	14815	13620

4.1.1. Pre-processing

Health data measured by wearables and sensors usually require additional pre-processing steps before being AI-ready and usable as input data for machine learning models. In particular, the data collected under free-living conditions may contain noise, missing data values, and larger data gaps. Gaps can be caused by different data storage methods, different estimation intervals, noisy data samples, or if the wearable was not worn for the entire study duration. A literature review shows that for filling in missing values, linear methods are most often applied while defining a limit for allowed consecutive missing values. In contrast, larger gaps are completely removed. From tables 3 and 4, it can be seen that over the 8 weeks of measurements, multiple days of missing data per patient are present. The duration of missing data is especially increased for the 2020 cohort. The maximum gap is obtained in subject 552 with more than 12 days, which does not

mean that there are missing measurements for consecutive days but the total sum of gaps accounts for 12 days. Thus, those gaps require data imputation or pre-processing. In this context, Bertachi et al. imputed data with gaps less or equal to 120 minutes, and Jaloli et al. only imputed missing data of 60 minutes [64, 57]. For data imputation, most use linear interpolation for training data and linear extrapolation for testing data to guarantee that future data would not be observed by the model [71]. Nemat et al. further substitute non-reported insulin data with zeros [71]. Finally, the data can be normalized for better generalization, to avoid overfitting, and to enable better training. Here, the input sequence can be scaled for each value to the minimum and maximum value over the entire training set of that variable [71], or all parameters can be scaled to a range between 0 and 1 [57].

Handling Missing Data: Since multiple variables, such as glucose values, given insulin dosages, and acceleration data are utilized, this thesis down-sampled the data to fit the time of glucose leading to more uniform time series sequences. Furthermore, it reduced the amount of missing values. Therefore, first, the time points were rounded to 5 minutes and then resampled to 5-minute intervals. Considering previous work, this thesis decided to use linear interpolation which is a method filling missing data points between two known values of the same parameter so that they are connected by a straight line [72]. The following formula is used for a basic linear interpolation, in which $f(x1)$ and $f(x0)$ represent the known values of the independent variables $x1$ and $x0$, while $f1(x)$ is the missing value:

$$f1(x) = f(x0) + \frac{f(x1) - f(x0)}{x1 - x0} * (x - x0). \tag{7}$$

In addition, linear extrapolation including values that are not in the range of $f(x1)$ and $f(x0)$ was tested, but the outcome was the same as interpolation for the OhioT1DM dataset. On average, most missing values appeared for glucose and then for the acceleration data. Only missing values less than 120 minutes being 24 consecutive instances were interpolated, since there are many missing glucose values. The remaining missing values for glucose were removed so that each patient had a collection of multiple data-frames without any time gaps and without any missing glucose data. Contrariwise, missing acceleration values were replaced with a -1, to indicate missing values, since glucose is the ground truth for defining the classes. Whereas, it is natural that wearable data will not be available for the entire time, as why the system should learn to identify and handle missing data points in wearables.

Pre-processing: Before removing the gaps, the data of single parameters were pre-

processed, and then all data was concatenated. For basal insulin, the continuously applied dosage as well as the temporally applied dosage was reported. Thus, first, the basal insulin was resampled to 5-minute intervals by filling values in between with the previous value. Then, both data-frames were merged. In the next step, all indexes of the basal dosage for which the temporal basal dosage had a corresponding entry were replaced with the entry. Missing data of basal insulin was filled forwards and backward with previously known values since basal insulin is given periodically. For bolus insulin, the starting and the ending time of infusion were reported. For those reported intervals, the given dosage was applied, and the remaining values were filled with a 0 as not reported times indicate that no bolus was infused.

This thesis wanted to include data of physical activity, but as reported in 4.1, the OhioT1DM dataset used two different wristbands for each cohort leading to the estimation of two different parameters for activity data. As a result, both cohorts' data could not be merged that easily. The 2018 cohort collected the step size while the 2020 cohort collected the magnitude of acceleration. Since the magnitude of acceleration can be approximately computed given the step count and time information, the following formulas were used to have uniform data for both cohorts taken [73, 74]:

$$
\begin{aligned}
distance(x) &= step\_count(x) * 0.75 \\
velocity(x) &= distance(x)/time \\
acceleration(x1) &= (velocity(x1) - velocity(x0))/time\_intervall.
\end{aligned}
\tag{8}
$$

Acceleration is the change in velocity divided by the change in time. Since the steps were measured every 5 minutes, the time for each distance and the time difference between velocity changes in 5 minutes [73, 74]. Furthermore, the unit for distance is meters, as why the step had to be multiplied by 0.75 to be converted into meters. This number was chosen after a literature review reporting values from 0.74-0.76 meters for the standard equality of one step and one meter. Lastly, to have uniform data, the acceleration data of both cohorts were scaled to values between 0 and 1.

Annotation of Classes: To define hypoglycemia, some studies suggest event-based classification over sample-based classification, in which consecutive data samples meeting a requirement are defined as an event. In a sample-based classification all samples under one defined threshold are annotated with the class [52]. However, since this work classifies multiple prediction horizons and predicts the risk of a possible hypoglycemic

state, event-based classification would first reduce the number of samples, and second would not enable better accuracy. The literature review has shown that the most often utilized threshold, which is also the standard threshold for defining hypoglycemia, is 70 mg/dL for patients with T1D. Only Georga et al. and Piersanti et al. defined hypoglycemia by a threshold of 60 mg/dL [52, 6]. Thus, each sample less and equal to 70 mg/dL was annotated with class 0 to describe the hypoglycemic state. From there, 5-15 minutes before hypoglycemia was assigned to class 1, more than 15-30 minutes before was classified as class 2, more than 30-60 minutes before as class 3, more than 1-2 hours before as class 4, more than 2-4 hours before as class 5, more than 4-8 hours before as class 6, more than 8-12 hours before as class 7, more than 12-24 hours before as class 8, and lastly more than 24-48 hours before as class 9, as also summarized in table 5. Class 0 is not that sensible to use for prediction but as hypoglycemic events can be asymptomatic, the class was still included to alert the patient. Most importantly, as a hypoglycemic event can occur for a longer duration it had to be ensured that no instance is overwritten with new classes. Therefore, all instances were first assigned a value of -1 as the class. Then, the hypoglycemia threshold was applied over all samples to define the hypoglycemic state. Thereafter, the time condition was only used for instances that were assigned to class -1. Hence, only instances which do not belong to any class already could be assigned a new class. For instance, if samples are assigned to class 0, they are not taken into consideration again for the requirement and thus, cannot be reassigned.

Table 5: Assignment of the classes

Prediction Horizon	Label
0 (hypoglycemic event)	Class 0
5-15 minutes	Class 1
20-30 minutes	Class 2
35-60 minutes	Class 3
65-120 minutes	Class 4
125-240 minutes	Class 5
245-480 minutes	Class 6
485-720 minutes	Class 7
725-1440 minutes	Class 8
1445-2880 minutes	Class 9

4.1.2. Correlation Analysis

The correlation coefficient presents the association between two given variables. Considering all data points, the relation in the change of two parameters is computed, measuring the degree to which both variables would fit on a straight line. It is to note that even if a correlation analysis does not depend on the measurement units, it varies with the range of observations [75] which could lead to a bias and differences in this work since the number of samples varies between subjects. The strength of the correlation is evaluated with the absolute value. Here, the score mostly ranges from -1 and $+1$ in which a value of $+1$ indicates a complete correlation. Thus, for given parameters X and Y, Y would be positively correlated to X most strongly. A score of 0 means that both variables are not dependent and do not have any associated behavior, while negative values indicate inverse correlation. Consequently, while one parameter increases the other parameter tends to decrease. It is reported that in general, a correlation score above 0.60 can be evaluated as strong and above 0.80 as very strong [76]. Table 6 summarizes the interpretation for given correlation score ranges in the biological domain given by Miot which is also reported to be the rule of thump in behavioral science by Mukaka [77, 78]. The interpretation of the correlation coefficient depends on the context and domain as to

Table 6: Rule of thump for correlation interpretation [77]

Correlation Range	Interpretation
0 to 0.3 (0 to -0.3)	negligible
0.31 to 0.5 (-0.31 to -0.5)	weak
0.51 to 0.7 (-0.51 and -0.7)	moderate
0.71 to 0.9 (-0.71 to 0.9)	strong correlations
$>$ 0.9 ($<$ -0.9)	very strong

why even if guidelines define a score as weak, it could be still of significance, especially in the medical domain [75]. Medical events can be multi-factorial and in the diabetes domain, a hypoglycemic event can be impacted by multiple parameters.
One of the most often utilized methods for quantitative correlation analysis is the Pearson correlation assuming a normal distribution and a linear relationship between the parameters. The formula for Pearson correlation is presented in the following, in which $COV(X, Y)$ is the covariance of X and Y, σ_X and σ_Y are the deviations of X and Y,

and μx and μy are the respective means [76]:

$$\rho_{X,Y} = \frac{COV(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X - \mu x)(Y - \mu y)]}{\sigma_X \sigma_Y} \tag{9}$$

Contrariwise, the Spearman correlation is a non-parametric method and an extended version of Pearson. It is based on ranks and not the actual value of the observations making it robust to outliers [75, 79]. If both variables are ranked, the following formula can be used presented in [79], in which d^i is the difference between each pair of the ranked variables and N represents the total number of samples:

$$r_s = 1 - \frac{6 \sum d_i^2}{N(N^2 - 1)} \tag{10}$$

$$d_i = X_i^r - {}_i^r$$

For this thesis, the Pearson correlation between glucose, basal insulin dosage, bolus insulin dosage, and acceleration data was computed. Furthermore, as a comparison, the Spearman correlation was analyzed since a linear relationship cannot be directly assumed. The correlation analysis aims to investigate the relation of parameters during the chosen classes, which could give insights into the distinctions or similarities between classes.

4.2. Model Architectures

This work developed a population-based deep-learning model to solve a multi-class classification problem and to classify the time to the onset of hypoglycemia. In the first step, a test subject was randomly chosen to decide on the main model architectures. These models were then tuned on the test subject. Data from the remaining 11 subjects was split into training and validation data and was shuffled. The validation data consisted of the last 20% from each training person's data. Then, the trained models were tested on the whole data of the selected test subject 552. In total four models were approached. The best LSTM architecture was compared to a BiLSTM model, while the best ResNet architecture was compared to a 1DCNN model.

For all models, early stopping, the saving and restoring of best weights, and the decrease of the learning rate were applied as callback functions observing the validation loss. In addition, the patience of early stopping was set to 5 for the RNN models due to the longer computation time, while it was set to 10 for the CNN models. The patience of the learning rate reduction was set to 3 for all models.

As a starting point, a simple LSTM and ResNet model were developed which were then further tuned. Both architectures were tested with different batch-sizes and various input time sequence lengths of 12 hours which are 144 data-points, 8 hours which are 96 data-points, 6 hours which are 72 data-points, and 4 hours which are 48 data-points. Moreover, each model used the Adam optimizer and the sparse-categorical-cross-entropy loss function. Additionally, weights were applied for each class due to the high imbalance of classes. The weights were computed with the following function, in which "class_occurrence" counts the number of samples for the considered class x, "total" counts the number of all available samples among all classes, while "number_of_classes" refers to the set of classes utilized:

$$weight_x = (1/class\_occurrence) * (total/number\_of\_classes) \tag{11}$$

The functions for calculating the weights are provided by TensorFlow in [80]. Moreover, to ensure reproducible results, a seed value of 42 was chosen.

The model was first trained with all 10 proposed classes, however, it had poor performance and many misclassified instances. Therefore, the last class was removed, and all subsequent experiments used only nine classes classifying up to 24 hours before the hypoglycemic event. A comparison of confusion matrices using 10 and 9 classes for the best RNN models and CNN models can be seen in figure 10 and in figure 12, respectively showing

a slight improvement when using 9 classes.
Finally, in the final population-based approach, which is explained in subsection 4.2, the superior models were compared. Additionally, a hybrid model consisting of the better CNN and better RNN model was tested.

LSTM vs BiLSTM Model

Three LSTM models were tested to select the basic architecture that was tuned for the test data. The first model consisted of one layer with 128 units, and the second model consisted of two layers of which the first had 128 units and the second had 64 units. The third model consisted of three stacked layers with 128 units followed by 64 units and 32 units. As a result, the performance of the third was best. Additionally, a model with three stacked layers with units of 256, 128, and 64 was tested which was superior but had a longer computation time. All models utilized the tanh activation function because other functions, such as rectified linear unit (ReLu), exponential linear unit (ELU), or swish could not produce any results. Thereafter, the model was tested with dropout layers, removing 20% of data-points after the first and after the second layers, which resulted in a worse classification performance. In addition, a global average 1D pooling layer was applied after the final LSTM layer, resulting in a worse performance. Among the tested batch-sizes of 32, 64, and 128, batch-size 128 and 64 performed similarly. However, batch-size 64 showed better overall performance. The model was also tested with a dense layer of 100 units following the last LSTM layer, which led to an increase in the metric values. For all reported experiments, an input sequence length of 12 hours was utilized. Thereafter, different sequence lengths were explored, and 8 hours and 12 hours yielded similar results, as shown in table 7. However, the short computation time of a smaller sequence length was considered superior. After deciding on the best LSTM model, a BiLSTM model with 128 units followed by an LSTM layer with 64 units was compared with input sequences of 12 and 8 hours. The architecture of compared RNN models can be seen in figure 9. The performances, when utilizing 8 hours of prior measurements, were similar. Whereas, the accuracy was increased with the BiLSTM model as can be seen in table 7. Nevertheless, LSTM was seen as superior due to the shorter computation time and better recall for the first classes, and class 5. Finally, the confusion matrices of the best LSTM and BiLSTM models are presented in figure 10, revealing that the model is not capable of classifying the latter classes well.

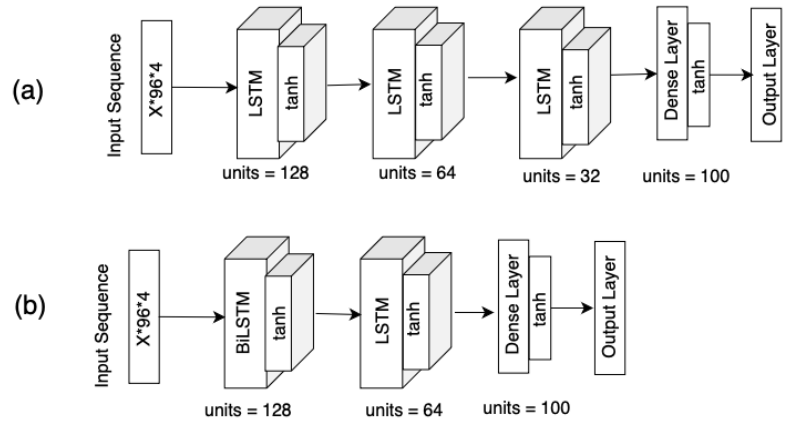


Figure 9: Architecture of applied RNN models
a) LSTM, b) BiLSTM
Abbreviation: tanh = hyperbolic tangent function (Activation function)

Table 7: Comparison of LSTM and BiLSTM models trained with a batch-size of 64 across 9 classes

Model	**Input-Length**	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-Measure (%)**
LSTM	12 hours	30	33	42	35
LSTM	8 hours	30	33	42	35
BiLSTM	12 hours	31	32	41	33
BiLSTM	8 hours	32	33	42	35

1DCNN vs ResNet Model

Different kernel sizes and layer configurations were experimentally tested for the 1DCNN and the ResNet models. In short, it was observed that 1DCNN models train and learn faster, especially with greater kernel-sizes. However, the validation and test accuracy are poor. Therefore, a ResNet model was used as a basic model. As a starting point, different block sizes and kernel sizes were tested, and a model with 5 residual blocks of which each block consisted of 2 1DCNN layers with the same kernel-size was selected. The kernel-sizes for blocks 1, 2, 3, 4, and 5 were 7, 5, 3, 3, and 3, respectively. Before the first block, a 1DCNN layer with a kernel-size of 9 was used without a shortcut connection. The filter-sizes of each 1DCNN layer were 64. In addition, each 1DCNN layer followed a batch-normalization and the ReLu activation function. The basic model was tested with

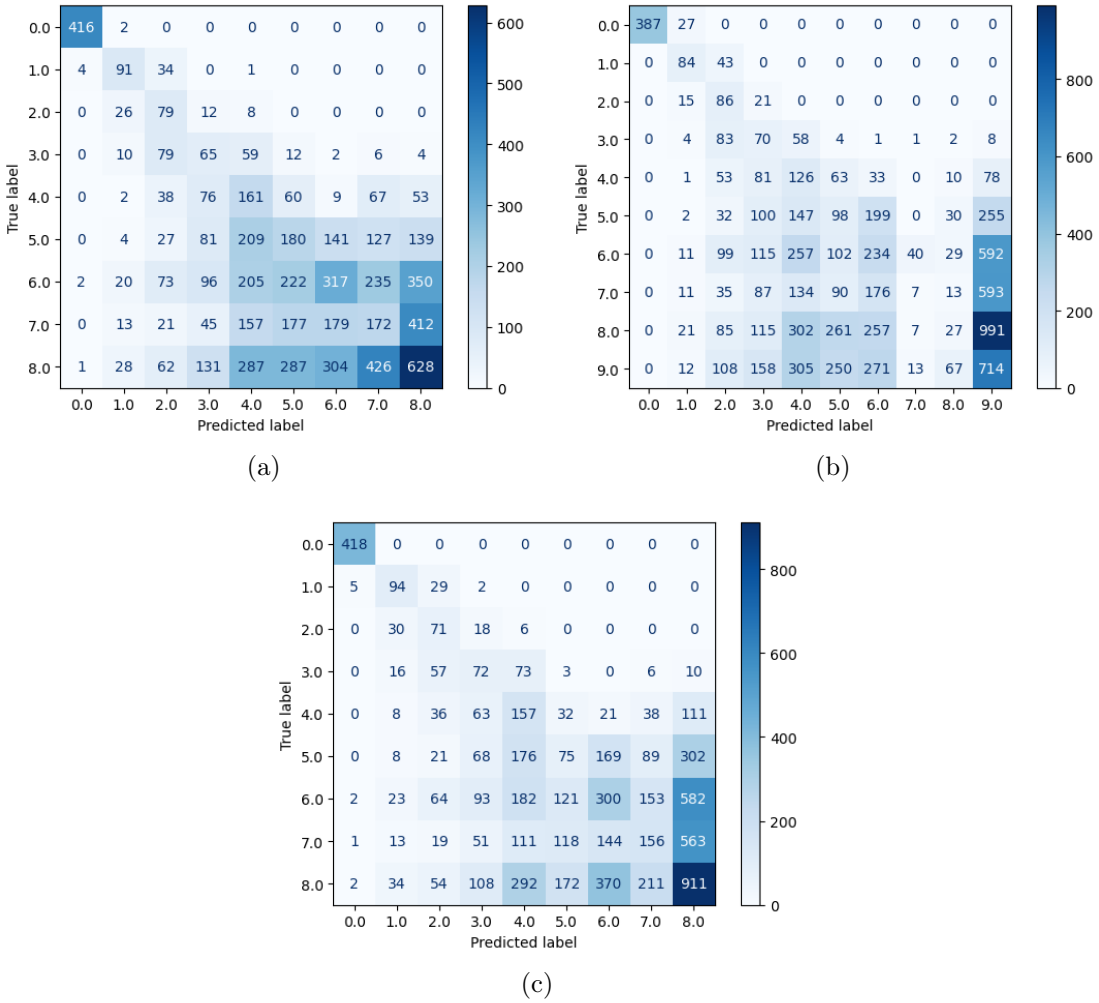


Figure 10: Confusion-matrices of RNN models
(a) LSTM model trained with an input length of 8 hours and 9 classes (b) LSTM model trained with an input length of 12 hours and 10 classes (c) BiLSTM model trained with an input length of 8 hours and 9 classes

a 1D max-pool layer of pool-size 2 after the first convolutional layer before the residual blocks, which resulted in worse metrics. Next, the model was tested with and without a dense layer with 100 units following the global 1D average pooling, which resulted in improved performance. Batch-sizes 32, 64, 128, and 256 were also tested. Here batch-size 128 outperformed the others. In contrast to the LSTM model, dropout layers after each residual block except the last block increased the classification performance. Finally, the best model was tested with various activation functions including ELU, swish, and tanh but none outperformed the ReLu activation function. Thereafter, three blocks were tested with kernel-sizes of 7, 5, and 3, respectively. The model with three blocks was

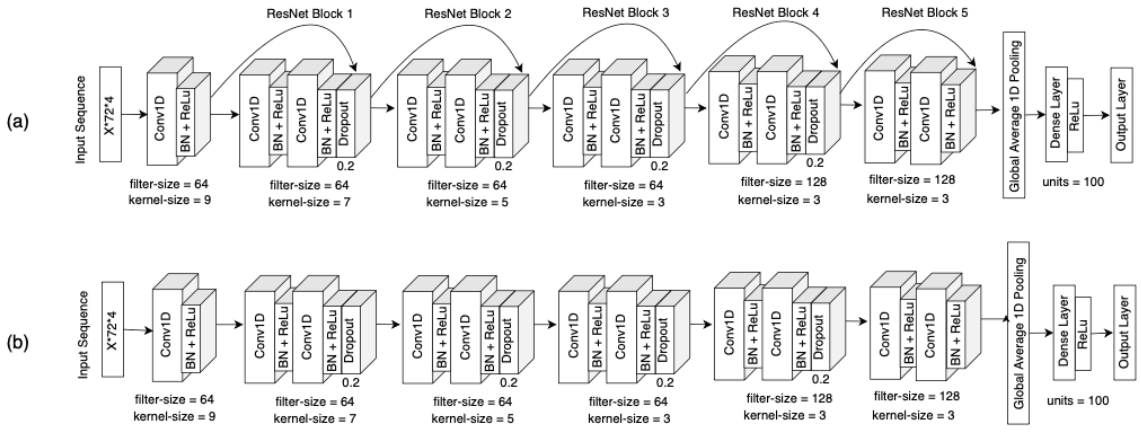


Figure 11: Architecture of applied CNN models
a) ResNet, b) 1DCNN
Abbreviations: BN = Batch-normalization, ReLU = Rectified Linear Units (Activation function)

tested with a filter-size of 128 for the last residual block, and the model with five blocks was tested with a filter-size of 128 for the 4th and the 5th residual blocks, respectively. Furthermore, each model was tested with three stacked convolutional layers for each block. However, the results were worse. A model with filter-sizes of 64, 64, 64, 128, and 128 and kernel sizes of 7, 5, 3, 3, and 3 for blocks 1, 2, 3, 4, and 5, respectively performed best. Lastly, different input sequence lengths were tested, and a length of 6 hours achieved the best outcome, while 12 hours was similar, and 9 hours had the worst results. The metrics for 12 and 6 hours of prior measurements as the input data are summarized in table 8. The advantage of using only 6 hours of prior measurements is that all data-frames can be used, while 12 hours lead to lost samples.
Thereafter, the same architecture without the shortcut connections was tested as a simple 1DCNN model of which the results can be seen in table 8 as well. The architecture of compared CNN models is visualized in figure 11. It can be seen that the ResNet model has a slightly better classification performance, thus it was chosen as superior. Additionally, the ResNet model depicted in [38] and visualized in figure 6 (b) was tried but resulted in a longer computation time (70-90 minutes). The confusion matrices of the best ResNet and the best 1DCNN model are presented in figure 12, which show a similar pattern to the experiments with RNN models. It is further noticed that the metrics of the best LSTM and the best ResNet model do not vary much.

Table 8: Comparison of ResNet and 1DCNN models trained with a batch-size of 128 across 9 classes

Model	**Input-Length**	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-Measure (%)**
ResNet	12 hours	29	32	36	33
ResNet	6 hours	28	33	42	34
1DCNN	12 hours	27	27	33	28
1DCNN	6 hours	28	30	37	31

Hybrid Model

The hybrid model depicted in figure 13, is a stacked model consisting of the ResNet model built by one 1DCNN layer followed by 5 residual blocks, a 1D average pooling layer, and 3 LSTM layers, followed by a global 1D average pooling and a dense layer. This model architecture was not tested prior on subject 552.

Experiments

For the final implementation, the three chosen models were tested on the whole dataset on all subjects with a LOOCV. Here, as visualized in figure 14, each subject gets the test person, while the model is tested with the remaining 11 persons. In total, 12 folds were built since the dataset consists of 12 subjects. Each test fold held the data of a different test person for whom the data was not present in the training fold. Furthermore, for the validation fold, the last 20% of each subject's data is removed from the training fold to include all train persons' data in the validation. The ResNet and Hybrid models were both trained with an input sequence length of 6 hours, and a batch-size of 128, while the LSTM model was trained with an input sequence length of 8 hours, and a batch-size of 64. For all models, a patience of 10 was applied for the early stopping. Then, as a comparison, transfer learning was utilized in which the model was further trained with 50% of the test subject's data. 20% of the data was used for validation and 30% for testing. These "individualized" models, utilized a batch-size of 16. Furthermore, the patience for early stopping was increased to 20 epochs for the subject-specific approach. Transfer learning is a method applied often for small datasets, in which an already trained model with similar data is reused for a new dataset. It can either be used for further training or for validating a model with unseen data. The proposed method is also applied by Deng et

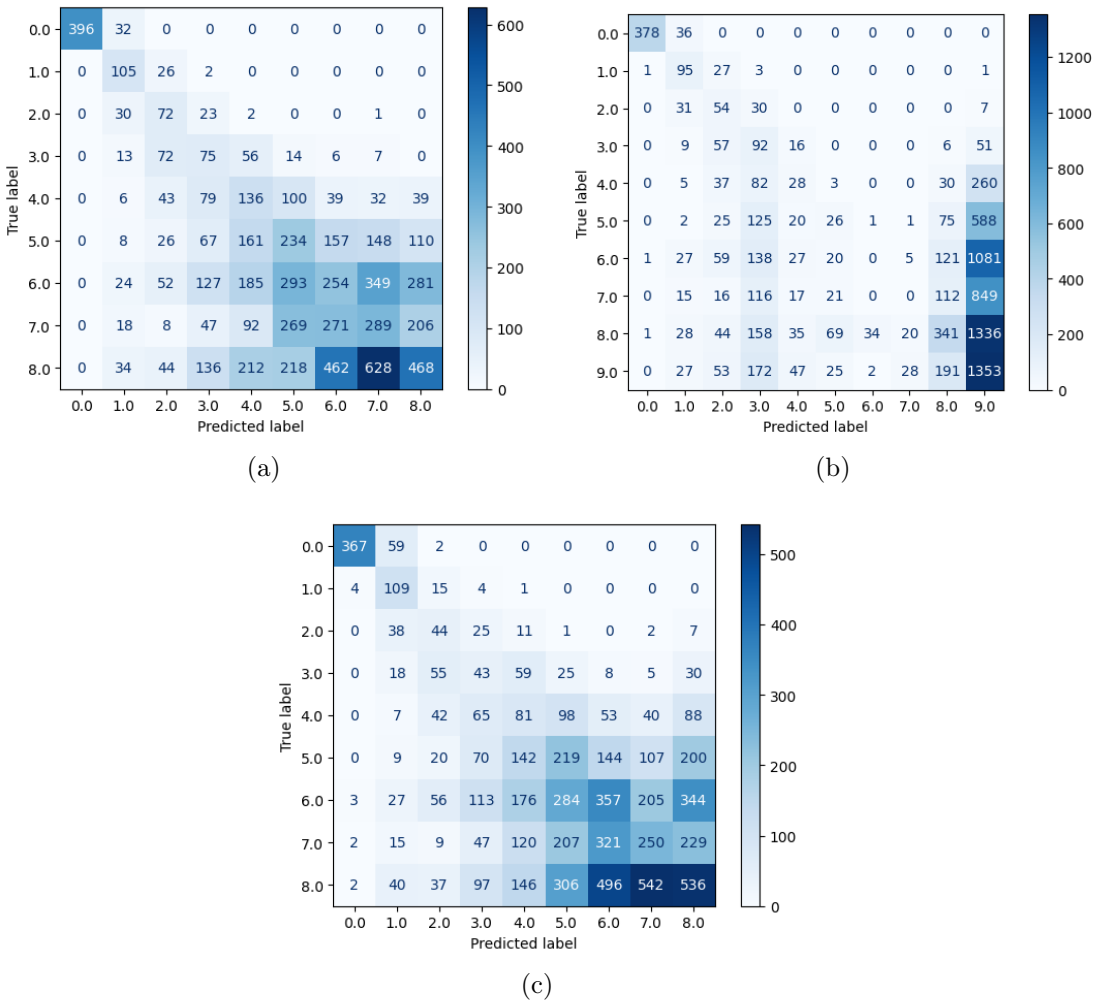


Figure 12: Confusion-matrices of CNN models
(a) ResNet model trained with an input length of 6 hours and 9 classes (b) ResNet model trained with an input length of 12 hours and 10 classes (c) 1DCNN model trained with an input length of 6 hours and 9 classes

al. who train a model with the training dataset and data of the test subject is used for further training [81]. Transfer learning was chosen since not every subject had enough data to train individual models. Moreover, the population-based models were tested with the same reduced test data to enable a fair comparison. Lastly, since the best LSTM and ResNet models did not reveal a very promising performance, the population-based and subject-specific models were also trained by utilizing only 6 classes classifying up to 4 hours before hypoglycemia. The batch-size for all models was half of the prior used batch-size, because the removal of classes resulted in fewer samples. Furthermore, the input sequence length of all models was set to 4 hours being 48 data-points.

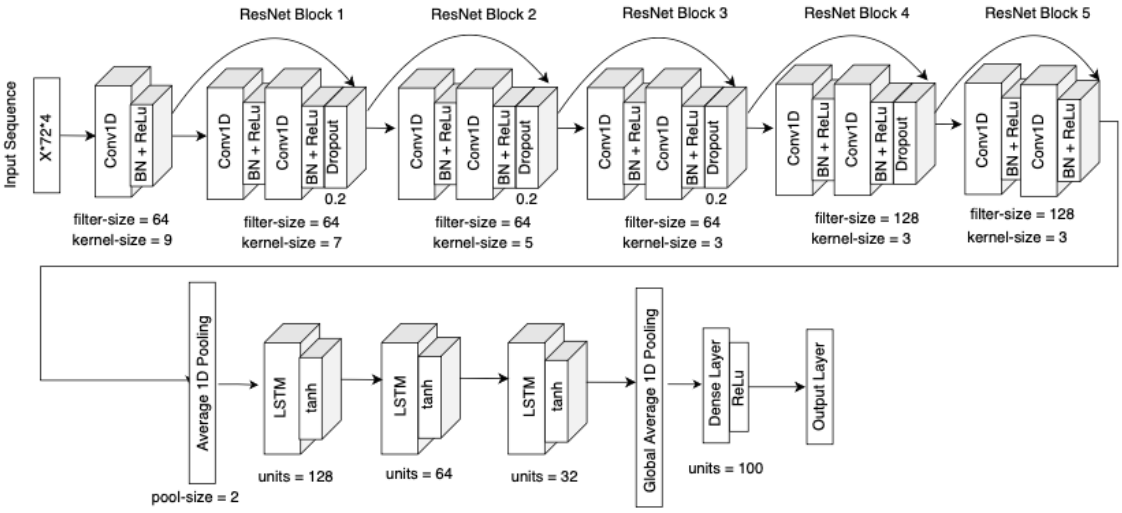


Figure 13: Architecture of applied hybrid model: ResNet + LSTM
Abbreviations: BN = Batch-normalization, ReLU = Rectified Linear Units (Activation function), tanh = hyperbolic tangent function (Activation function)

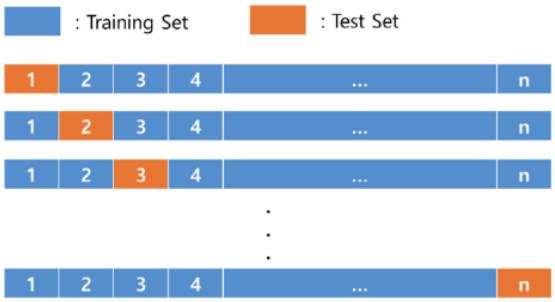


Figure 14: Architecture of LOOCV [82]

4.3. Metrics used for the Model Evaluation

This thesis is based on a multi-class classification task. Therefore, the developed models need to be validated with metrics such as accuracy, precision, recall, and F1-measure. All utilized metrics are in the range of 0 and 1. Furthermore, the confusion matrices for some models are visualized which is a cross table showing the true and predicted classes of each sample as already seen in the previous subsection. The true positives (TP), the false positives (FP), the false negatives (FN), and the true negatives (TN) can be extracted from the confusion matrix and give more insight into the classification performance of the single classes. In this case, TP are values that are correctly identified

as positive by the model, whereas FP are predicted to be positive but belong to another class. FN are classified as negative but belong to the considered class. Lastly, TN is correctly classified as negative. The equations for the proposed metrics can be seen in the following, utilizing the described components [83]:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{12}$$

$$Precision = \frac{TP}{TP + FP} \tag{13}$$

$$Recall = \frac{TP}{TP + FN} \tag{14}$$

$$F1 - Measure = \frac{2 \times Precision}{Precision + Recall} \tag{15}$$

From the given formulas, it can be seen that the precision computes the portion of truly positive samples from all as positive classified instances. Whereas, recall gives the portion of identified samples of one class [83]. To illustrate the difference, precision tells how many alarms for a hypoglycemic event are indeed true alarms, while recall considers how many of the hypoglycemic events are indeed recognized, and thus could be prevented. Furthermore, the F1-Score is the harmonic mean between precision and recall. Accuracy computes the portion of correctly classified data-samples considering the whole dataset. Hence, it calculates the average probability of a correct classification for a random sample. One major disadvantage of the accuracy metric is that it is biased with imbalanced datasets, especially in multi-class classification problems [83]. To illustrate the drawback, if half of the entire dataset belongs to one specific class and those are all classified correctly, the accuracy would be around 50%. However, if other classes have poor performance, it cannot be concluded that a random sample is classified correctly with an accuracy of 50%. Finally, this work will report the macro averages for each metric, due to the high imbalance in data. The macro average considers all classes equally without having a bias of popular and less popular classes [83].

The literature review has shown that often relevant events are detected but most systems have a low precision inducing many false alarms. Bertachi et al. state that too many false alarms could cause the patient to consume unnecessary carbohydrates, especially at night time, which could lead to hyperglycemia and weight gain [64]. Hence, the false alarm rate must be in acceptable ranges. In relation, recall is more important compared to precision in clinical settings which is also reported by Zhu et al. [59]. Consequently, the F1-Measure is of high relevance.

4.4. Expected Limitations

From the previous subsections, it could be seen that the dataset itself has some drawbacks and that the reported models do not perform that well. There are many gaps in the data which could cause the loss of important information and the dataset is not that large. Furthermore, the distribution of classes is imbalanced which could lead to a worse classification of underrepresented classes. Besides, the experienced hypoglycemic events differ between the patients as why the model could be biased and is limited in learning the data of the population equally. The data is split into training and validation, but because the last 20% is assigned as validation data and since the classes are highly imbalanced, it is not known if the validation data represents all classes equally well. Lastly, as also depicted in chapter 2, deep learning could cause overfitting with small datasets which was also observed in the reported models in subsection 4.2.

5. Results

This chapter first explores the pre-processed dataset and visualizes the parameters of selected hypoglycemic events for each patient in subsection 5.1. In the second part, the population- and individual-based correlation results are analyzed in subsection 5.2. Lastly, subsection 5.3 reports the results of applied deep learning models.

5.1. Preliminary Data Analysis

Time series sequences were created for the train and test datasets to train and evaluate the deep learning models. Here, a sequence length was defined and the last value of the series decided the class of the sequence. The literature review showed that the input series should be as long as the maximal prediction horizon. Nevertheless, the experiments in chapter 4.2 revealed that sequence lengths of 6 and 8 hours yield better performances when classified up to 24 hours before hypoglycemia. Longer sequences either have inefficient computation time or result in lost data samples due to unusable data-frames. The distribution of samples per class before pre-processing can be seen in table 9. Contrariwise, table 10 presents the classes' distribution after pre-processing, gaps removal, and after the creation of time series sequences with a sequence length of 8 hours. The comparison illustrates the number of lost instances since the removal of gaps leads to unusable data-frames and samples because not every data-frame meets the requirement of the defined sequence length. Thus, most subjects have reduced samples using an input sequence of 8 hours. Most samples are lost in the latter classes. Whereas, linear interpolation also increases the number of instances of some subjects in class 0. Using an input sequence of 12 hours or 24 hours resulted in even fewer samples.

Data Exploration: Figures 15, 16, 17, and 18 visualize the collected pre-processed parameters 48 hours before a hypoglycemic event. Here, only one hypoglycemic data point was selected for each patient. It can be seen that the applied insulin dosage is highly associated with the decrease of glucose and the hypoglycemic event in most of the cases presented.

For subject 540, it can be observed that three hypoglycemic events were experienced within 48 hours, one of the most recent events was severe hypoglycemia. Bolus insulin was administered prior to the depicted hypoglycemic event while exercise data did not seem to impact glucose values.

For subject 544, increased insulin dosages were injected and the maximum glucose values were higher in comparison. The chosen data point is the first hypoglycemic event within

Table 9: Samples per class with raw data

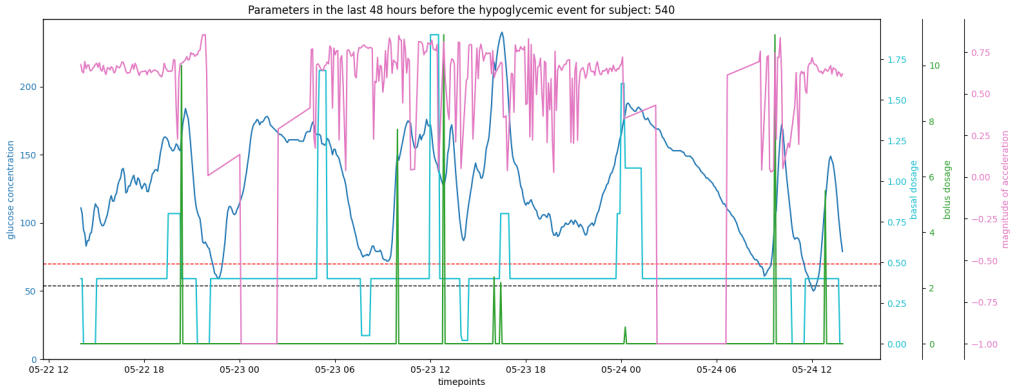
ID	Total	Class									
		0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
540	15229	986	314	298	566	1025	1707	2779	1952	3400	2202
544	8049	188	87	75	122	216	432	864	864	2409	2792
552	11377	408	142	137	261	514	1007	1769	1354	2814	2971
559	12725	518	184	167	318	612	1176	2156	1727	3281	2586
563	11099	329	130	122	223	427	709	1336	1206	2805	3812
567	14406	925	140	125	238	453	878	1722	1495	3638	4792
570	7543	227	87	77	137	250	420	768	760	1987	2830
575	14812	1173	291	263	485	930	1655	2734	1926	3385	1970
584	5384	137	72	64	126	236	406	593	528	1439	1783
588	6372	136	77	68	132	241	456	805	693	1621	2143
591	14357	570	233	213	401	767	1326	2258	1952	3798	2803
596	13245	300	167	143	260	516	990	1774	1602	3501	3992

Table 10: Samples per class for pre-processed time series data of 8 hours

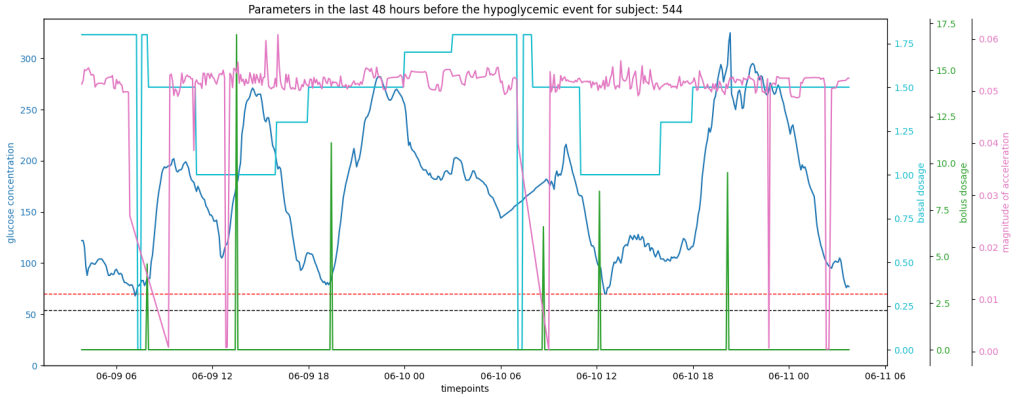
ID	Total	Class									
		0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
540	13722	894	291	277	523	933	1521	2418	1640	3086	2139
544	7366	184	78	69	110	192	398	842	861	2073	2559
552	9168	418	130	125	237	466	908	1520	1176	2154	2034
559	11450	441	152	138	268	522	1037	1950	1589	3066	2287
563	10613	327	124	118	217	415	685	1283	1114	2577	3753
567	11747	932	137	120	226	429	830	1495	1242	2698	3638
570	7050	227	87	77	135	238	396	720	686	1804	2680
575	13738	1158	261	236	437	834	1488	2538	1777	3140	1869
584	4921	130	70	64	126	236	404	589	528	1370	1404
588	6078	138	77	68	132	238	432	757	645	1448	2143
591	12919	600	214	197	382	731	1263	2110	1807	3419	2196
596	10384	259	136	122	218	425	788	1412	1273	2850	2901

the presented 48 hours. Furthermore, increased physical activity is identified at some points but because bolus insulin is also increased at the same time, the direct dependence cannot be clarified.
Subject 552 was in a hypoglycemic state one day before the chosen event and possibly had increased physical activity for a longer duration one day before, which could have an impact on the decrease in glucose values. Moreover, basal insulin was constantly administered with the same dosage.
Subject 559 experienced five hypoglycemic events one day before and had high variations in glucose. Most events seem to result from the bolus and basal insulin dosages.
For subject 563, in general, moderate activity data and a constant basal insulin dosage of 0.70 can be seen. Within the last 48 hours, four hypoglycemic events were experienced of which the first one was severe hypoglycemia, and the last hypoglycemia occurred right before the presented event.
For subject 567 a trend for glucose data is noticed which immediately decreases with the applied bolus insulin. No hypoglycemic events were experienced before.
Likewise, subject 570 had not experienced any hypoglycemia. Exercise data as well as the bolus insulin dosage seem to be increased, while basal insulin was administered in constant dosages in the last 48 hours. Besides, during the night and in the morning the subject had increased glucose values between 200-300 mg/dL.
Subject 575 had one severe hypoglycemia two days before the chosen event. Before the illustrated hypoglycemic value which is experienced during the night, bolus insulin was administered, although the glucose values were not that increased and had already shown a pattern of decrease. Additionally, activity data seems to be slightly increased.
Moving to subject 584, it can be observed that one hypoglycemia was experienced one day before the presented event. Glucose values seem to be increased before the first hypoglycemic event with values up to 400 mg/dL. Additionally, multiple bolus insulin dosages were infused within the day while also basal insulin was applied. During the decrease of glucose, physical activity seems to be increased as well. Then, before the presented event, bolus insulin was applied twice for glucose values of approximately 200 mg/dL. As a result, glucose stayed moderate between 175 and 225 mg/dL for the next few hours until it suddenly decreased at midnight.
Subject 588 had great fluctuations in glucose data and compared to the other subjects, the infused bolus insulin dosages were not that increased while basal insulin was more adjusted and not constant. Increased exercise can be seen at some time points but no hypoglycemic events are experienced in the last 48 hours and in general the glucose

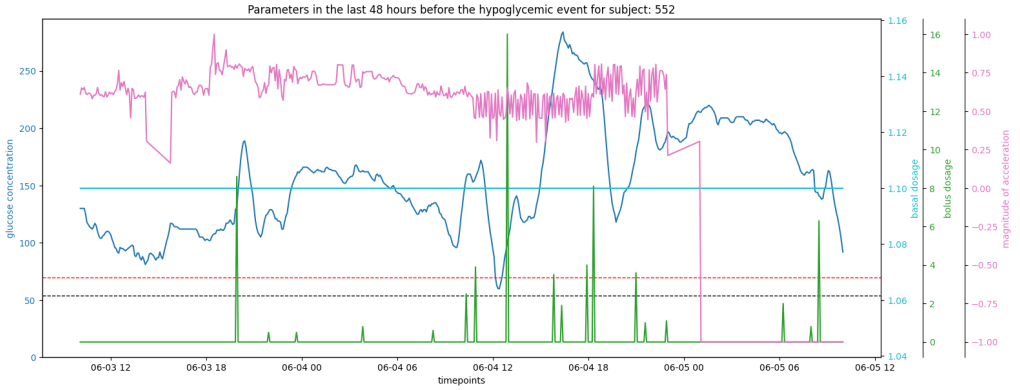
values are not that elevated. Therefore, the direct cause of the chosen hypoglycemic event cannot be recognized.
For subject 591, five hypoglycemic events can be seen in the last two days of which three were severe events including the most recent state. Additionally, increased activity is noticed one day before the chosen event. The subject had increased glucose values between 200-250 mg/dL during the night which decreased with the infused bolus insulin before sleeping at around 6 a.m.. Another dosage of bolus insulin was administered after some hours which happened to be right before the chosen event. Moreover, basal insulin was infused in small dosages, although the patient had glucose values under 100 mg/dL. Lastly, subject 596 has more often normal glucose values between 100-150 mg/dL. In general, decreased insulin dosages were infused compared to the other subjects, while basal insulin was given in constant dosages. Two hypoglycemic events were experienced in the last two days and the last event was recent to the chosen hypoglycemic state. Before the hypoglycemic states, the glucose levels were increased up to 300 mg/dL. Furthermore, both hypoglycemic events seem to be associated with the bolus dosage.
From the presented visualizations of selected hypoglycemic data points, it can be summarized that some of the events could be prevented with prediction algorithms, especially in subjects 570, 575, 584, 591, and 596. In those patients, it is asserted that the event was the result of short-term actions right before the hypoglycemia. Moreover, the number of prior experienced hypoglycemic states also seems to be of relevance looking at all subjects. Physical activity could impact the behavior of glucose but a direct relation could not be highlighted since most often insulin was also increased at the same time.



(a)



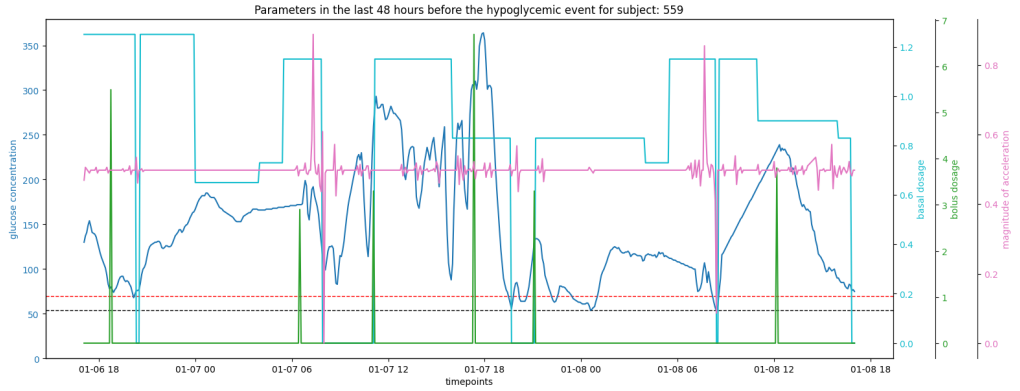
(b)



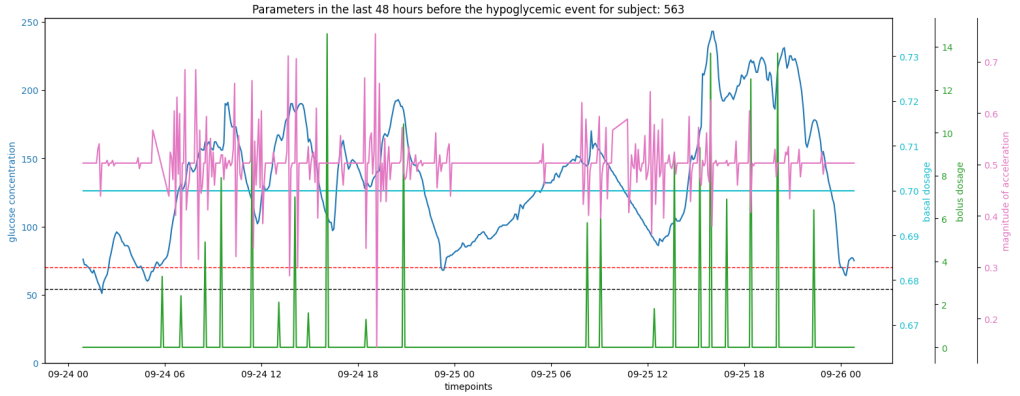
(c)

(d)

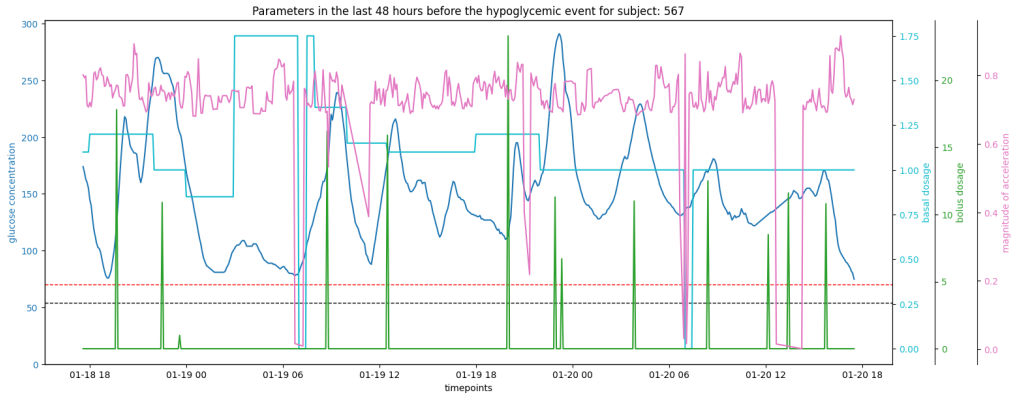
Figure 15: Parameters in the last 48 hours before hypoglycemia (1) (a) Subject 540 (b) Subject 544 (c) Subject 552 (d) Legend



(a)



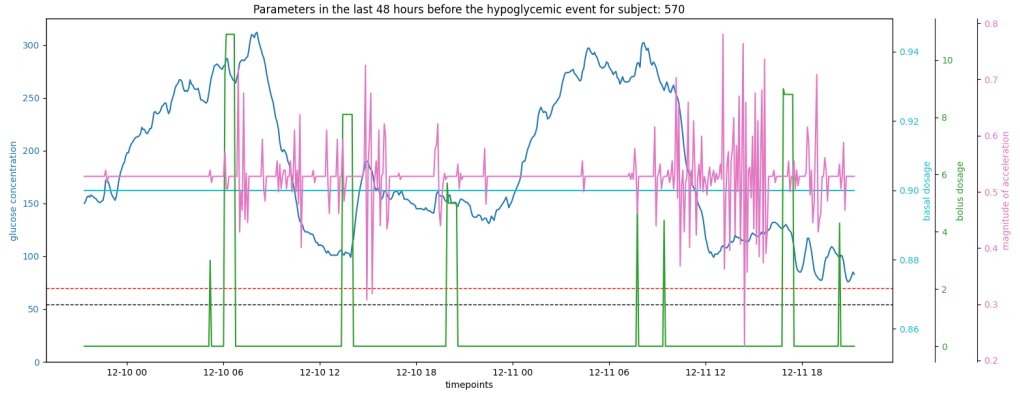
(b)



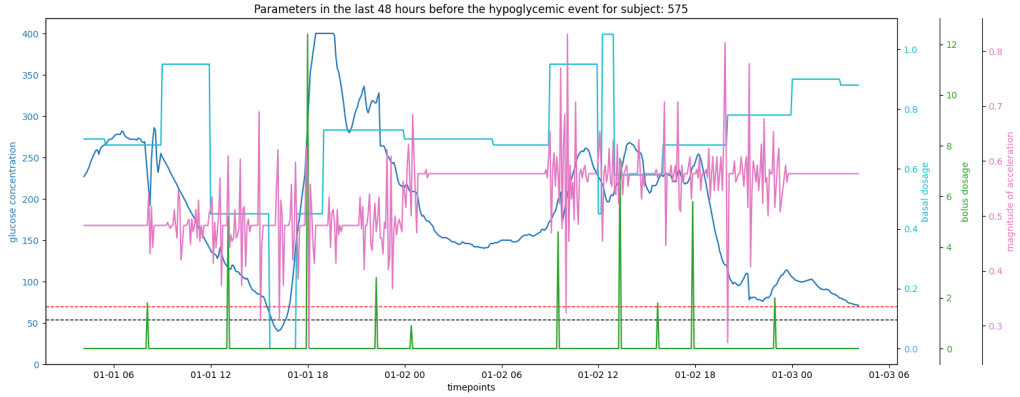
(c)

(d)

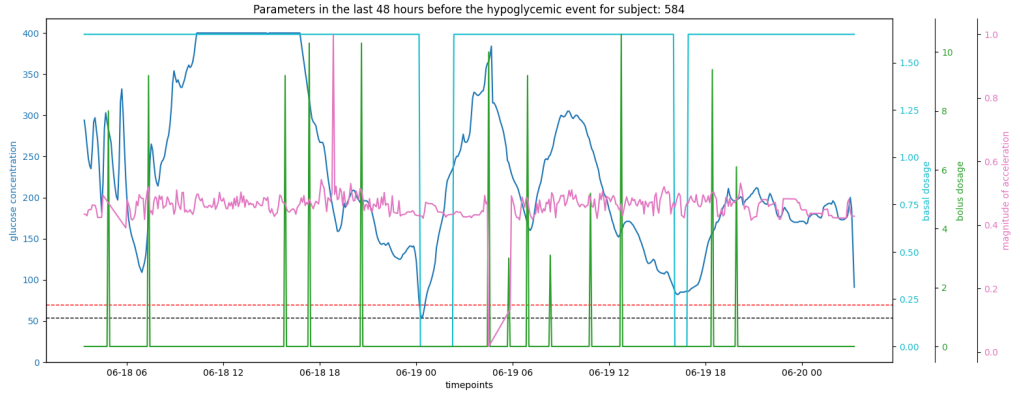
Figure 16: Parameters in the last 48 hours before hypoglycemia (2) (a) Subject 559 (b) Subject 563 (c) Subject 567 (d) Legend



(a)



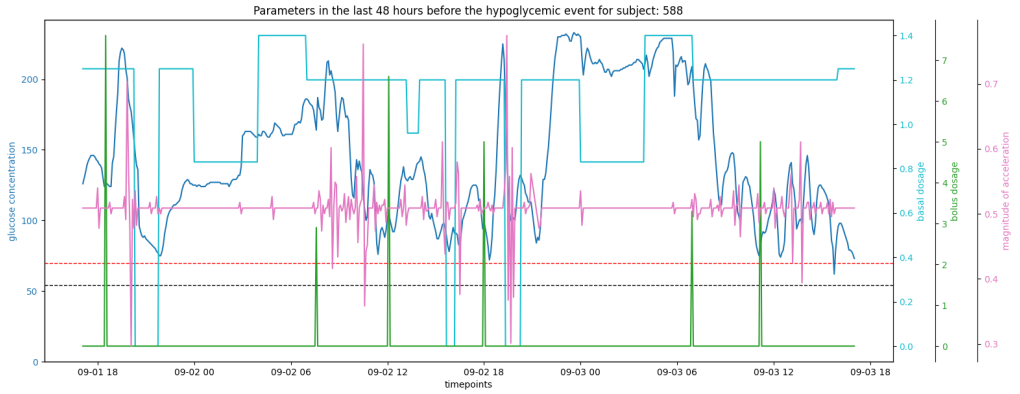
(b)



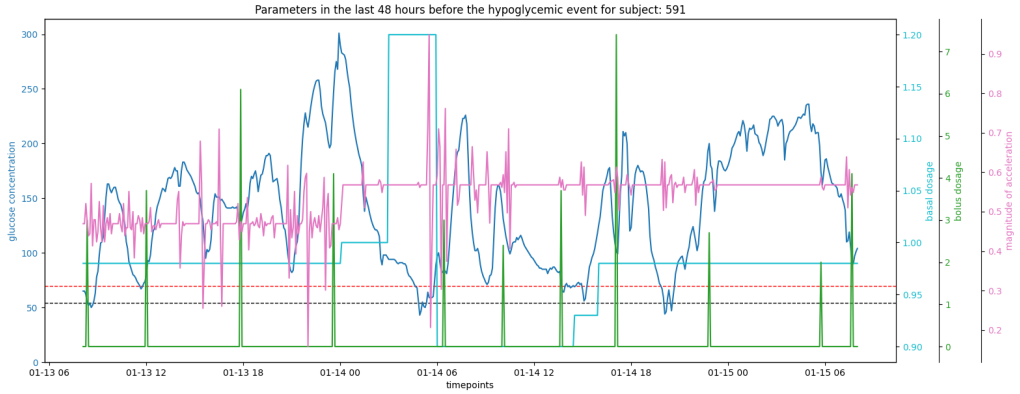
(c)

(d)

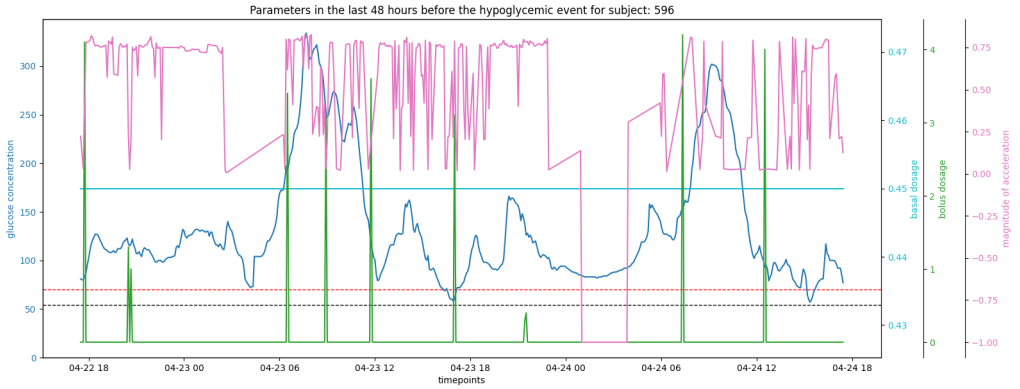
Figure 17: Parameters in the last 48 hours before hypoglycemia (3) (a) Subject 570 (b) Subject 575 (c) Subject 584 (d) Legend



(a)



(b)



(c)

(d)

Figure 18: Parameters in the last 48 hours before hypoglycemia (4) (a) Subject 588 (b) Subject 591 (c) Subject 596 (d) Legend

5.2. Results of the Correlation Analysis

Subsection 5.2.1, describes the results of the Pearson and Spearman correlation which are applied to the whole population. Furthermore, population-based pairwise plots of given variables are illustrated. Lastly, subsection 5.2.2 observes the individual-based results of the Pearson correlation method.

5.2.1. Population-based Correlation

To estimate a population-based correlation, the data-frames of all patients were concatenated and then grouped by classes. The computed Pearson and Spearman coefficients per class are presented in table 11. First looking at the Pearson coefficient, it can be seen that the correlation score between glucose and basal insulin is negligible and there is no significant relation among the classes. The latter classes achieve the maximum scores with 0.108, 0.149, and 0.160, showing an increase from class 9 to class 7, respectively. However, even 0.160 only represents a very weak to negligible relation. Likewise, the relation between glucose and bolus insulin is not significant with a maximum score of 0.228 in class 8. Therefore, glucose and bolus insulin seem to have a very weak association 12 to 24 hours before the hypoglycemic event. Glucose and the Magnitude of Acceleration (macc) indicate a negligible and irrelevant correlation among all classes with a maximum value of -0.075 in class 9. The strongest coefficient in the Pearson correlation can be seen between basal insulin and macc, but it does not follow a specific trend among the classes. Class 9 and 8 show a negative very weak dependence of -0.223 and -0.241, respectively. The other classes can be interpreted as non-relevant, while the correlation score shows a negative very weak association in class 5 increasing to the maximum score of -0.298 and -0.282 in classes 2 and 1, respectively. To summarize the relation between basal insulin and macc, all coefficients are negative and the average correlation among the classes is -0.215, indicating the possibility of a very weak dependence, especially 15 to 60 minutes before the hypoglycemic event. The Pearson coefficients between bolus insulin and macc, and between bolus and basal insulin do not reveal any relevant dependence.
Comparing the Pearson coefficients with the Spearman coefficients, it is noticed that there is not a significant difference between most of the parameters. However, the dependence between glucose and bolus insulin, and between glucose and macc have some variations among the classes. While the maximum score is 0.228 in class 8, which is 12-24 hours before hypoglycemia, using the Pearson correlation, it is 0.256 in class 6, which is 4-8 hours before hypoglycemia, using the Spearman correlation, which is a difference of at

Table 11: Population-based Pearson and Spearman correlation analysis

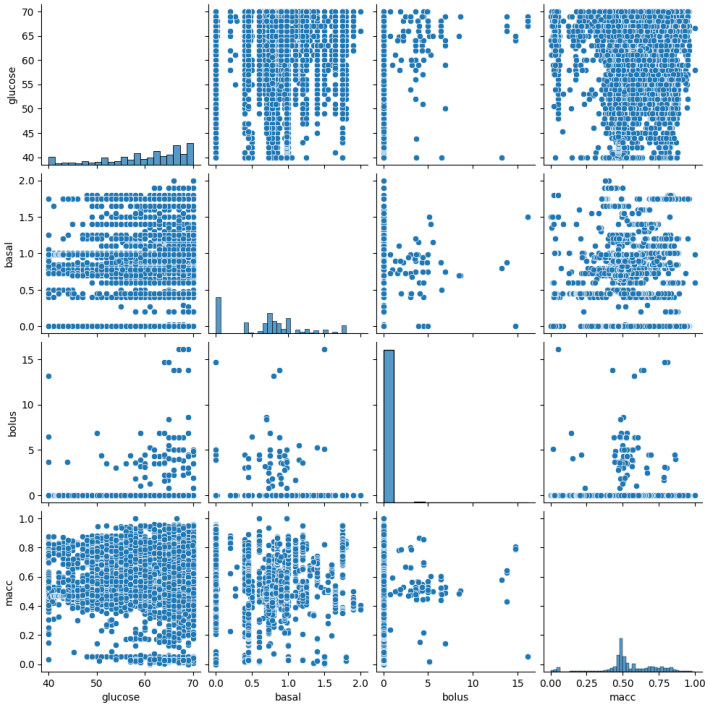
	Correlations	**Classes**									
		0	1	2	3	4	5	6	7	8	9
Pearson	glucose/ basal	0.059	0.088	0.090	0.034	0.040	0.060	0.092	**0.160**	0.149	0.108
	glucose/ bolus	0.047	-0.030	0.003	0.021	0.010	0.022	0.030	0.052	**0.228**	0.042
	glucose/ macc	-0.063	-0.033	-0.008	-0.034	0.020	0.016	-0.030	-0.033	-0.055	**-0.075**
	basal/ macc	-0.083	-0.282	**-0.296**	-0.260	-0.247	-0.222	-0.114	-0.186	-0.241	-0.223
	bolus/ macc	-0.035	**-0.042**	0.004	-0.013	0.00	**0.042**	0.037	0.00	-0.014	0.003
	basal/ bolus	0.020	0.017	-0.018	0.007	**-0.022**	0.00	-0.012	0.013	0.007	0.003
Spearman	glucose/ basal	0.042	0.067	0.044	0.049	0.079	0.069	0.095	**0.160**	0.147	0.094
	glucose/ bolus	0.052	-0.003	0.003	0.016	0.024	0.001	**0.246**	0.048	0.039	0.046
	glucose/ macc	-0.025	-0.059	-0.033	-0.034	0.011	**-0.120**	-0.034	-0.049	-0.029	-0.082
	basal/ macc	-0.043	**-0.259**	-0.257	-0.226	-0.210	-0.177	-0.084	-0.136	-0.203	-0.204
	bolus/ macc	0.004	0.018	**0.057**	-0.006	0.011	0.035	0.045	0.006	0.001	0.015
	basal/ bolus	0.026	-0.016	**-0.032**	0.015	-0.030	0.013	-0.012	0.012	0.00	-0.009

least 8 to 16 hours. Lastly, the relation between glucose and macc is not relevant using the Pearson correlation, while the maximum score using the Spearman correlation is -0.120 in class 5 which could be interpreted as a very weak dependence.

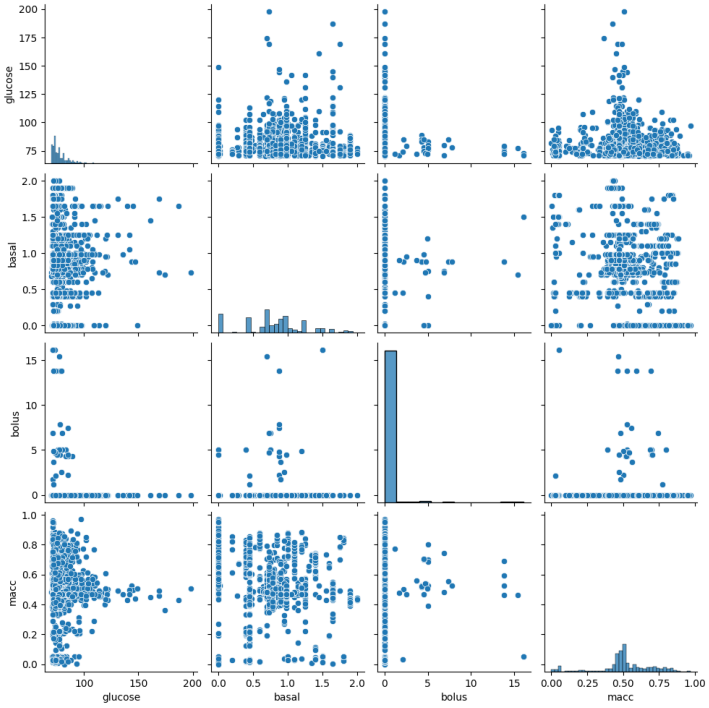
To conclude, comparing the classes, the most significant Pearson coefficients can be seen in classes 8, and 7 with a very weak relation between glucose and basal insulin, and between glucose and bolus insulin, respectively. While classes 1-9 show a very weak to weak dependence between basal insulin and macc. As a difference in the Spearman coefficients, class 5 shows a very weak relation between glucose and macc, and class 6 between glucose and bolus insulin. Lastly, it is worth mentioning that between glucose and basal insulin, there are only positive coefficients, while for basal insulin and macc, the coefficients are negative.

Moreover, pairwise plots are presented to visualize the degree to which the data fits into a straight line. Here, the data of two parameters are plotted against each other. If a trend of a straight line is visible, a linear dependence is indicated [75]. The pairwise plots can be seen in figures 19, 20, 21, 22, 23 illustrating very high variations, especially with increasing classes and time horizons.

For class 0 which is visualized in figure 19 (a), there cannot be seen any particular strong relationship among the parameters as also shown by the Pearson and Spearman correlation analysis. Higher bolus insulin is more often applied with glucose values between 60 and 70 mg/dL. In addition, for basal and bolus insulin, there seems to be a very weak linear relation for few values since more often there is no bolus insulin infused. If basal insulin is applied, there could be a very weak relation between basal insulin and macc. Likewise, the pairwise plots for class 1 visualized in figure 19 (b) shows a possible dependence between bolus and basal insulin. Moreover, macc seems to be more often increased with normal glucose values between 70 and 100 mg/dL. In particular, increased dosages are more often noticed for glucose levels under 100 mg/dL, if any bolus insulin is administered. The plots of class 2 which can be seen in figure 20 (a) behave similarly. Furthermore, in class 3, which is shown in figure 20 (b), glucose values ranging up to 200 mg/dL are more often associated with increased basal insulin dosages. In the plots of class 4, which can be seen in figure 21 (a), nothing significant is recognized. The values of macc are increased in general, and a negative relation is noticed between glucose and bolus insulin for some instances. It can be further observed that the pairwise plots of glucose and basal insulin, and glucose and macc resample a histogram from class 1 to 4 in which, for a specific threshold, both parameters tend to increase, and after the threshold glucose tends to decrease.

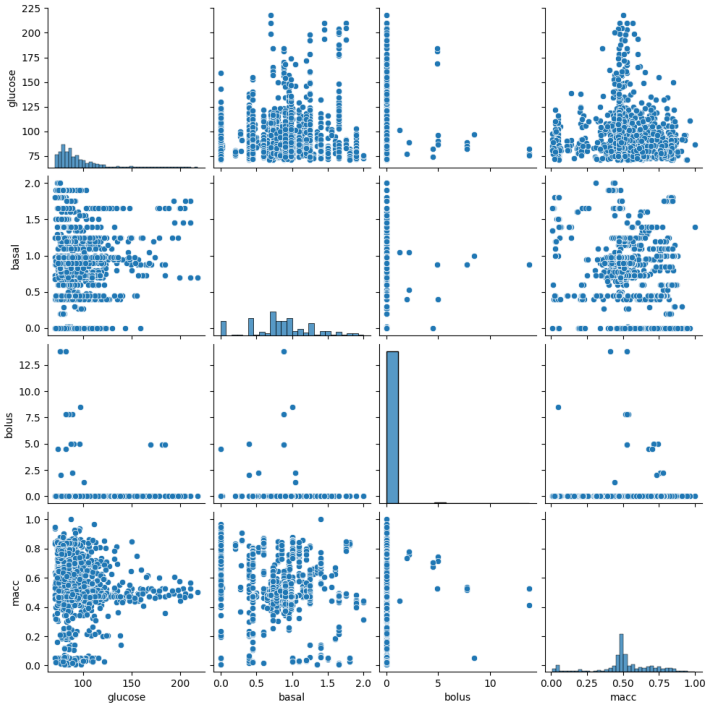


(a)

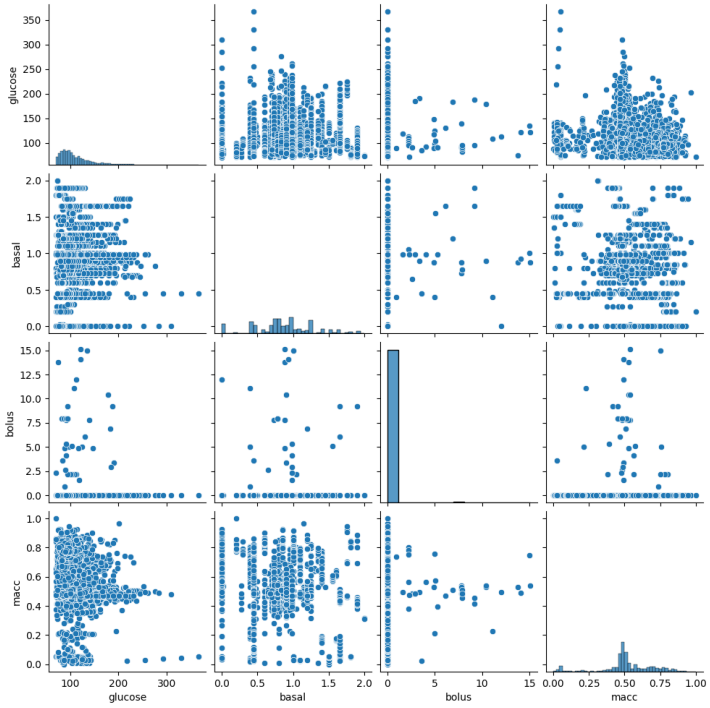


(b)

Figure 19: Pairwise-plots (1)
(a) Class 0 (b) Class 1

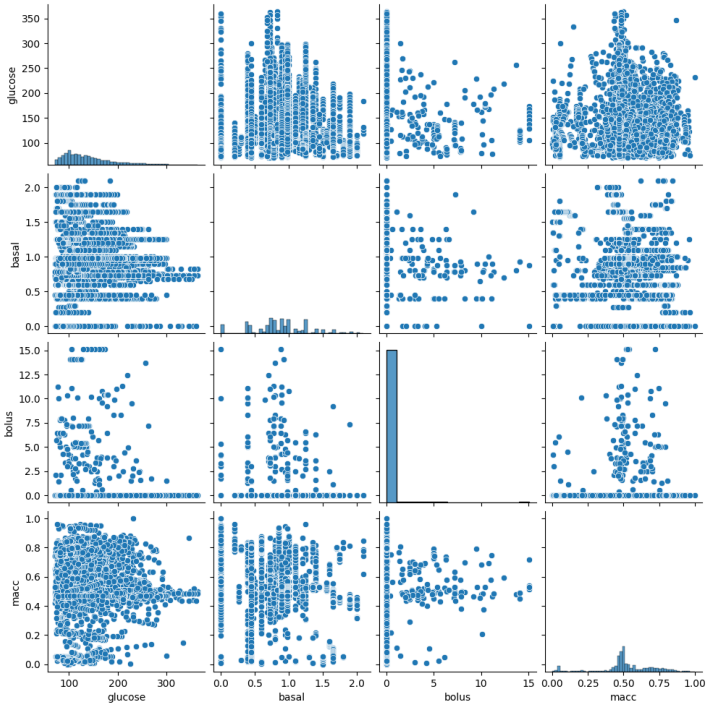


(a)

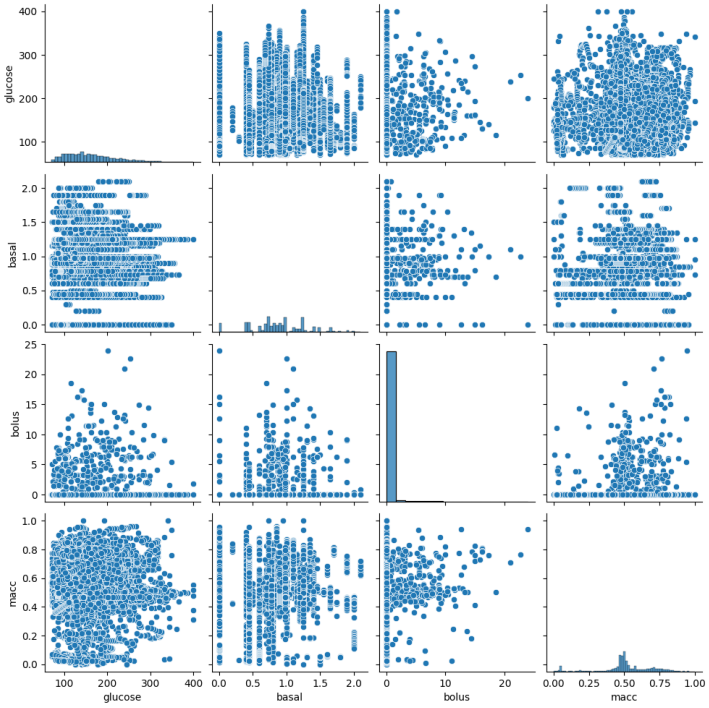


(b)

Figure 20: Pairwise-plots (2)
(a) Class 2 (b) Class 3

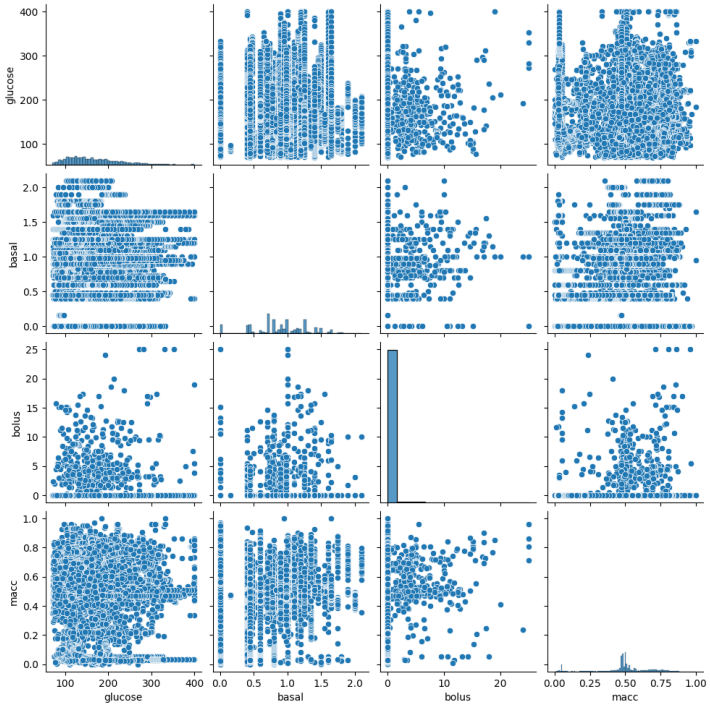


(a)

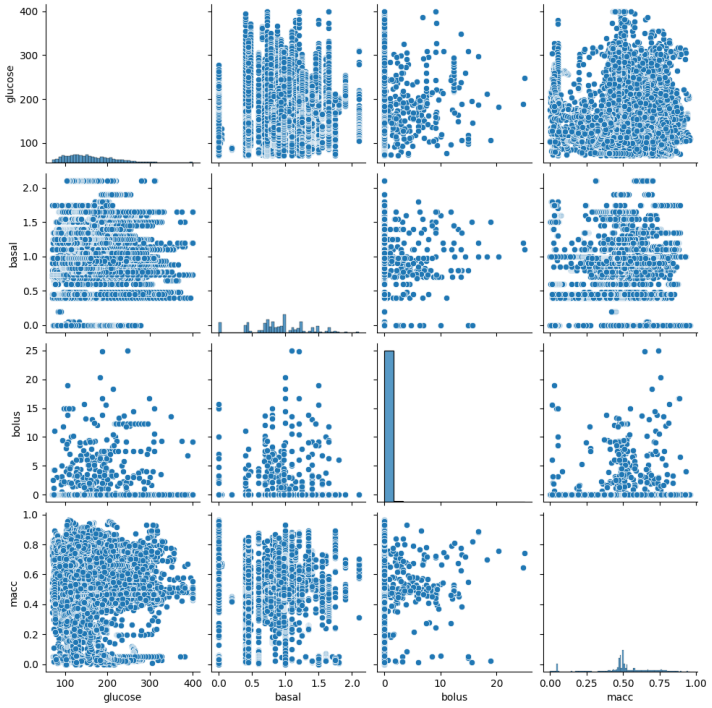


(b)

Figure 21: Pairwise-plots (3)
(a) Class 4 (b) Class 5

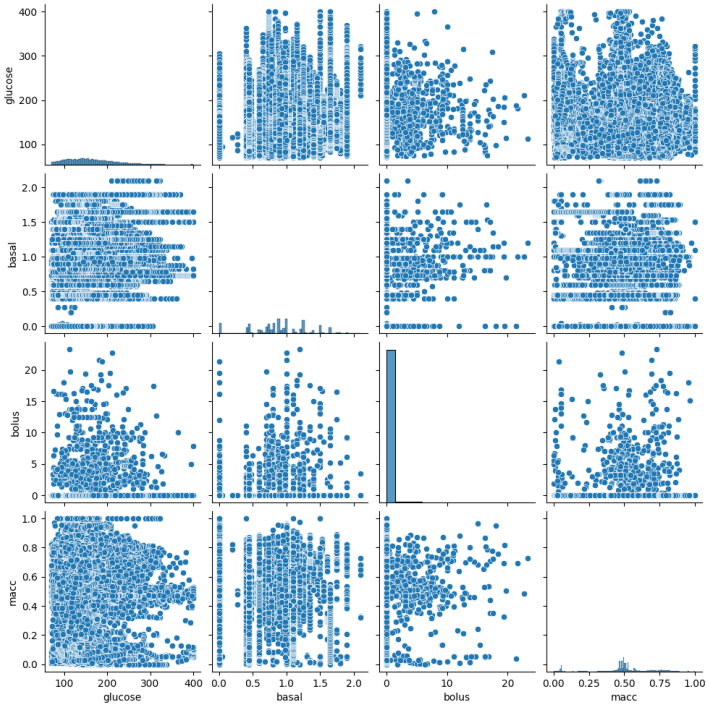


(a)

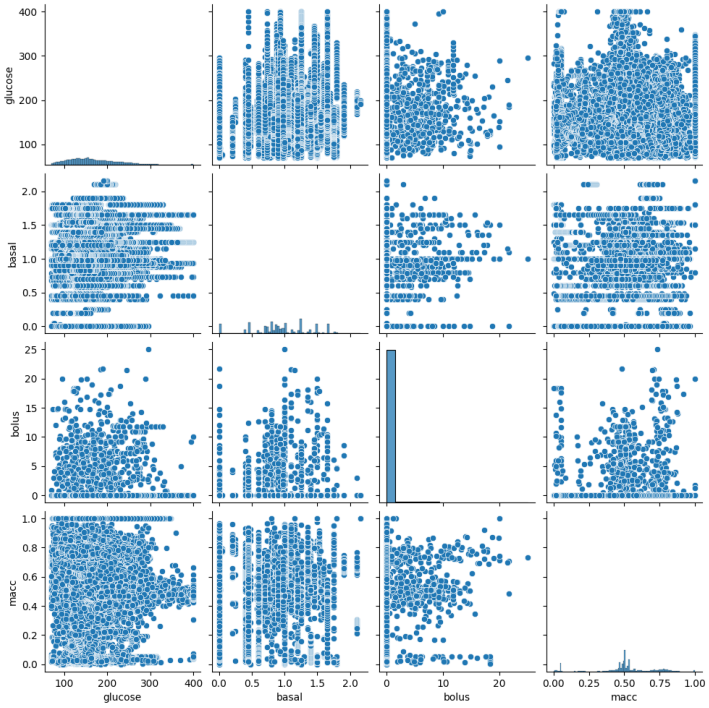


(b)

Figure 22: Pairwise-plots (4)
(a) Class 6 (b) Class 7



(a)



(b)

Figure 23: Pairwise-plots (5)
(a) Class 8 (b) Class 9

Bolus and basal insulin, and glucose and bolus insulin seem to be (very) weakly related in a negative way, for class 5, which is presented in figure 21 (b). In both plots, bolus insulin increases with the decrease of the other parameter. In general, macc is increased and high bolus seems to be associated with increased acceleration. Lastly, bolus and basal insulin are most increased with glucose values between 200 and 300 mg/dL, whereas higher glucose is more often associated with lower bolus and moderate basal insulin. Classes 6 and 7 visualized in figure 22, do not variate much from class 5 and a similar positive relation can be observed between bolus insulin and macc. However, there does not seem to be any negative association between glucose and bolus insulin, or any relation between basal and bolus insulin. Likewise, for classes 8 and 9 which can be seen in figure 23 no significant differences are noticed. Bolus insulin and macc could have a very weak negative dependence. Besides, glucose and bolus insulin seem to be negatively associated after a glucose threshold above 200 mg/dL. Moreover, bolus and basal insulin could have a negative correlation after a threshold above 1.0 for bolus. Lastly, higher glucose values, especially above 300 mg/dL, are more often associated with increased basal dosages.
Conclusively, it can be asserted that a negative correlation score between glucose and insulin most possibly indicates the decrease of glucose and increase of basal or bolus insulin. Increased macc may be related to decreased glucose levels.

5.2.2. Individual-based Correlation

For individual-based analysis, the data-frames of each patient were concatenated and grouped by classes before the correlation was computed. Tables 26, 27, 28 and 29 in the appendix present the individual correlation coefficients for each class. Here, it can be seen that the coefficients vary between the subjects while some indicate strong relations between parameters and some show irrelevant behavior between the same parameters.
Looking at the correlation coefficients of subject 540, a weak positive relation between glucose and basal insulin is noticed in classes 9 and 8. Then, the dependence decreases and is very weak in class 5, and non-relevant from classes 4-1. Hence, the increase in glucose seems to be weakly or moderately associated with the increase of basal insulin 12-48 hours before the event. For basal insulin and macc, the coefficients do not present any notable relevance. Only classes 9 and 1 have a very weak score of -0.151 and -0.198, respectively. Lastly, glucose and macc have a very weak to weak correlation among the classes with the maximum coefficients of -0.287 and -0.226 for class 9 and 0, respectively. Furthermore, class 5 is very weak, but other classes do not indicate any significant dependence. Thus, it is noticed that the maximum coefficients of glucose and macc

overlap with the maximum coefficient of glucose and basal insulin. Other parameters do not indicate any significant relation.

Contrariwise in subject 544, it can be observed that the correlation between glucose and basal insulin is not significant during classes 9 and 8. Classes 7 and 6 have a very weak relation, which is negative and weak in classes 5 and 4 with a score of -0.365 and -0.361, respectively. Thereafter, a moderate dependence can be seen in class 3 with a coefficient of -0.611. Starting from class 2 the score decreases and is very weak in class 0. Likewise, there is a strong association between basal insulin and macc. Class 9 starts with a weak relation increasing to -0.603 in class 7 and then slowly decreases from moderate to very weak in class 0. Therefore, there is a negative dependence 30-60 minutes before hypoglycemia for glucose and basal insulin, while 8-12 hours before the event seems most significant for the correlation between basal insulin and macc. Glucose and bolus insulin show a weak dependence in class 2 with 0.310, decreasing to -0.124 in class 1 and to 0.109 in class 0. Glucose and macc indicate a very weak relationship for classes 9, 6, and 2, respectively. Then, the score increases to -0.222 in class 1 and changes to 0.210 in class 0. Additionally, the correlation analysis between bolus insulin and macc presents a very weak positive dependence in classes 6 and 5 with 0.237 and 0.280, respectively. Similarly, the coefficient for basal insulin and bolus insulin is very weak in class 5 with -0.274 and class 0 with -0.187.

For subject 552 it can be seen that the relation between glucose and basal insulin is not very relevant with a maximum coefficient of 0.241 and 0.130 in classes 5 and 4, respectively. Furthermore, between glucose and bolus insulin, between bolus and macc, and between basal and bolus insulin, no significant dependence is evident. The achieved maximum scores are 0.042 in class 9, -0.150 in class 4, and -0.165 in class 4, respectively. There is a very weak to weak relation between glucose and macc. The maximum coefficient is -0.315 in class 7 indicating a negative weak dependence 8-12 hours before hypoglycemia, which decreases to -0.207 in class 3. Then, it is non-significant in class 0. Likewise, basal insulin and macc indicate a weak relation starting with a very weak coefficient in classes 6 and 5, increasing to 0.358 in class 4. Thereafter, the correlation score decreases and is very weak with -0.232 and -0.277 in classes 1 and 0, respectively. To conclude, class 4, which represent 1-2 hours before hypoglycemia, seems most significant for subject 552 considering the pairwise correlations of all parameters, while class 7, which represents 8-12 hours before the event, is most relevant for glucose and macc.

The only relevant relation in subject 559 can be seen between glucose and basal insulin which reveals a very weak positive dependence starting from class 5 with 0.293. The

coefficient then decreases and is suddenly increased to a weak positive relation in class 0 with 0.495.

Likewise, subject 563 does not have any strong correlation coefficients. The most significant scores are seen between glucose and basal insulin with 0.299 and 0.274 in class 7 and 6, respectively. Other classes are negligible or not relevant. Between glucose and macc there is a very weak relation in classes 2 and 1 with 0.174 and -0.124, respectively, showing a change from positive to negative in 15 minutes. Basal insulin and macc have only a very weak coefficient in classes 3 and 2 with 0.170 and 0.110, respectively, and lastly, between basal insulin and bolus insulin, there seems to be a weak dependence in class 3 with 0.327. As a result, the first classes are more significant and only the correlation of glucose and basal insulin is more relevant in the latter classes.

Additionally, subject 567 does not show strong correlation scores. There is a very weak to weak correlation between glucose and basal insulin starting with -0.293 in class 6, increasing to 0.362 in class 4 which then decreases to -0.234 in class 0. For basal insulin and macc, there is a very weak relation with -0.178, 0.188, -0.125, and -0.175 in classes 7, 6, 5, and 4, respectively, which then increases to -0.202 in class 2. Lastly, the other parameters are less relevant with only very weak coefficients less than 0.160 or -0.160.

For subject 570, a positive very weak to weak dependence can be seen between glucose and basal insulin. Class 9 indicates a very weak relation with 0.236, while classes 7 and 6 illustrate a weak relation of 0.336 and 0.327, respectively. The coefficient is moderate with 0.547 in class 2 and then decreases to a very weak negative association. Bolus insulin and macc seem to have a positive weak association in class 1 with 0.224, while basal insulin and bolus insulin have a very weak association in class 7 with 0.184 and a weak to moderate relation in class 5 with 0.359. Glucose and macc have a very weak relation in classes 2 and 1 with 0.247 and -0.186, respectively. Other parameters are not that relevant.

Subject 575 indicates a possible weak relation between glucose and basal insulin in classes 7 and 6 with -0.273 and -0.168, respectively, and a weak association between basal insulin and macc in classes 9 and 8 with -0.164 and -0.127, respectively. Whereas, other parameters do not indicate any significant dependence.

The correlation analysis of subject 584 presents a very weak relationship between glucose and basal insulin from class 9-6. Then, the coefficient is moderate in class 5 with -0.628 and weak in class 3 with -0.359. It starts decreasing in class 1 with -0.106 and is again moderate in class 0 with 0.528. Thus, the maximum coefficient is achieved in class 5 with a negative relation while class 0 is moderate with a positive dependence. Furthermore,

glucose and bolus insulin have a very weak association from class 7-5 and reach their maximum score in class 3 with 0.227. For glucose and macc, a very weak relation starts in class 6 with -0.169 increasing to 0.266, 0.126, 0.242, and 0.330 from class 5-2. For basal insulin and macc, a very weak relation can be observed with -0.227 in class 5 decreasing to -0.127 in class 2. Thereafter, class 1 has no relevance, and class 0 indicates a very weak relation of -0.175. Other parameters do not reveal a relevant dependence.

Subject 588 indicates less relevant dependence between glucose and basal insulin with a very weak score of -0.193 in class 5 which increases to 0.204, and 0.205 in class 4 and 3. Then, the relation decreases and is increased with the maximum value in class 0 with 0.287. Glucose and bolus insulin do not indicate any dependence, likewise glucose and macc, bolus insulin and macc and bolus and basal insulin. For basal insulin and macc starting from class 8, there is a strong relation with -0.752 which is the maximum score among all classes. The correlation then decreases to -0.670 in class 7. Class 6 only indicates a weak relation of -0.258, while class 5 is moderate with -0.581. Thereafter, the coefficient slowly decreases to -0.105 in class 2, and increases again to -0.212 in class 1. Class 0 is very weak with only -0.123.

Subject 591 shows a very weak relation between glucose and basal insulin in class 0 with a score of 0.201. For glucose and macc there is a very weak relation in class 5 with -0.125, with -0.210 in class 2, and with 0.140 in class 0. For basal insulin and macc, there is a very weak to weak relation starting from class 9 with -0.259, which increases to -0.304 in class 7. Class 5 is irrelevant, while class 4 is increased with -0.369, which then decreases to -0.113 in class 0. Other parameters do not show any significant association.

Lastly, looking at the correlation coefficients of subject 596 it can be seen that there is a very weak relation between glucose and basal insulin with 0.172 in class 9, and with -0.130 in class 6. Glucose and macc reveal a very weak relation in classes 5 and 3 with 0.201 and -0.257, respectively. Basal insulin and macc have a fragile relation in classes 8 and 5 with 0.189 and 0.119, respectively. Lastly, bolus insulin and macc show a very weak relation in classes 3 and 1 with -0.169 and -0.115, respectively. Other parameters do not indicate any dependence.

To summarize, it can be stated that the most significant classes change from subject to subject and that the maximum scores do not always overlap within the participants. Furthermore, basal insulin seems to be most impactful, followed by bolus insulin and then the macc. In particular, the maximum scores are obtained in subjects 544, 584, and 588, but in different classes for either glucose and basal insulin or basal insulin and macc.

5.3. Deep Learning models

This subsection reports the population-based deep learning models consisting of a ResNet, an LSTM, and a hybrid model utilizing 9 classes in subsection 5.3.1. Then, the person-specific models are compared in subsection 5.3.2. Finally, the same experiments are investigated with reduced classes classifying the onset of hypoglycemia up to 4 hours before in subsections 5.3.3 and 5.3.4.

5.3.1. Population-based Models using 9 Classes

First, up to 24 hours before the hypoglycemic event was classified while utilizing classes 0-8 for training. Table 12 presents the results of the macro average metrics for each subject and each model and highlights the best values. On average, the LSTM model performs best however, the values of the metrics differ between the subjects. Thus, no model is superior for all individuals. The hybrid model tested on subject 570 obtained the best accuracy of 43% with a great difference. Nevertheless, most samples of the popular class were detected in subject 570, as to why the accuracy does not reveal the classification capacity of the model. The second best accuracy is obtained at 38% with the hybrid model as well for subject 567, followed by an accuracy of 37% for subject 563 with the ResNet model, and for subject 544 with the LSTM model. Moreover, the ResNet model reports the best macro average precision with 41% for subject 575 while the other models only reach a precision of 29%, followed by a precision of 40% in subject 567 with the hybrid model. Subject 567 has the best macro average recall of 47% with the LSTM model, again with a great difference. Thereafter, the second best recall is reported for subject 559 with 42%, who has the best values in every with the LSTM model. Lastly, the best macro average F1-measure is achieved with 39% for subject 567 with the LSTM model as well, followed by subjects 552, 559, and 563 having an F1-measure of 34% with the LSTM model. Consequently, the LSTM model shows a better harmony between precision and recall. In total, the LSTM model is better with at least three metrics for subject 540 with a great difference, subjects 552, 559, and 563 with a moderate difference, and subject 591 with a minimal difference only. Therefore, the onset of hypoglycemia is better classified with an LSTM model for 5 out of 12 persons. Whereas, the ResNet model is slightly better with at least three metrics for subjects 584, 588, and 596 making 3 out of 12 persons. For subject 588, the ResNet model and the LSTM share the same precision and recall, and the F1-measure is better with 0.01 in the LSTM model but the ResNet has better accuracy. The hybrid model is only better

with at most two metrics and often with no strong differences from the other models. Conclusively, the LSTM model obtained better results on average, especially for the recall and the F1-measure, while ResNet does not show great variations. Moreover, it can be seen that on average, the precision is worse than the recall and the accuracy is the worst metric. Overall, the best performance for all metrics is obtained for subject 567. Nevertheless, considering all maximum values, the model has a poor overall classification ability and cannot even achieve a recall of 50%, whereas the best F1-Measure is only 39%. In addition, the training accuracy of the LSTM model was between 30% and 40%. In the following, the performance of the single classes is investigated for a population-based LSTM model since the LSTM model was selected as superior considering the overall results. Table 13 presents the precision, recall, and F1-measure of each class and each subject, in which the best values of each metric among the subjects are highlighted. Considering all subjects, it is noticed that the first classes are better identified while the latter are less distinguishable. The first class which represents the hypoglycemic event itself is mostly detected correctly. In detail, a recall of at least 80% for subjects 544, 570, and 591 is achieved, while the remaining subjects obtain a recall of at least 97%. The precision behaves very similarly leading to the worst F1-measure of 88% in subject 544 and the best F1-measure of 100% in subject 567. Turning now to the performance of class 1, a decrease in the classification ability and more variation between the subjects is evident. In total, 6 subjects have a recall of at least 70%, and the population recall considering all subjects is 66%. In contrast, the precision is very low with 39% for all subjects, inducing many false alarms. Only subject 567 has a precision greater than 50% with 58%, who has the best performance with a recall of 77% and an F1-measure of 66%. Whereas, the worst performance is noticed in subject 544 with an F1-measure of 40%, a recall of 47%, and a precision of 35%, indicating a great difference. Thus, on average more than 60% of all hypoglycemic cases are detected up to 15 minutes before but with an average F1-measure of 49%. Class 2 shows even more variations and reveals a population-based precision, recall, and F1-measure of 16%, 49%, and 24%, respectively. The worst metrics are obtained in subjects 575 and 588 with an F1-measure of 17%. The precision is 12% and 11%, while the recall is 34% and 36%, respectively. Contrariwise, the best precision can be seen in subject 567 with only 25%, who also has the best F1-measure with 37%, while the best recall is seen in subject 544 with 71% but with a precision of only 16%. Consequently, the model is not capable of alerting hypoglycemic cases 30 minutes as well as 15 minutes before the event since too many false alarms occur. Then, the performance of classes decreases radically with increasing prediction

horizons. For class 3, the best values can be seen in subject 567 with precision, recall, and F1-measure of only 21%, 44%, and 29%, while the population-based mean values are 13%, 31%, and 18%, respectively. Thus, a great difference between recall and precision is recognized. Besides, the worst metrics are 8%, 18%, and 11% for precision, recall, and F1-measure, respectively in subject 584. Likewise, class 4 behaves similarly. The best precision is 24% in subject 570, the best recall is 55% in subject 588 and the best F1-measure is 25% in subject 575. Hence, subject 567 does not have the best results anymore and only achieves 15%, 41%, and 22% for precision, recall, and F1-measure, respectively. The worst performance is seen in subject 544 with 0% in all metrics. The population-based precision, recall, and F1-measure are 14%, 27%, and 18%, respectively. Class 5 performs very poorly with precision, recall, and F1-measure of 15% considering the population. Subject 552 is above the average with 18%, 40%, and 25% for precision, recall, and F1-measure, respectively. Moving to class 6, the best values are seen in subject 579 with 22%, 57%, and 32% and the population-based metrics are 24%, 20%, and 22%, respectively for precision, recall, and F1-measure. Besides, class 7 obtains its best metrics in subject 567 with an F1-measure of 23% while the population F1-measure for all subjects is 16%. Lastly, in comparison, class 8 has an improved performance with average metrics of 42%, 31%, and 35%, for precision, recall, and F1-measure, respectively. The best precision is seen at 48%, the best recall at 63%, and the best F1-measure at 54% in subject 544. Contrariwise, the worst performance is seen in subject 570 with 40%, 9%, and 14% for precision, recall, and F1-measure, respectively.

To conclude, it can be seen that the first classes can be better classified, despite having fewer samples, and the latter classes cannot be well differentiated. Thus, the model can only identify the hypoglycemic event 0-30 minutes before with an acceptable recall but decreased precision. Since early stopping is used, it is identified that subjects who trained for more epochs also achieved better classification results. Contrariwise, subjects 544, 570, 575, 588, 591, and 596 trained for less than 5 epochs, and do not show any improvement in the validation loss. Subject 567 had its best validation loss after 9 epochs but training for more epochs led to overfitting.

As reported previously, the LSTM model is not superior for every subject and some subjects have better average metrics with a ResNet model. Thus, considering the metrics of the single classes for each subject using a ResNet model which can be seen in the appendix in table 30, it is noticed that the maximum values in average decrease and that subject 567 has not the best results anymore. The best values for class 0 are obtained with subjects 584 and 591, with an F1-measure of 97% each, lower than the

maximum value using an LSTM model. However, subject 570 has an increased recall and F1-measure, while the precision is decreased in comparison. Other subjects have worse or similar results compared to the results of the LSTM model. For class 1, the best recall is seen in subject 559 with 82%, the best F1-measure in subject 591 with 54% while the precision is also best in subject 567 but with 53%. A decrease in the overall performance is evident, while the performance in subject 591 increases using the ResNet model. The worst results are obtained in subject 588, the recall is decreased to 28% for subject 567, and the F1-measure is 36%. Similar results can be seen for the other classes. Overall, there are stronger variations between the subjects and the classes using the ResNet model. Subject 563 has the best F1-measure for class 4 with 20%, while the best recall is noticed in subject 544 with 35%. The best precision is 16% in subjects 563 and 575. Lastly, the best recall of 63% and F1-measure of 63% can be seen in subject 588 in class 8. Overall, the precision is even more decreased than with the LSTM model. Considering the hybrid model, it can be seen that the recall in class 0 is strongly decreased in most of the subjects, and decreases to at least 48% for subject 563. The metrics of subject 567 are still the highest but with decreased values. For class 1, the best precision is 41%, the best recall is 64% and the best F1-measure is 46%, illustrating again a great decrease compared to the results of the other approaches. Still, despite having worse results, most subjects behave similarly to the performance of the LSTM model.
Moreover, the comparison of the population performances for each class and each model presented in table 14 reveals that the LSTM model is indeed superior considering all subjects. Notably, classes 0-2, 4, and 8 are better classified. Classes 3, 5, and 7 have slightly increased metrics with the hybrid approach, while the ResNet model produces better results for class 6.
Finally, figure 24 visualizes the confusion matrices of the best subject for each model. Here, subject 563 was selected for the ResNet model, and subject 567 was selected for the LSTM and hybrid models. For the LSTM model, class 0 is only misclassified with class 1, while most of the samples of class 1 are classified correctly or as class 2. Furthermore, the model cannot distinguish samples belonging to class 3 from class 1 or 2 with high confidence while more instances are classified correctly. Thus, even if some samples are misclassified in the first classes, the hypoglycemic event could still be identified but with shifted time. Starting with class 4, the performance decreases and low precision is noticed. In short, the LSTM model indeed outperforms the other models, with a better recall, especially for the first classes. In addition, misclassifications are more often with the nearest neighbors.

Table 12: Macro average metrics of each model for each subject using 9 classes
Abbreviations: ACC = accuracy; M-PR = macro-precision; M-RC = macro-recall; M-F1-M = macro-F1-measure

Subject	**Model**	**Metric**			
		ACC	M-PR	M-RC	M-F1-M
540	ResNet	0.26	0.23	0.31	0.25
	LSTM	**0.28**	**0.31**	**0.38**	**0.32**
	Hybrid	0.27	0.29	0.30	0.28
544	ResNet	0.24	0.29	**0.38**	**0.29**
	LSTM	**0.37**	0.27	0.35	0.28
	Hybrid	0.23	**0.30**	0.35	**0.29**
552	ResNet	0.26	0.32	0.37	0.31
	LSTM	**0.27**	**0.33**	**0.41**	**0.34**
	Hybrid	0.24	0.31	0.31	0.28
559	ResNet	0.25	0.30	0.37	0.29
	LSTM	**0.28**	**0.32**	**0.42**	**0.34**
	Hybrid	0.23	**0.32**	0.35	0.25
563	ResNet	**0.37**	**0.33**	0.37	0.33
	LSTM	0.31	**0.33**	**0.41**	**0.34**
	Hybrid	0.34	0.29	0.33	0.26
567	ResNet	0.34	0.31	0.28	0.28
	LSTM	0.36	0.37	**0.47**	**0.39**
	Hybrid	**0.38**	**0.40**	0.41	0.34
570	ResNet	0.25	0.27	0.26	0.25
	LSTM	0.26	**0.32**	0.35	**0.31**
	Hybrid	**0.43**	0.27	**0.36**	0.29
575	ResNet	0.31	**0.41**	0.36	**0.31**
	LSTM	0.28	0.29	0.36	0.29
	Hybrid	**0.32**	0.29	**0.38**	0.30
584	ResNet	**0.21**	**0.32**	**0.33**	**0.28**
	LSTM	**0.21**	0.26	0.31	0.27
	Hybrid	0.20	0.28	0.30	0.26
588	ResNet	**0.34**	**0.30**	**0.37**	0.29
	LSTM	0.25	**0.30**	**0.37**	**0.30**
	Hybrid	0.21	0.28	0.30	0.26
591	ResNet	0.22	0.29	**0.35**	**0.30**
	LSTM	**0.27**	**0.31**	0.34	**0.30**
	Hybrid	0.21	0.28	0.29	0.26
596	ResNet	**0.26**	**0.31**	**0.40**	**0.32**
	LSTM	**0.26**	0.30	0.39	0.31
	Hybrid	**0.26**	0.27	0.30	0.30

Table 13: Population-based LSTM results for each subject using 9 classes
Abbreviation: F1-M = F1-measure

Subject	**Metric**	**Class**								
		0	1	2	3	4	5	6	7	8
540	Precision	**1.00**	0.42	0.16	0.11	0.12	0.15	0.29	0.19	0.35
	Recall	0.98	0.70	0.54	0.18	0.26	0.15	0.20	0.18	0.20
	F1-M	0.99	0.53	0.25	0.14	0.16	0.15	0.23	0.19	0.26
544	Precision	0.97	0.35	0.16	0.09	0.00	0.00	0.20	0.16	0.48
	Recall	0.80	0.47	**0.71**	0.23	0.00	0.00	0.24	0.05	**0.63**
	F1-M	0.88	0.40	0.26	0.13	0.00	0.00	0.22	0.07	**0.54**
552	Precision	0.99	0.47	0.18	0.13	0.15	0.18	0.28	0.10	0.43
	Recall	0.97	0.66	0.62	0.36	0.22	**0.40**	0.22	0.06	0.23
	F1-M	0.98	0.55	0.28	0.19	0.18	**0.25**	0.25	0.07	0.30
559	Precision	**1.00**	0.40	0.16	0.13	0.14	**0.19**	0.26	0.16	0.42
	Recall	0.97	0.70	0.61	0.37	0.32	0.17	0.15	0.16	0.32
	F1-M	0.99	0.51	0.25	0.20	0.19	0.18	0.19	0.16	0.36
563	Precision	**1.00**	0.40	0.16	0.13	0.15	0.15	0.21	0.25	0.48
	Recall	0.99	0.75	0.51	0.29	0.32	0.13	0.19	0.17	0.35
	F1-M	0.99	0.52	0.24	0.18	0.20	0.14	0.20	0.20	0.40
567	Precision	**1.00**	**0.58**	**0.25**	**0.21**	0.15	0.15	**0.33**	0.22	0.48
	Recall	**1.00**	**0.77**	0.69	0.44	0.41	0.24	0.12	0.25	0.32
	F1-M	**1.00**	**0.66**	**0.37**	**0.29**	0.22	0.18	0.18	**0.23**	0.39
570	Precision	**1.00**	0.36	0.16	0.18	**0.24**	0.12	0.22	0.17	0.40
	Recall	0.80	0.60	0.33	0.30	0.20	0.06	**0.57**	0.25	0.09
	F1-M	0.89	0.45	0.22	0.23	0.22	0.08	**0.32**	0.20	0.14
575	Precision	0.94	0.30	0.12	0.19	0.17	0.11	0.25	**0.26**	0.32
	Recall	**1.00**	0.71	0.34	0.26	0.46	0.00	0.24	0.05	0.37
	F1-M	0.97	0.42	0.17	0.22	**0.25**	0.00	0.24	0.08	0.34
584	Precision	0.99	0.41	0.11	0.08	0.08	0.13	0.12	0.09	0.38
	Recall	**1.00**	0.58	0.37	0.18	0.16	0.10	0.14	0.07	0.24
	F1-M	0.99	0.48	0.17	0.11	0.11	0.11	0.13	0.08	0.29
588	Precision	0.93	0.44	0.11	0.11	0.13	0.15	0.20	0.12	**0.55**
	Recall	**1.00**	0.59	0.36	0.23	**0.55**	0.11	0.28	0.02	0.21
	F1-M	0.96	0.50	0.17	0.15	0.21	0.13	0.23	0.03	0.30
591	Precision	**1.00**	0.37	0.14	0.13	0.11	0.13	0.29	0.18	0.42
	Recall	0.80	0.50	0.44	0.40	0.08	0.21	0.20	0.14	0.32
	F1-M	0.89	0.43	0.21	0.19	0.09	0.16	0.23	0.15	0.37
596	Precision	0.99	0.44	0.19	0.09	0.11	0.11	0.13	0.19	0.40
	Recall	**1.00**	0.75	0.46	**0.49**	0.11	0.14	0.01	**0.26**	0.33
	F1-M	0.99	0.56	0.27	0.15	0.11	0.12	0.02	0.22	0.36

Table 14: Performance of each class for each model using 9 classes
Abbreviations: Avg = Average; F1-M = F1-measure

	Metric	**Avg**	**Class**								
			0	1	2	3	4	5	6	7	8
ResNet	Precision	0.30	0.92	0.35	0.15	0.13	0.13	0.14	0.24	0.18	0.43
	Recall	0.34	0.93	0.60	0.36	0.19	0.13	0.11	0.27	0.30	0.21
	F1-M	0.30	0.92	0.44	0.21	0.15	0.13	0.12	0.25	0.22	0.28
LSTM	Precision	0.31	0.98	0.39	0.16	0.13	0.14	0.15	0.24	0.18	0.42
	Recall	0.39	0.96	0.66	0.49	0.31	0.27	0.15	0.20	0.14	0.31
	F1-M	0.33	0.97	0.49	0.24	0.18	0.18	0.15	0.22	0.16	0.35
Hybrid	Precision	0.30	0.95	0.27	0.16	0.14	0.13	0.16	0.23	0.22	0.42
	Recall	0.35	0.83	0.47	0.39	0.30	0.25	0.21	0.09	0.33	0.26
	F1-M	0.30	0.88	0.34	0.22	0.19	0.17	0.18	0.13	0.26	0.32

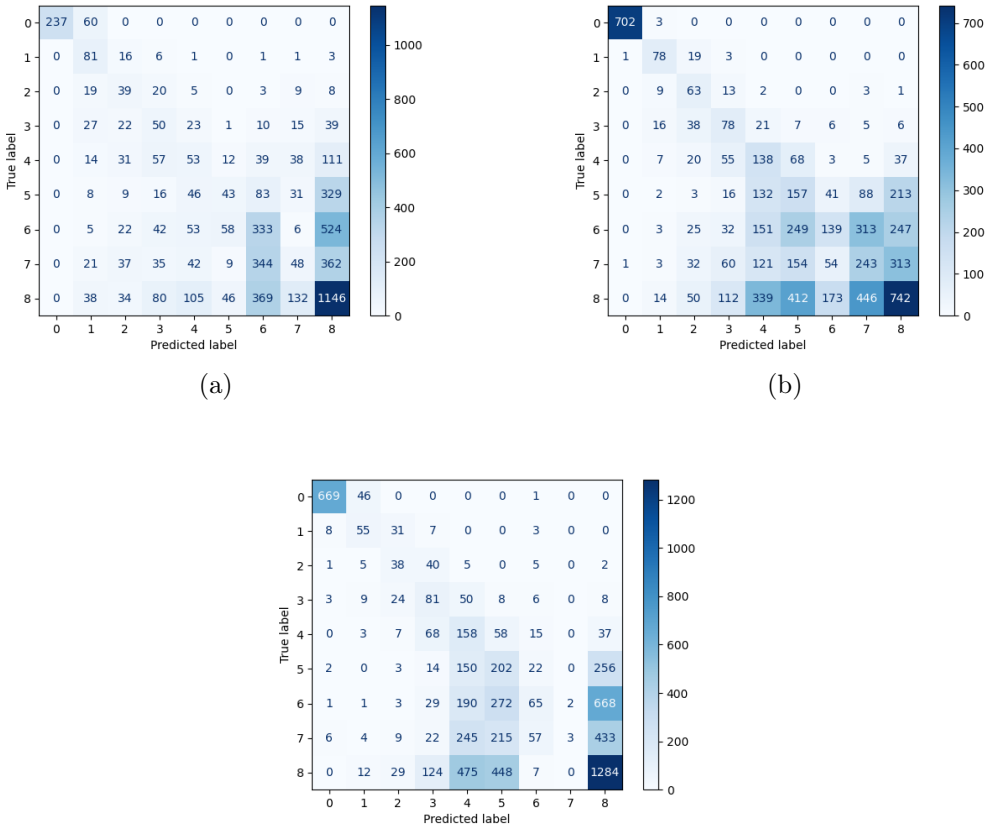


Figure 24: Population-based confusion-matrices across 9 classes
(a) Subject 563 trained with the ResNet model (b) Subject 567 trained with the LSTM model (c) Subject 567 trained with the hybrid model

5.3.2. Subject-specific and Population-based Models using 9 Classes

As demonstrated above, the metrics of the subjects vary, as to why subject-specific models were approached utilizing transfer learning. The macro averages of each subject trained with person-specific models are summarized in table 15. Here, the best performance is achieved in subject 567 with an accuracy of 39% using a ResNet model. The other metrics obtain their best values using the LSTM model with a precision of 38%, a recall of 46%, and an F1-measure of 38%. Besides, the performance between the ResNet and the LSTM model does not vary much for subject 567. On average, it can be seen that the LSTM model does not produce the best outcomes for most of the subjects anymore, and is only superior with at least three metrics in four subjects. Whereas, the ResNet

model is better with at least three metrics in five subjects. Subject 540 behaves very similarly in both models. Nevertheless, considering all subjects, it is noticed that the best F1-measures are achieved with the LSTM model. Overall, there do not seem to be great differences between the results of the LSTM and the ResNet models. However, if great variations are present, the LSTM model is superior such as in subjects 563, 570, 584, and 588. In contrast, the ResNet model is only significantly better in subjects 575 and 591. The hybrid model is not superior in any subject. Subsequently, LSTM is selected again for the detailed comparison of population-based and subject-specific approaches. The performance of single classes for each patient obtained with the population-based LSTM model is presented in table 17, while the results of the subject-specific LSTM model can be seen in table 16. When comparing both approaches, it is noticed that the overall performance is improved for most of the subjects with person-specific models. Particularly, for class 0, the lowest values which are obtained in the population-based models improve from 93%, 76%, and 86%, to 97%, 89%, and 94% for the subject-specific models for precision, recall, and F1-measure, respectively. Altogether, subject 540 has an increased recall, subject 544 has a decreased precision but increased recall and F1-measure, subjects 559, 588, and 596 have increased metrics, and subjects 591 and 570 have significantly increased metrics. The recall improved from 76% to 94% and from 79% to 89%, while the F1-measure improved from 86% to 96% and from 88% to 94%, respectively. Contrariwise, subjects 552, 567, and 575 have decreased performances. Likewise, most subjects perform slightly better with the subject-specific models in class 1. Whereas, subjects 559, and 588 have worse outcomes. The most significant change can be observed in subject 570. Here, the precision increases from 40% to 44%, the recall from 62% to 72%, and the F1-measure from 48% to 55%. Furthermore, the precision, recall, and F1-measure of subject 591 improves from 39%, 53%, and 45% to 44%, 78%, and 56% with the person-specific models. Thus, 10% and 25% more hypoglycemic cases can be predicted up to 15 minutes before hypoglycemia using an individualized model, respectively for subjects 570 and 591. Finally, the minimum values of the population-based models change from 25%, 46%, and 33% to 20%, 52%, and 29% for precision, recall, and F1-measure, respectively for the subject-specific models but the metrics are observed in different subjects in both approaches. The precision worsens while the other metrics improve. Likewise, similar outcomes can be described for class 2. Subjects 599, 570, 591, and 596 decrease in performance, whereas the remaining subjects have better metrics with the person-specific models. The most significant increase is seen in subject 540 with a precision, recall, and F1-measure from 15%, 57%, and 23% to 16%, 64%,

and 26%. The lowest values do not vary much but are obtained for different persons in both approaches. Contrariwise, the maximum values are the same and obtained for the same persons. Moving to class 4, fewer variations are seen. Subjects 559, 584, and 588 decrease significantly, and have a very poor performance in the person-specific models. Furthermore, subjects 570, 575, and 596 increase in performance. Similarly, the highest values are obtained for the same persons for precision and F1-measure but with decreased values compared to the population-based model. In addition, the best recall can be seen in subject 575 with 42% in the population-based model decreasing to 37% in the personalized model in subject 563. The lowest values do not change. Then, class 5 shows similar results but while no instances are classified correctly in subject 575 in the population-based model, the precision increases to 16%, the recall to 41%, and the F1-measure to 23% with a subject-specific approach. In addition, classes 6 and 7 behave similarly as well, most subjects have slightly increased values, and subjects 544, 559, and 591 have decreased performance. Finally, in class 8, the performance only increases for subjects 567, 570, and 588 and decreases for the remaining test persons.

The learning processes are similar as well. Subjects 552, 575, and 584 stopped after reaching 100 epochs and could further train, subject 540 had its best loss value after 50 epochs, subject 591 after 19 epochs, while the remaining subjects had their best validation loss only after at most 4 epochs.

In short, it was demonstrated that most subjects achieve better results with subject-specific models, especially in the first classes and in the recall metric. From the population metrics considering all patients, presented in table 18, it is noticed that the accuracy of both LSTM approaches is very similar. Besides, all metrics in classes 0, 1, and 5 increase for the subject-specific model, class 2 has only a better recall for the subject-specific model, while the precision and F1-measure are decreased. Class 3 remains the same in all metrics, while classes 4, 6, 7, and 8 are better with the population-based model. Furthermore, again it can be seen that the LSTM model outperforms the other models. Even the population-based LSTM model is better for classes 0-2 than the ResNet and the hybrid model. The population-based hybrid model and the subject-specific ResNet model are slightly better in class 3, the subject-specific ResNet is slightly better in class 4, the subject-specific hybrid and the ResNet model are better in class 5, while the subject-specific ResNet is better in class 6. Then, the population-based hybrid model has increased metrics in class 7, and lastly, the population-based LSTM model is superior in class 8. It can be further seen that the individualized ResNet model is significantly increased in all macro average metrics compared to the population-based ResNet model.

Lastly, the hybrid model increases in accuracy and the macro average recall for the subject-specific model, while the population-based hybrid model does not vary much from the results of the LSTM model.
Thus, none model is suitable for classifying all classes well and the best metrics vary among the classes. The first classes are better classified with the LSTM model, while the person-specific ResNet model can better identify the latter classes. Still, the overall classification ability of long-term prediction horizons is not sufficient. In short, individualized training mostly impacts the performance of the ResNet model while the LSTM model obtains the same macro average metrics in both approaches and only profits in the first classes. Lastly, figure 25 reveals the best confusion matrices for the LSTM and ResNet models since the hybrid model was not superior for any specific patient. Subjects 567 and 570 are selected for the LSTM model, while subject 575 is selected for the ResNet model. Commonly, classes 4 to 9 cause the most misclassification while the first classes seem to have good precision and recall. Figure 26 shows the confusion matrices for each population-based and subject-specific model reflecting the described behaviors. While the LSTM models indeed behave similarly it can be observed that the ResNet model has an increased recall and the instances are better classified, with less misclassification in the first classes using the personalized approach. The hybrid models behave similarly as well and it can be seen that class 6 is better classified but still with a poor classification ability.

Table 15: Subject-specific macro average metrics of each model for each subject using 9 classes

Subject	**Models**	**Metrics**			
		ACC	M-PR	M-RC	M-F1-M
540	ResNet	**0.33**	**0.33**	0.39	0.32
	LSTM	0.30	0.32	**0.40**	**0.33**
	Hybrid	0.30	0.31	0.39	0.31
544	ResNet	0.24	0.26	**0.46**	0.29
	LSTM	**0.36**	0.27	0.39	0.29
	Hybrid	0.25	**0.31**	0.51	**0.32**
552	ResNet	0.26	**0.32**	**0.38**	**0.31**
	LSTM	**0.23**	0.31	0.37	**0.31**
	Hybrid	0.21	0.28	**0.38**	0.28
559	ResNet	**0.27**	0.31	**0.38**	**0.31**
	LSTM	0.23	0.30	0.36	0.30
	Hybrid	0.26	**0.35**	0.37	0.27
563	ResNet	**0.24**	0.24	0.32	0.25
	LSTM	0.23	**0.28**	**0.39**	**0.30**
	Hybrid	0.23	0.25	0.34	0.26
567	ResNet	**0.39**	0.34	**0.46**	0.37
	LSTM	0.38	**0.38**	**0.46**	**0.38**
	Hybrid	0.34	0.28	0.37	0.29
570	ResNet	0.24	0.27	0.31	0.26
	LSTM	**0.36**	**0.39**	**0.40**	**0.36**
	Hybrid	0.29	0.25	0.32	0.26
575	ResNet	**0.30**	**0.33**	**0.44**	**0.34**
	LSTM	0.23	0.27	0.41	0.30
	Hybrid	0.24	0.28	0.39	0.28
584	ResNet	0.20	0.34	0.32	0.27
	LSTM	**0.22**	**0.35**	**0.38**	**0.34**
	Hybrid	0.15	0.22	0.29	0.23
588	ResNet	0.25	0.26	0.33	0.27
	LSTM	**0.29**	**0.29**	**0.34**	**0.30**
	Hybrid	0.26	0.27	0.32	0.27
591	ResNet	**0.24**	**0.31**	0.34	**0.31**
	LSTM	0.21	0.28	0.33	0.28
	Hybrid	0.24	0.30	**0.35**	0.28
596	ResNet	**0.27**	**0.33**	**0.41**	**0.33**
	LSTM	0.25	0.30	0.40	0.32
	Hybrid	0.24	0.25	0.30	0.26

Table 16: Subject-specific LSTM results for each subject using 9 classes

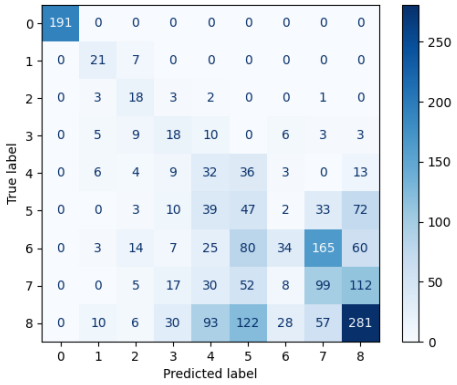
Subject	**Metric**	**Class**								
		0	1	2	3	4	5	6	7	8
540	Precision	**1.00**	0.35	0.16	0.11	0.12	0.13	0.40	0.21	0.39
	Recall	**1.00**	0.79	0.64	0.20	0.18	0.11	0.26	0.19	0.24
	F1-M	**1.00**	0.49	0.26	0.14	0.14	0.12	0.31	0.20	0.30
544	Precision	0.97	0.20	0.12	0.09	0.00	0.00	0.24	0.23	**0.60**
	Recall	0.92	0.54	0.60	0.39	0.00	0.00	0.48	0.19	0.42
	F1-M	0.94	0.29	0.19	0.14	0.00	0.00	0.32	0.21	**0.50**
552	Precision	**1.00**	0.59	0.15	0.08	0.06	0.13	0.17	0.15	0.48
	Recall	0.93	0.52	**0.71**	0.37	0.12	0.20	0.18	0.09	0.25
	F1-M	0.96	0.55	0.25	0.13	0.08	0.15	0.17	0.11	0.33
559	Precision	**1.00**	0.38	0.16	0.09	0.08	0.12	0.33	0.17	0.36
	Recall	0.99	0.65	0.57	0.16	0.19	0.20	0.10	0.29	0.13
	F1-M	**1.00**	0.48	0.25	0.11	0.12	0.15	0.16	0.21	0.19
563	Precision	**1.00**	0.43	0.16	0.13	0.16	0.12	0.02	0.18	0.29
	Recall	0.99	**0.82**	0.58	0.36	**0.37**	0.08	0.01	0.18	0.15
	F1-M	**1.00**	0.56	0.25	0.19	0.22	0.10	0.02	0.18	0.20
567	Precision	**1.00**	0.44	**0.27**	0.19	0.14	0.14	**0.42**	0.28	0.52
	Recall	**1.00**	0.75	0.67	0.33	0.31	0.23	0.09	**0.31**	**0.45**
	F1-M	**1.00**	0.55	**0.39**	0.24	0.19	0.17	0.14	**0.29**	0.48
570	Precision	**1.00**	0.44	0.13	**0.20**	**0.25**	**0.30**	0.29	**0.32**	0.58
	Recall	0.89	0.72	0.18	0.33	0.29	0.15	**0.68**	0.16	0.20
	F1-M	0.94	0.55	0.15	**0.25**	**0.27**	0.20	**0.41**	0.21	0.30
575	Precision	0.98	0.33	0.14	0.15	0.14	0.16	0.22	0.18	0.10
	Recall	0.97	0.67	0.44	**0.51**	0.35	**0.41**	0.22	0.15	0.01
	F1-M	0.98	0.44	0.21	0.23	0.20	**0.23**	0.22	0.16	0.02
584	Precision	**1.00**	**1.00**	**0.27**	0.10	0.15	0.06	0.08	0.02	0.45
	Recall	**1.00**	0.75	0.67	0.28	0.25	0.07	0.17	0.01	0.27
	F1-M	**1.00**	**0.86**	**0.39**	0.14	0.19	0.06	0.11	0.01	0.34
588	Precision	0.98	0.33	0.08	0.07	0.14	0.16	0.38	0.09	0.42
	Recall	**1.00**	0.56	0.29	0.08	0.17	0.25	0.42	0.01	0.30
	F1-M	0.99	0.41	0.13	0.08	0.15	0.20	0.40	0.01	0.35
591	Precision	0.99	0.44	0.10	0.06	0.02	0.10	0.32	0.12	0.34
	Recall	0.94	0.78	0.36	0.20	0.04	0.10	0.10	0.12	0.33
	F1-M	0.96	0.56	0.15	0.10	0.03	0.10	0.16	0.12	0.33
596	Precision	**1.00**	0.52	0.18	0.09	0.14	0.13	0.06	0.17	0.45
	Recall	**1.00**	**0.82**	0.40	0.45	0.19	0.23	0.01	0.21	0.28
	F1-M	**1.00**	0.64	0.25	0.15	0.16	0.17	0.01	0.19	0.34

Table 17: Population-based LSTM results with less test data for each subject using 9 classes

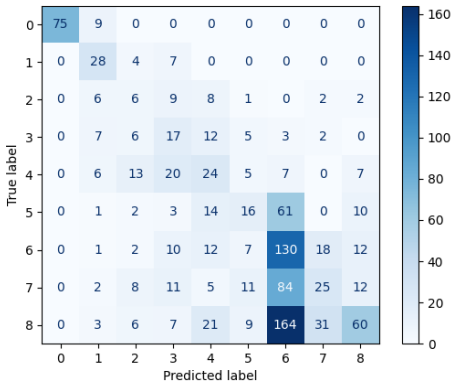
Subject	**Metric**	**Class**								
		0	1	2	3	4	5	6	7	8
540	Precision	**1.00**	0.37	0.15	0.11	0.14	0.13	0.35	0.25	0.43
	Recall	0.99	0.76	0.57	0.19	0.25	0.12	0.21	0.28	0.23
	F1-M	**1.00**	0.50	0.23	0.14	0.18	0.12	0.27	0.26	0.30
544	Precision	**1.00**	0.26	0.11	0.10	0.00	0.00	0.21	0.50	**0.65**
	Recall	0.81	0.46	0.60	0.33	0.00	0.00	0.45	0.18	**0.59**
	F1-M	0.90	0.33	0.18	0.16	0.00	0.00	0.29	0.26	**0.62**
552	Precision	**1.00**	0.52	0.14	0.08	0.05	0.15	0.14	0.09	0.59
	Recall	0.95	0.52	**0.71**	0.40	0.08	**0.32**	0.13	0.05	0.27
	F1-M	0.97	0.52	0.24	0.14	0.06	**0.20**	0.13	0.06	0.36
559	Precision	**1.00**	0.39	0.19	0.13	0.10	0.19	0.36	0.22	0.46
	Recall	0.97	0.68	0.68	0.27	0.26	0.20	0.21	0.25	0.27
	F1-M	0.98	0.49	0.29	0.17	0.14	0.19	0.26	0.24	0.34
563	Precision	**1.00**	0.42	0.16	0.13	0.16	0.12	0.02	0.18	0.28
	Recall	0.98	0.78	0.58	0.36	0.37	0.07	0.02	0.18	0.15
	F1-M	0.99	0.54	0.25	0.19	0.22	0.09	0.02	0.18	0.20
567	Precision	**1.00**	0.43	0.25	**0.20**	0.13	0.12	**0.45**	0.28	0.49
	Recall	**1.00**	0.71	0.63	0.33	0.31	0.19	0.08	**0.30**	0.44
	F1-M	**1.00**	0.53	0.36	**0.25**	0.19	0.15	0.13	**0.29**	0.46
570	Precision	**1.00**	0.40	0.18	0.14	**0.30**	**0.30**	0.27	0.31	0.55
	Recall	0.79	0.62	0.24	0.21	0.29	0.15	**0.67**	0.17	0.18
	F1-M	0.88	0.48	0.20	0.17	**0.30**	**0.20**	**0.38**	0.22	0.28
575	Precision	0.93	0.25	0.13	0.13	0.12	0.00	0.18	0.31	0.39
	Recall	**1.00**	0.71	0.46	0.29	**0.42**	0.00	0.18	0.04	0.43
	F1-M	0.97	0.37	0.20	0.18	0.19	0.00	0.18	0.08	0.41
584	Precision	**1.00**	**1.00**	**0.27**	0.10	0.16	0.06	0.08	0.02	0.45
	Recall	**1.00**	0.75	0.67	0.28	0.25	0.07	0.17	0.02	0.28
	F1-M	**1.00**	**0.86**	**0.39**	0.14	0.19	0.06	0.11	0.02	0.34
588	Precision	0.95	0.41	0.07	0.13	0.11	0.17	0.29	**0.80**	0.19
	Recall	**1.00**	0.60	0.17	0.21	0.34	0.14	0.38	0.05	0.06
	F1-M	0.97	0.48	0.10	0.16	0.16	0.16	0.33	0.10	0.09
591	Precision	**1.00**	0.39	0.13	0.09	0.11	0.08	0.36	0.17	0.43
	Recall	0.76	0.53	0.38	0.35	0.09	0.12	0.26	0.11	0.39
	F1-M	0.86	0.45	0.20	0.14	0.10	0.10	0.30	0.13	0.41
596	Precision	**1.00**	0.55	0.21	0.10	0.16	0.15	0.00	0.18	0.41
	Recall	0.99	**0.79**	0.40	**0.44**	0.18	0.16	0.00	0.26	0.36
	F1-M	0.99	0.65	0.27	0.16	0.17	0.16	0.00	0.21	0.38

Table 18: Comparison of population-based and subject-specific approaches for each model using 9 classes
Abbreviations: Avg = Average; PB = Population-based; SS = Subject-specific; F1-M = F1-measure

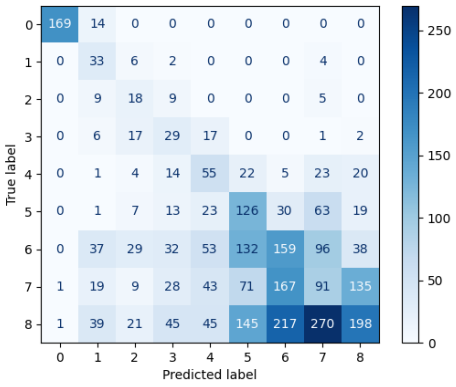
	Metrics	**Avg**	**Classes**								
			0	1	2	3	4	5	6	7	8
PB ResNet	Precision	0.28	0.94	0.33	0.15	0.11	0.10	0.11	0.19	0.18	0.37
	Recall	0.32	0.90	0.61	0.34	0.17	0.17	0.08	0.20	0.31	0.17
	F1-M	0.28	0.92	0.43	0.21	0.14	0.10	0.09	0.19	0.23	0.24
SS ResNet	Precision	0.30	0.92	0.28	0.14	0.13	0.13	0.17	0.29	0.21	0.45
	Recall	0.38	0.91	0.65	0.43	0.31	0.26	0.21	0.26	0.19	0.21
	F1-M	0.31	0.92	0.39	0.21	0.18	0.18	0.19	0.27	0.20	0.29
PB LSTM	Precision	0.31	0.99	0.39	0.16	0.11	0.13	0.14	0.24	0.21	0.45
	Recall	0.39	0.95	0.68	0.49	0.29	0.25	0.13	0.20	0.16	0.33
	F1-M	0.33	0.97	0.49	0.24	0.16	0.15	0.13	0.22	0.18	0.38
SS LSTM	Precision	0.31	1.00	0.40	0.15	0.11	0.12	0.14	0.25	0.18	0.48
	Recall	0.39	0.98	0.72	0.50	0.29	0.21	0.18	0.19	0.17	0.24
	F1-M	0.32	0.99	0.52	0.23	0.16	0.15	0.16	0.21	0.17	0.31
PB Hybrid	Precision	0.28	0.95	0.26	0.15	0.14	0.11	0.16	0.22	0.23	0.35
	Recall	0.33	0.80	0.47	0.38	0.30	0.24	0.20	0.07	0.34	0.21
	F1-M	0.29	0.87	0.33	0.21	0.19	0.15	0.18	0.11	0.28	0.26
SS Hybrid	Precision	0.28	0.92	0.26	0.12	0.11	0.11	0.19	0.22	0.23	0.36
	Recall	0.36	0.89	0.59	0.37	0.31	0.23	0.23	0.07	0.33	0.19
	F1-M	0.29	0.91	0.36	0.18	0.16	0.15	0.21	0.11	0.27	0.25



(a)

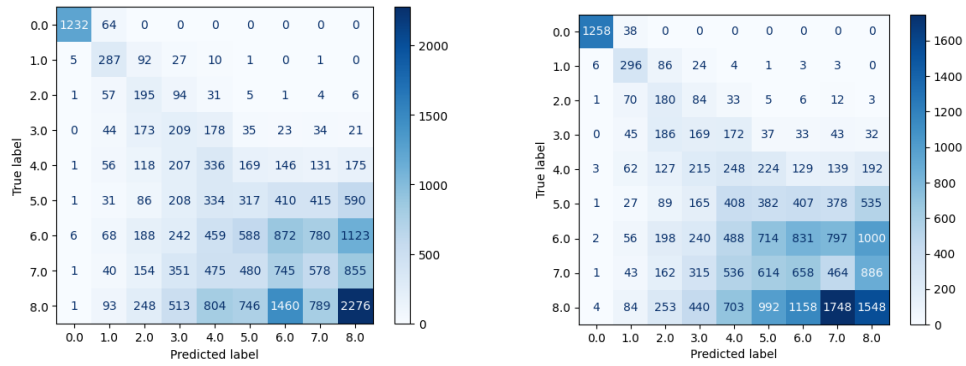


(b)

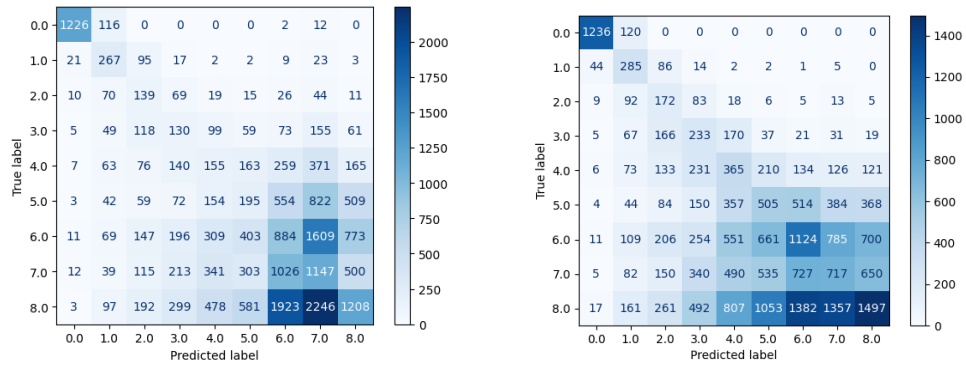


(c)

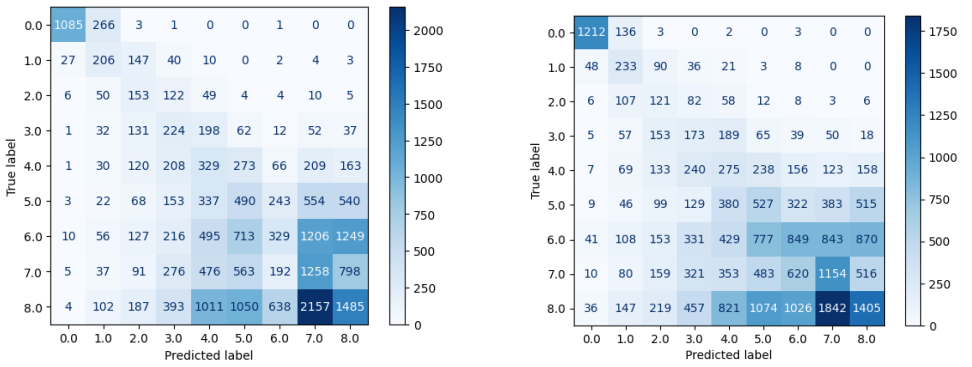
Figure 25: Subject-specific confusion-matrices across 9 classes
(a) Subject 567 trained with the LSTM model (b) Subject 570 trained with the LSTM model (c) Subject 575trained with the ResNet model



(a) (b)



(c) (d)



(e) (f)

Figure 26: Population-based and subject-specific confusion-matrices across 9 classes (a) Population-based LSTM model (b) Subject-specific LSTM model (c) Population-based ResNet model (d) Subject-specific ResNet model (e) Population-based hybrid model (f) Subject-specific hybrid model

5.3.3. Population-based Models using 6 Classes

The comparison in the prior subsections revealed that the proposed models have difficulties in distinguishing the latter classes. The first classes had better performance and induced that around 60-70% of all hypoglycemic events could be predicted on short notice. However, the models obtained low precision causing many false alarms. Thus, the same models were run for only 6 classes classifying up to 4 hours before hypoglycemia to investigate if the precision increases.

Table 20 presents the macro average metrics for each subject and each model. Similarly to the previous experiments, it can be seen that LSTM is superior and has even better metric values. The best average values are still obtained for subject 567 with an accuracy of 75%, a precision of 63%, a recall of 67%, and an F1-measure of 64% using the LSTM model. On average, around 70% of all events are classified but the precision is still lower than the recall causing an F1-measure below 65%. The lowest best metrics are an accuracy of 52% for subject 584 using a ResNet model, a precision of 51%, a recall of 49%, and an F1-measure of 49% using an LSTM model. Thus, at worst around 50% of all instances can be classified. Moreover, the LSTM model is superior in at least three metrics for all subjects but subject 596, who is slightly better classified with the hybrid model. On average, an accuracy of 60%, a precision of 55%, a recall of 58%, and an F1-measure of 55% is achieved considering the maximum values in each subject. In addition, the LSTM model had a training accuracy between 60-65%.

Table 21 reveals the performance of single classes obtained with the LSTM model. Also here, high variations between the subjects can be seen. Similarly, class 0 is distinctive and can be classified with at least 94% which is a significant increase by 14%. Almost all hypoglycemic events are detected with an average of 99%. Turning now over to the performance of class 1, it is evident that the results start to vary among the subjects. The best result is noticed in subject 567 with a precision of 66% and an F1-measure of 70%, while subject 552 has the best recall with 78% but a precision of only 61%. Compared to the results obtained with 9 classes, the precision and F1-measure of subject 567 increase by 8% and 4%, respectively, while the recall decreases by 2%. For subject 552, the performance is significantly better with 6 classes and increases from an F1-measure of 55% to 69%. Furthermore, subjects 540 and 544 show better performance in all metrics. Subjects 570, 575, 584, 588, and 591 reveal a significant increase from an average recall of 60% to 69% and an average precision of 38% to 58%. Thus, 9% more hypoglycemic states can be predicted 15 minutes before with fewer false alarms. The worst performance is seen with a precision of 46%, a recall of 52%, and an F1-measure of 49% in subject 544.

Moving to class 2, subject 567 has still the best results but with increased performance. The precision improves by 16% and the F1-measure by 14%. Nevertheless, the precision is still low with under even 50%. In some subjects only the precision and F1-measure increase, while most of the subjects increase in all metrics. The population performance considering all subjects of class 2 is 27%, 53%, and 36% for precision, recall, and F1-measure, respectively which varies significantly from the performance of class 1 especially in precision by a difference of 33%. The worst performance can be seen with 17%, 37%, and 23% in subject 584 who had an even worse performance with 9 classes. For classes 3 and 4, the results are even more decreased and the best recall is obtained at only 42% in subject 567 and 44% in subject 588, respectively. The best precision is 43% in subject 567 and 48% in subject 563. Furthermore, the best F1-measure can be observed in subject 567 with 42% and 41%, respectively for classes 3 and 4. Overall, most subjects have improved performance compared to when using 9 classes. The average F1-measure improves from 18% to 31% for class 3, and from 18% to 34% for class 4. Nevertheless, the classification ability is still insufficient. Subjects 552 and 567 have a worse recall for class 3, and subjects 559, 567, 575, and 588 have a worse recall for class 4, while the precision is significantly increased. Lastly, a significant difference can be seen for class 5 which increases from a population F1-measure considering all subjects of 15% using 9 classes to 69% using 6 classes. Thus, all subjects reveal increased performance. The best performance is noticed in subject 567 with 80%, 79%, and 80% while the worst performance is seen in subject 584 with 63%, 48%, and 55%, for precision, recall, and F1-measure, respectively. To sum up, subject 567 has the best metric values on average for all classes. While the worst performance is seen for subjects 584 and 544 with low precision and lower recall than the other patients. However, even with the reduced classes, the model cannot distinguish better between classes 3 and 4, while class 5 is significantly increased compared to the model trained with 9 classes.

The training process shows that training with fewer classes, on average most subjects stopped after 13-14 epochs, subjects 544, 567, and 575 stopped before 10 epochs, and lastly, subject 584 trained for 22 epochs which shows an improvement in learning and decrease of the validation loss over epochs.

Now coming to the ResNet model, a significant improvement can be seen as well. Subject 540 has better values for all metrics and the F1-measure increases from 53% and 21% to 80% and 54% for classes 1 and 2, respectively. Subjects 544, 552, 563, 567, 575, 588, and 596 also improve in all metrics. The best F1-measures are 99%, 68%, 44%, 39%, 38%, and 81%, for classes 0, 1, 2, 3, 4, and 5 respectively improving by 1%, 14%, 16%, 19%,

18%, and by 59% in comparison to the model trained with 9 classes. It is noticed that only class 5 has a better performance than using LSTM.
Contrariwise, the best obtained F1-measures using the hybrid model improve from 95% to 98%, from 46% to 64%, from 32% to 50%, from 28% to 41%, from 21% to 42%, from 24% to 76% for classes 0, 1, 2, 3, 4, 5 respectively. Hence, also the hybrid model shows a performance improvement and has better values than the ResNet model in classes 2, 3, and 4, whereas only the F1-measure in class 4 is better than the LSTM model.
The confusion matrices of subject 567 for each model are presented in 28. Likewise, it can be seen that most misclassifications are within the neighbor classes, especially for classes 0-2. Class 4 cannot be well differentiated. Furthermore, while the hybrid and LSTM models perform similarly, the ResNet model has more misclassifications for classes 2-4, while for class 0 all instances are classified, and most samples of class 5 are correctly classified.
Finally, the macro averages of all models in table 19 reveal an improvement with reduced classes while the LSTM model is superior in all classes. On average, 60% of all hypoglycemic events can be predicted with a precision of 55%.

Table 19: Performance of each class for each model using 6 classes

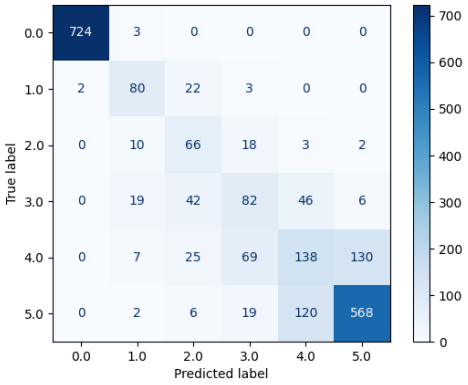
	Metric	**Avg**	**Class**					
			0	1	2	3	4	5
ResNet	Precision	0.51	0.95	0.47	0.27	0.29	0.37	0.68
	Recall	0.53	0.96	0.60	0.38	0.31	0.23	0.73
	F1-M	0.51	0.95	0.53	0.32	0.30	0.28	0.71
LSTM	Precision	0.55	0.99	0.60	0.27	0.29	0.39	0.74
	Recall	0.59	0.99	0.71	0.53	0.34	0.31	0.65
	F1-M	0.56	0.99	0.65	0.36	0.31	0.34	0.69
Hybrid	Precision	0.51	0.96	0.46	0.26	0.29	0.37	0.70
	Recall	0.55	0.95	0.69	0.41	0.31	0.26	0.68
	F1-M	0.52	0.96	0.55	0.32	0.30	0.31	0.69

Table 20: Macro average metrics of each model and each subject using 6 classes

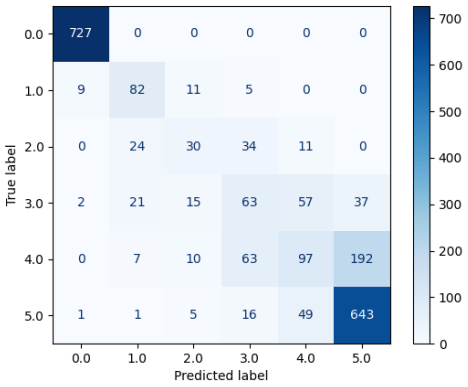
Subject	**Model**	**Metric**			
		ACC	M-PR	M-RC	M-F1-M
540	ResNet	0.56	0.51	**0.57**	0.52
	LSTM	**0.59**	**0.53**	**0.57**	**0.55**
	Hybrid	0.56	0.49	0.55	0.50
544	ResNet	0.60	0.51	0.52	0.51
	LSTM	**0.61**	**0.52**	0.53	**0.52**
	Hybrid	0.60	0.51	**0.55**	0.51
552	ResNet	0.57	0.50	0.57	0.52
	LSTM	**0.59**	**0.53**	**0.59**	**0.55**
	Hybrid	**0.59**	0.49	0.55	0.51
559	ResNet	**0.60**	0.48	0.51	0.49
	LSTM	**0.60**	**0.54**	**0.60**	**0.55**
	Hybrid	0.58	0.51	0.56	0.53
563	ResNet	0.62	0.54	0.56	0.54
	LSTM	**0.65**	**0.63**	**0.63**	**0.61**
	Hybrid	0.62	0.54	0.58	0.55
567	ResNet	0.74	0.59	0.59	0.58
	LSTM	**0.75**	**0.63**	**0.67**	**0.64**
	Hybrid	0.71	0.59	0.65	0.61
570	ResNet	0.55	0.45	0.46	0.43
	LSTM	**0.60**	**0.53**	**0.55**	**0.53**
	Hybrid	0.54	0.46	0.49	0.46
575	ResNet	0.66	0.55	0.54	0.52
	LSTM	**0.67**	**0.58**	**0.61**	**0.58**
	Hybrid	0.65	0.53	0.56	0.53
584	ResNet	**0.52**	0.46	0.45	0.45
	LSTM	0.47	**0.51**	**0.49**	**0.49**
	Hybrid	0.43	0.44	0.45	0.44
588	ResNet	**0.55**	0.50	0.49	0.49
	LSTM	0.54	**0.53**	**0.55**	**0.53**
	Hybrid	0.52	0.48	0.49	0.48
591	ResNet	0.55	0.47	0.47	0.46
	LSTM	**0.56**	**0.52**	**0.57**	**0.53**
	Hybrid	**0.56**	0.49	0.52	0.50
596	ResNet	0.60	**0.56**	0.58	0.56
	LSTM	0.60	0.55	0.58	0.56
	Hybrid	**0.62**	0.55	**0.59**	**0.57**

Table 21: LSTM results for each subject using 6 classes

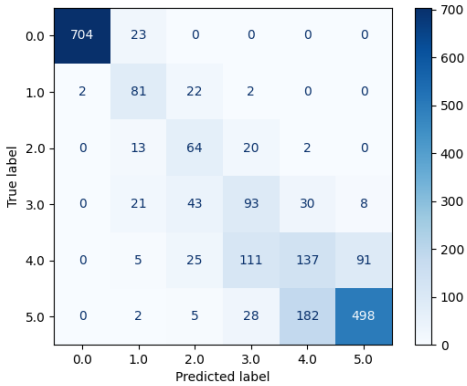
Subject	**Metric**	**Class**					
		0	1	2	3	4	5
540	Precision	**1.00**	0.60	0.28	0.26	0.38	0.69
	Recall	0.99	0.72	0.52	0.32	0.31	0.59
	F1-M	0.99	0.66	0.37	0.29	0.34	0.64
544	Precision	0.96	0.46	0.25	0.21	0.43	0.79
	Recall	0.94	0.52	0.37	0.23	0.35	0.76
	F1-M	0.95	0.49	0.30	0.22	0.38	0.78
552	Precision	**1.00**	0.61	0.26	0.24	0.36	0.74
	Recall	0.99	**0.78**	0.62	0.28	0.28	0.62
	F1-M	0.99	0.69	0.36	0.26	0.32	0.67
559	Precision	**1.00**	0.62	0.27	0.26	0.37	0.73
	Recall	**1.00**	0.70	0.66	0.37	0.23	0.64
	F1-M	**1.00**	0.66	0.38	0.31	0.28	0.68
563	Precision	**1.00**	0.63	0.35	0.36	**0.48**	0.75
	Recall	0.99	0.74	0.55	0.40	0.34	0.76
	F1-M	0.99	0.68	0.43	0.38	0.40	0.75
567	Precision	**1.00**	**0.66**	**0.41**	**0.43**	0.45	**0.80**
	Recall	**1.00**	0.75	**0.67**	**0.42**	0.37	**0.79**
	F1-M	**1.00**	**0.70**	**0.51**	**0.42**	**0.41**	**0.80**
570	Precision	0.99	0.49	0.19	0.34	0.39	0.75
	Recall	0.98	0.64	0.38	0.31	0.27	0.72
	F1-M	0.99	0.59	0.25	0.32	0.32	0.73
575	Precision	0.99	0.64	0.27	0.33	0.45	0.77
	Recall	0.99	0.73	0.52	0.34	0.36	0.71
	F1-M	0.99	0.68	0.35	0.33	0.40	0.74
584	Precision	**1.00**	0.56	0.17	0.21	0.31	0.63
	Recall	**1.00**	0.65	0.37	0.27	0.28	0.48
	F1-M	**1.00**	0.60	0.23	0.24	0.29	0.55
588	Precision	**1.00**	0.59	0.21	0.27	0.37	0.75
	Recall	0.99	0.66	0.40	0.31	**0.44**	0.53
	F1-M	**1.00**	0.63	0.27	0.28	0.40	0.62
591	Precision	0.98	0.61	0.27	0.28	0.30	0.66
	Recall	**1.00**	0.76	0.48	0.35	0.25	0.55
	F1-M	0.99	0.68	0.34	0.31	0.27	0.60
596	Precision	**1.00**	0.57	0.30	0.30	0.37	0.77
	Recall	0.97	0.63	0.55	0.37	0.32	0.67
	F1-M	0.98	0.60	0.39	0.33	0.34	0.72



(a)



(b)



(c)

Figure 27: Population-based confusion-matrices across 6 classes
(a) Subject 567 trained with the LSTM model (b) Subject 567 trained with the ResNet model (c) Subject 567 trained with the hybrid model

5.3.4. Subject-specific and Population-based Models using 6 Classes

The macro average metrics of each subject for the subject-specific models using 6 classes which are presented in table 22, show that LSTM is still the best model. The best values are achieved for subject 563 with the LSTM model with accuracy, precision, recall, and F1-measure of 70%, 66%, 69%, and 65%. Additionally, the LSTM model outperforms the other models for all patients but one who is better with the hybrid model. The ResNet model is not superior for any patient, unlike the subject-specific models using 9 classes. The performances of each model vary in most of the patients. Notably, the LSTM model is better with a visible difference except for subjects 544, 552, and 559. The LSTM model is significantly better in subjects 563, 567, 570, 575, 588, and 591. Thus, for the comparison of single classes, again the LSTM model is selected. The performance of the population-based approach can be seen in table 24, whereas the results of the individualized models are presented in table 23. Unlike in the previous experiment with 9 classes, the differences are not that strong and a significant improvement cannot be observed. For class 0, it is evident that the performance for most of the patients is the same, subjects 563, 570, and 596 have better values, while subjects 552 and 591 decrease in performance. The maximum values are obtained in five patients with 100% for all metrics in the subject-specific and population-based model. Turning now to class 1, only the performance of subjects 540, 563, 570, 588, 591, and 596 improve. In contrast, the metrics of the other subjects either remained the same or decreased. Notably, the recall of subject 567 improved from 77% to 81% while the precision and F1-measure decreased. The maximum values are the same for both approaches for the same subject with 100%, 83%, and 91% for precision, recall, and F1-measure, respectively. Likewise, in class 2 more subjects have the same or decreased values. Only subjects 544, 559, and 588 improve in performance from an F1-measure of 44% to 49%, 39% to 42%, and 19% to 25%, respectively. The maximum values also do not change and are seen in subject 563 with 74%, and 49% for recall, and F1-measure, respectively. The best precision is increased by 1% and seen in subject 544 in the individualized approach. Contrariwise, the metrics of three subjects decrease. Furthermore, similar behavior can be observed in class 3. Subjects 567 and 575 increase in performance, subjects 591 and 596 decrease, while the remaining subjects vary just slightly or remain the same. The maximum precision is the same, but recall and F1-measure are decreased in the subject-specific model. In general, it is noticed that the recall and F1-measure are decreasing after class 2. In class 4, subjects 552 and 575 are slightly increased, and subjects 567 and 591 are decreased in performance. The maximum values increase using the subject-specific model and are seen

in different subjects from 58%, 34%, 42% to 59%, 41%, and 44% for precision, recall, and F1-measure, respectively. Lastly, in class 5 it is visible that almost all subjects behave similarly with only little variations. The best precision increase from 93% to 95%, the recall is the same with 84% and the F1-measure increases from 82% to 83%.
The training process for the subject-specific LSTM model reveals that most subjects trained only for 1 epoch which supports the similar metrics for both approaches. Subject 588 trained for 12 epochs, subject 559 for 29 epochs, subject 575 for 59 epochs, and finally, subjects 544, 567, and 591 stopped at 100 epochs.
Figure 28 presents the best confusion matrix for each model. Here, subjects 563 and 567 are chosen for the LSTM model, while subject 575 is selected for the ResNet and the hybrid model. The confusion matrices of both LSTM models illustrate that the performance decreases after class 3 and that classes 3 and 4 have poor performance. The ResNet model is also poor for classes 1 and 3. Here, the misclassification is only with the nearest neighbor but on average fewer samples are correctly identified. However, classes 4 and 5 are better classified as most instances for class 4 would be predicted 2 hours before but not after, while most events of class 5 are classified correctly in comparison to the other models. For the hybrid model, similar observations are made.
From table 25 it can be seen that both approaches with the LSTM model are indeed very similar in the population performance considering all subjects, while the population-based model is better for classes 1, 2, and 4. Furthermore, the individual model is only better in class 4. Comparing the macro averages of all models, it is visible that both LSTM approaches perform best, followed by the hybrid population-based model. Likewise, the individualized models do not show a great difference and more often decrease in values using the other models. It can be seen that especially the individualized hybrid model has a significant decrease in all metrics for the first classes and is only better in class 4. The ResNet model has the worst macro average metrics and also here, the subject-specific model is only better in class 4. On average, when training with only 6 classes, more instances are classified correctly with a population-based model. The LSTM model is capable of classifying 60% of all hypoglycemic events at the right time, while 73% of all events can be predicted up to 15 minutes before with a false alarm possibility of 40%.
Lastly, figure 29 shows the confusion matrices for each population-based and subject-specific model reflecting the observed behaviors. As also recognized in the other experiments, it is noticed that the LSTM models have the best recall.

Table 22: Subject-specific macro average metrics of each model and each subject using 6 classes

Subject	**Model**	**Metric**			
		ACC	M-PR	M-RC	M-F1-M
540	ResNet	0.56	0.51	0.56	0.52
	LSTM	**0.59**	**0.54**	**0.59**	**0.55**
	Hybrid	0.53	0.50	0.56	0.50
544	ResNet	0.56	0.42	0.45	0.42
	LSTM	0.63	0.53	0.56	0.53
	Hybrid	**0.64**	**0.55**	**0.59**	**0.54**
552	ResNet	0.54	0.52	0.57	0.54
	LSTM	**0.56**	**0.56**	**0.58**	**0.56**
	Hybrid	0.54	0.50	0.54	0.51
559	ResNet	0.56	0.48	0.52	0.49
	LSTM	**0.57**	**0.52**	**0.58**	**0.53**
	Hybrid	0.56	0.51	0.56	0.52
563	ResNet	0.63	0.53	0.57	0.54
	LSTM	**0.70**	**0.66**	**0.69**	**0.65**
	Hybrid	0.63	0.55	0.59	0.54
567	ResNet	0.64	0.49	0.53	0.49
	LSTM	**0.71**	**0.56**	**0.63**	**0.58**
	Hybrid	0.59	0.47	0.51	0.47
570	ResNet	0.50	0.39	0.44	0.41
	LSTM	**0.59**	**0.57**	**0.55**	**0.55**
	Hybrid	0.46	0.39	0.40	0.39
575	ResNet	**0.69**	0.59	0.59	0.57
	LSTM	**0.69**	**0.60**	**0.62**	**0.61**
	Hybrid	**0.69**	0.58	0.61	0.58
584	ResNet	**0.51**	0.53	**0.55**	0.53
	LSTM	**0.51**	**0.56**	**0.55**	**0.55**
	Hybrid	0.42	0.49	0.45	0.47
588	ResNet	0.42	0.36	0.41	0.37
	LSTM	**0.51**	**0.46**	**0.50**	**0.47**
	Hybrid	0.46	0.38	0.40	0.39
591	ResNet	0.55	0.49	0.46	0.46
	LSTM	**0.57**	**0.54**	**0.57**	**0.55**
	Hybrid	**0.57**	0.52	0.51	0.50
596	ResNet	0.52	0.47	0.54	0.49
	LSTM	**0.57**	**0.52**	**0.59**	**0.54**
	Hybrid	0.56	0.51	**0.59**	**0.54**

Table 23: Subject-specific LSTM results for each subject using 6 classes

Subject	**Metric**	**Class**					
		0	1	2	3	4	5
540	Precision	**1.00**	0.55	0.30	0.30	0.40	0.66
	Recall	**1.00**	0.79	0.54	0.32	0.35	0.55
	F1-M	**1.00**	0.65	0.39	0.31	0.37	0.60
544	Precision	0.98	0.48	**0.38**	0.25	0.39	0.68
	Recall	0.96	0.53	0.69	0.17	0.23	0.79
	F1-M	0.97	0.50	**0.49**	0.20	0.29	0.73
552	Precision	**1.00**	0.71	0.36	0.23	0.36	0.70
	Recall	0.95	0.63	0.64	0.37	0.30	0.60
	F1-M	0.97	0.67	0.46	0.29	0.33	0.64
559	Precision	**1.00**	0.60	0.30	0.27	0.32	0.66
	Recall	0.99	0.65	0.70	0.32	0.21	0.60
	F1-M	**1.00**	0.62	0.42	0.29	0.25	0.62
563	Precision	**1.00**	0.82	0.36	0.36	**0.59**	0.81
	Recall	**1.00**	0.74	**0.74**	**0.44**	0.34	**0.84**
	F1-M	**1.00**	0.78	**0.49**	**0.40**	0.43	**0.83**
567	Precision	**1.00**	0.52	0.34	0.33	0.38	0.80
	Recall	**1.00**	0.81	0.60	0.37	0.26	0.77
	F1-M	**1.00**	0.63	0.43	0.35	0.31	0.79
570	Precision	0.98	0.54	0.20	**0.37**	0.36	**0.95**
	Recall	0.98	0.69	0.29	0.41	0.31	0.63
	F1-M	0.98	0.61	0.23	0.39	0.33	0.76
575	Precision	0.99	0.69	0.34	0.36	0.47	0.75
	Recall	0.99	0.68	0.56	0.39	**0.41**	0.72
	F1-M	0.99	0.69	0.43	0.37	**0.44**	0.73
584	Precision	**1.00**	**1.00**	0.32	0.22	0.25	0.59
	Recall	**1.00**	**0.83**	0.40	0.20	0.23	0.62
	F1-M	**1.00**	**0.91**	0.35	0.21	0.24	0.61
588	Precision	**1.00**	0.34	0.18	0.27	0.36	0.63
	Recall	**1.00**	0.53	0.41	0.25	0.24	0.56
	F1-M	**1.00**	0.42	0.25	0.26	0.29	0.59
591	Precision	**1.00**	0.66	0.26	0.29	0.30	0.70
	Recall	0.97	0.77	0.49	0.32	0.21	0.69
	F1-M	0.99	0.71	0.34	0.31	0.25	0.69
596	Precision	**1.00**	0.53	0.29	0.34	0.28	0.70
	Recall	0.99	0.76	0.57	**0.44**	0.20	0.59
	F1-M	0.99	0.62	0.38	0.38	0.24	0.64

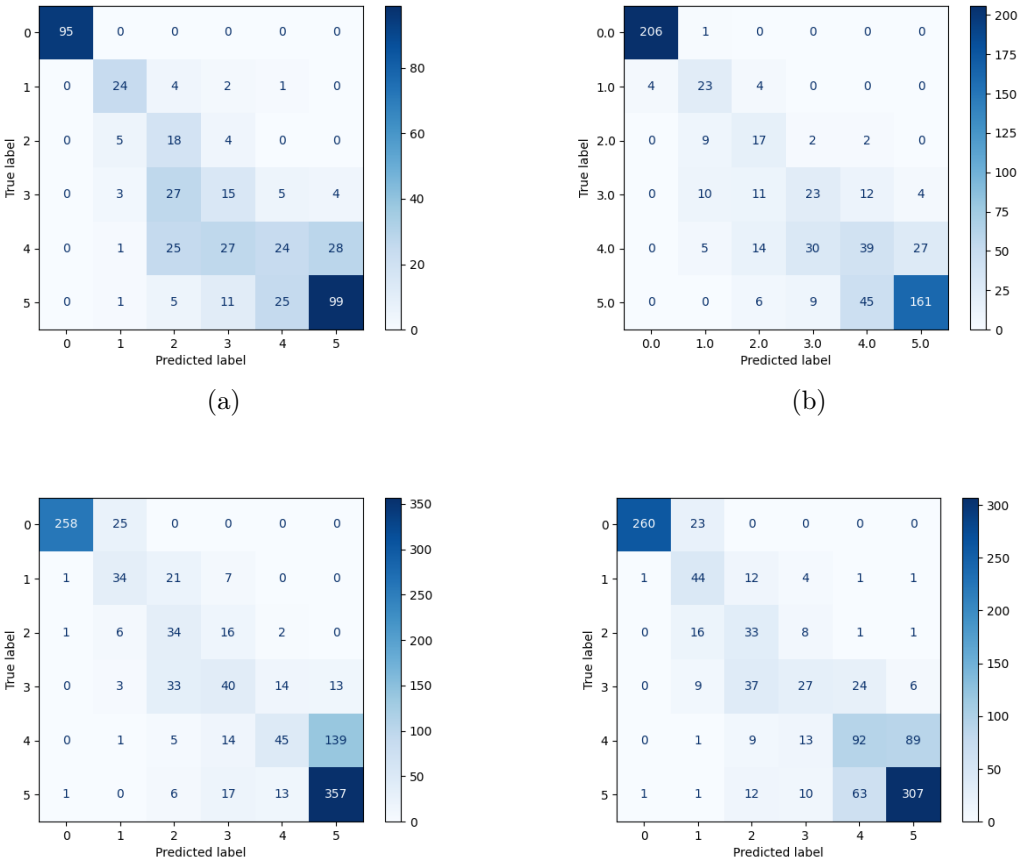
Table 24: Population-based LSTM results with less test data for each subject using 6 classes

Subject	**Metric**	**Class**					
		0	1	2	3	4	5
540	Precision	**1.00**	0.54	0.31	0.30	0.40	0.66
	Recall	**1.00**	0.79	0.54	0.33	**0.34**	0.55
	F1-M	**1.00**	0.64	0.39	0.32	0.37	0.60
544	Precision	0.98	0.48	0.35	0.25	0.38	0.68
	Recall	0.96	0.53	0.62	0.17	0.23	0.79
	F1-M	0.97	0.50	0.44	0.20	0.29	0.73
552	Precision	**1.00**	0.71	0.33	0.27	0.33	0.70
	Recall	0.99	0.71	0.69	0.35	0.27	0.60
	F1-M	0.99	0.71	0.45	0.30	0.30	0.65
559	Precision	**1.00**	0.62	0.27	0.26	0.35	0.67
	Recall	**1.00**	0.70	0.70	0.32	0.21	0.58
	F1-M	**1.00**	0.66	0.39	0.28	0.26	0.62
563	Precision	**1.00**	0.79	**0.37**	0.35	**0.58**	0.80
	Recall	0.99	0.74	**0.74**	0.44	0.33	**0.84**
	F1-M	0.99	0.77	**0.49**	0.39	**0.42**	**0.82**
567	Precision	**1.00**	0.55	**0.37**	0.32	0.42	0.79
	Recall	**1.00**	0.77	0.67	0.30	0.32	0.76
	F1-M	**1.00**	0.64	0.48	0.31	0.36	0.78
570	Precision	0.98	0.52	0.20	**0.37**	0.35	**0.93**
	Recall	0.95	0.69	0.29	0.41	0.29	0.63
	F1-M	0.97	0.60	0.23	0.39	0.32	0.75
575	Precision	0.99	0.69	0.36	0.35	0.43	0.74
	Recall	0.99	0.71	0.61	0.35	0.32	0.75
	F1-M	0.99	0.70	0.46	0.35	0.37	0.74
584	Precision	**1.00**	**1.00**	0.32	0.22	0.25	0.59
	Recall	**1.00**	**0.83**	0.40	0.20	0.23	0.62
	F1-M	**1.00**	**0.91**	0.35	0.21	0.24	0.61
588	Precision	**1.00**	0.37	0.14	0.24	0.35	0.64
	Recall	**1.00**	0.53	0.29	0.25	0.26	0.55
	F1-M	**1.00**	0.43	0.19	0.25	0.30	0.59
591	Precision	**1.00**	0.64	0.29	0.30	0.31	0.70
	Recall	0.99	0.77	0.45	0.35	0.24	0.68
	F1-M	**1.00**	0.70	0.35	0.33	0.27	0.69
596	Precision	**1.00**	0.51	0.31	0.36	0.28	0.69
	Recall	0.95	0.73	0.60	**0.48**	0.20	0.59
	F1-M	0.98	0.60	0.41	**0.41**	0.23	0.64

Table 25: Comparison of the population-based and subject-specific approach for each model using 6 classes

	Metric	**Avg**	**Class**					
			0	1	2	3	4	5
PB ResNet	Precision	0.51	0.94	0.47	0.30	0.34	0.35	0.66
	Recall	0.54	0.96	0.58	0.41	0.33	0.22	0.72
	F1-M	0.52	0.95	0.52	0.35	0.33	0.27	0.69
SS ResNet	Precision	0.49	0.94	0.39	0.27	0.30	0.37	0.68
	Recall	0.53	0.90	0.62	0.42	0.32	0.25	0.69
	F1-M	0.50	0.92	0.48	0.33	0.31	0.30	0.69
PB LSTM	Precision	0.55	1.00	0.60	0.31	0.31	0.37	0.71
	Recall	0.59	0.99	0.73	0.56	0.34	0.28	0.66
	F1-M	0.56	0.99	0.66	0.40	0.32	0.32	0.68
SS LSTM	Precision	0.55	1.00	0.60	0.30	0.31	0.38	0.71
	Recall	0.59	0.99	0.71	0.56	0.35	0.29	0.66
	F1-M	0.56	0.99	0.65	0.39	0.32	0.33	0.68
PB Hybrid	Precision	0.52	0.97	0.46	0.30	0.32	0.37	0.69
	Recall	0.56	0.95	0.70	0.44	0.34	0.27	0.68
	F1-M	0.53	0.96	0.55	0.36	0.33	0.31	0.69
SS Hybrid	Precision	0.50	0.96	0.39	0.28	0.30	0.38	0.70
	Recall	0.54	0.88	0.63	0.46	0.35	0.27	0.65
	F1-M	0.51	0.92	0.49	0.35	0.33	0.32	0.68

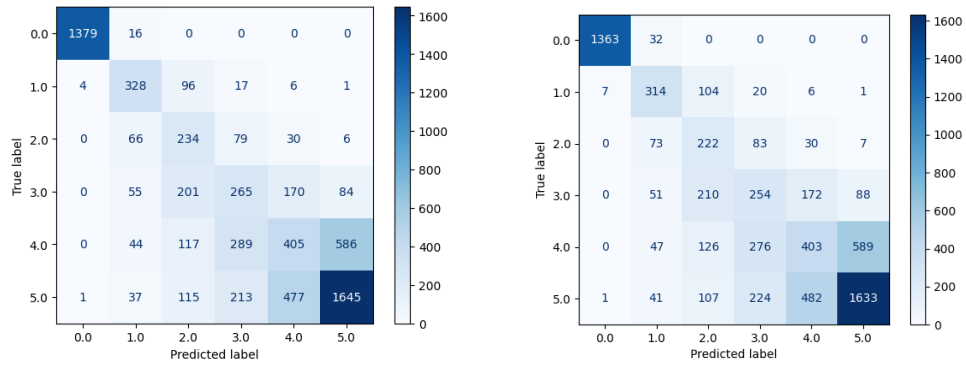
Conclusively, the proposed models cannot outperform the results of presented state-of-the-art studies, the same difficulties are faced and the precision is very low. Therefore, it is noticed that a classification system with multiple classes causes more misclassifications and worse recall. Nevertheless, as pointed out, the confusion matrices reveal that most of the miss-classifications are with the nearest neighbors when using 6 classes and up to 30 minutes before the event can be classified with a good recall.



(a) (b)

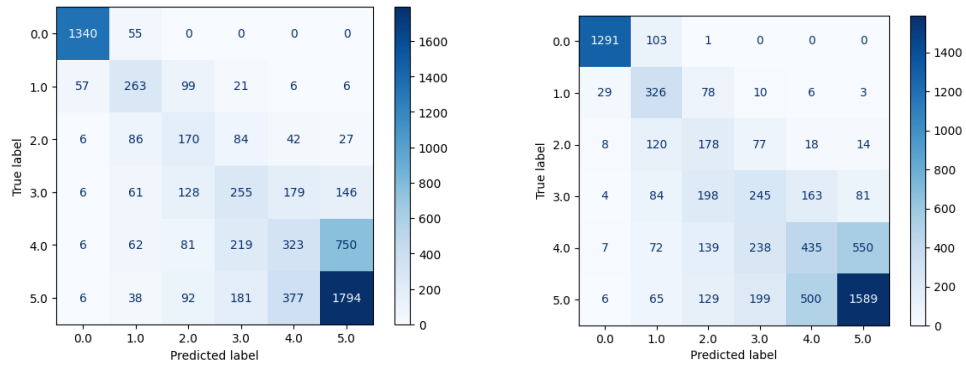
(c) (d)

Figure 28: Subject-specific confusion-matrices across 6 classes
(a) Subject 563 trained with the LSTM model (b) Subject 567 trained with the LSTM model (c) Subject 575 trained with the ResNet model (d) Subject 575 with the hybrid model



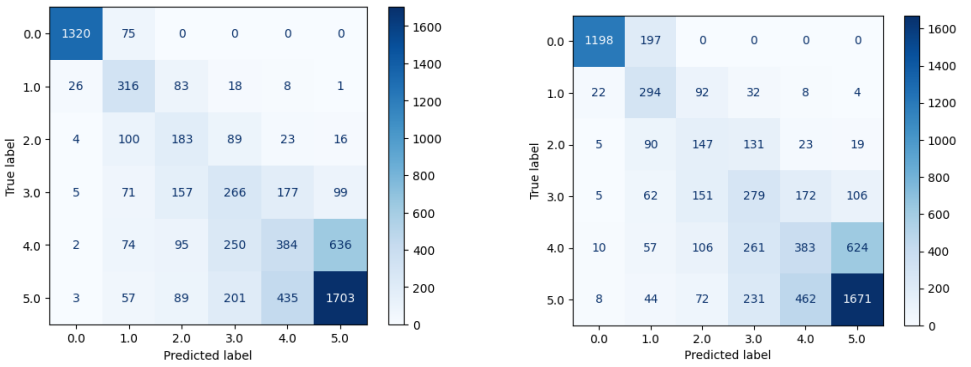
(a)

(b)



(c)

(d)



(e)

(f)

Figure 29: Population-based and subject-specific confusion-matrices across 6 classes (a) Population-based LSTM model (b) Subject-specific LSTM model (c) Population-based ResNet model (d) Subject-specific ResNet model (e) Population-based hybrid model (f) Subject-specific hybrid model

6. Discussion

As demonstrated in the previous chapters, hypoglycemia prediction is a major challenge in diabetes research. Notably, all experiments showed that hypoglycemic events, the causes leading to hypoglycemia, and the patterns before the event revealed great variations between and within the subjects. From the presented visualizations of chosen hypoglycemic data points, it can be concluded that some of the events were most possibly caused by excessive insulin dosages and could be prevented with prediction algorithms. Especially, in the data-plots of subjects 570, 575, 584, 591, and 596, it is asserted that hypoglycemia was the result of short-term actions right before the state. Moreover, the number of prior experienced hypoglycemic states could be of relevance as it was seen that most subjects experienced multiple events within the last 48 hours including severe hypoglycemia. Multiple hypoglycemic events in such a short time could increase insulin sensitivity and cause greater variations in glucose. Section 2.1.2 also reported that recent severe hypoglycemia can provoke another event. Besides, the patterns in glucose data are very individual, and often showed high variations within the day, which could depend on the lifestyle, fitness, and eating behavior of the person. Notably, the short-term decrease in glucose before the event is similar in most subjects. The data-plots illustrate that some patients did not have much variation in glucose data and had smaller insulin dosages infused, while others experienced multiple hypoglycemic events before. Glucose values were usually above 200 mg/dL and could rise to 300-400 mg/dL in selected patients, which could most possibly happen after the meal necessitating increased insulin dosages. Exercise-induced hypoglycemia could not be directly identified. Increased physical activity was noticed in some patients for a specific time interval within the last 48 hours or timely before the decrease of glucose values. However, the activity often overlapped with the administered insulin dosages. Here, for a better analysis, additional information is required, such as the activity type, intensity, and time interval of the sports session. Furthermore, data from athletes or children could reveal more insights into exercise-related hypoglycemia because the OhioT1DM is mostly represented by an age range of 40–60 years.

Additionally, the correlation analysis illustrated great variations. While some subjects revealed moderate to strong correlations, others only showed weak dependence for the same pairs of parameters. Furthermore, the maximum scores change concerning the classes. The best correlation scores considering the population were obtained between glucose and bolus insulin, between glucose and basal insulin, and between basal insulin

and magnitude of acceleration. It needs to be highlighted, that the correlation score between the latter pair is very weak to weak from classes 1-9. Thus, a direct relation for a specific PH is not evident. The maximum score was obtained with -0.296 between basal insulin and macc 15-30 minutes before hypoglycemia. Whereas, glucose and basal insulin only obtain a very weak score for at most three classes between 8-48 hours before hypoglycemia, and glucose and bolus insulin show only a relevant dependence 12-24 hours before the event. Moreover, minimal differences were observed between the Pearson and Spearman correlations, while the best coefficient between some pairs was shifted among the classes. Most values followed a similar trend. Thus, a linear dependence cannot be directly assumed, especially for glucose and bolus insulin. Considering the individual Pearson correlation coefficients, the best scores are seen between glucose and basal insulin and likewise between basal insulin and macc but with increased maximum scores indicating a weak to moderate dependence. In particular, subjects 544, 570, and 584 showed a moderate dependence between glucose and basal insulin in classes 3, 2, and 5, respectively. This demonstrates that the behavior of these parameters is more dependent on short-term prediction horizons. For subjects 554 and 584, negative behavior was observed, which could show an increase in insulin and a decrease in glucose, as concluded from the pairwise plots. When looking at all patients, increased scores more often show a positive correlation, which could be due to increased insulin administrations with increased glucose, leading to hypoglycemia, or an administration of insulin before the rise of glucose. The best coefficients between basal insulin and macc can be observed 8-24 hours before hypoglycemia for subjects 544 and 588 with a negative moderate to strong dependence. Most often, the scores are increased between classes 4-9 while in some subjects, the maximum score is noticed in short-term prediction horizons. However, the cause of dependence cannot be directly interpreted. A possible scenario is a decrease in insulin administration with planned exercise. In contrast, the relationship between glucose and macc varies greatly among subjects and classes, and the best achieved score indicates only a very weak to weak correlation. Thus, it is foreseen that not all subjects would perform well and that not all variables would show a significant association. The worst coefficients were observed in subjects 563, 575, and 596. Moreover, participants who experienced more hypoglycemic events could impact the deep learning model and lead to biased training. Hence, testing on subjects behaving differently from popular subjects could lead to worse results.

One major challenge of machine learning models in diabetes research is that data is not AI-ready and requires multiple pre-processing steps. In particular, this problem

is faced when using multiple features, which is suggested because other studies have presented that additional features such as insulin, exercise, or information about meal intake improve performance, as presented in 3.3. The OhioT1DM dataset collected data from two different cohorts and used two different wearable devices. Thus, the estimated exercise data differed and could not be used without data imputation. Here, feature engineering is required to enable uniform data representations. This work has decided to convert the step count to acceleration data, but variations between both parameters could still be noticed. Moreover, large gaps in glucose data were identified, which could cause significant information loss and disable the use of all data. Subjects 552 and 567 had the most missing glucose values with estimated gaps of 12.5 and 11 days, respectively. Consequently, Prof. Bunescu of the OhioT1DM dataset was asked about possible causes of missing data in glucose since CGM devices should continuously measure. He responded that gaps are most likely to occur if patients change their infusion set or if the sensor becomes detached or is not inserted correctly under the skin. Subsequently, these problems represent a great disadvantage of wearable sensors, leading to noise or missing values. In particular, monitoring systems are disabled to alert patients. To compensate for the gaps and reduce the number of missing values, linear interpolation was applied. As mentioned already, linear methods might not work well for glucose as to why a limit was set for allowed consecutive missing values. Better methods could help to impute more data. The described pre-processing steps require time and knowledge in feature engineering and would probably require more effort to have a standard representation that is not dependent on the sensor brands.

Moving to the deep learning models, the training process illustrated a significant difference between the training performance and the validation and test performance. If trained for sufficient epochs, almost all instances of the training data were classified correctly which demonstrates overfitting. Thus, early stopping had to be applied but resulted in shorter training periods of the model. It can be seen that some subjects were only trained for less than five epochs resulting in worse performance. Consequently, the training accuracy was between 30% and 40% when using 9 classes and between 60% to 65% when using 6 classes. Nevertheless, the validation loss did not directly show a learning behavior, could not achieve any sufficient value, and did not reflect the behavior of the training loss. These observations indicate great variations within the experienced hypoglycemic events and between the subjects. In addition, even if the model is only trained and validated on one subject, the model cannot identify the classes well in the validation data. This highlights the difference in glucose patterns within the same person. It is possible that

the validation data did not represent the learned patterns of the training data. This problem could be further led by the small size of the dataset and the nature of deep learning architectures. Another major cause could be the imbalance of classes or the bias of popular classes. Moreover, popular subjects could impact the model and lead to a bias so that less represented samples are not learned with strong connections. Nevertheless, even if having imbalanced data and fewer samples for the first classes, the latter classes had worse results using 9 classes. The recall in the first classes was significantly better but with poor precision indicating many false alarms. In this context, the applied weights could impact a bias of underrepresented classes but a model without any weights classified every instance to the popular class. Hence, in a population-based approach, the patterns leading to the onset of hypoglycemia in short-term prediction horizons are similar in most of the subjects. In contrast, it is more difficult to develop a model that is capable of predicting multiple hours before the event because each hypoglycemic event could follow a different trend. It is believed that a system based on long-term prediction horizons needs a larger dataset and individual training of the patients. It can be seen that the hypoglycemic event itself was almost classified correctly in all approaches, models, and patients while the LSTM model is superior and had the highest minimal value. This value was even increased with the individualized models and when training only with 6 classes. Furthermore, 15-30 minutes before the state obtained a sufficient recall.
Coming to the comparison of population-based and subject-specific models, it is observed that when using 9 classes, population-based models profit from transfer learning especially in short-term PHs. Thus, the prediction of hypoglycemia in 5-30 minutes seems very much possible while almost all hypoglycemic events are detected. In particular, the greatest performance improvement could be seen in subject 570. This subject interestingly had a small dataset with only 227 hypoglycemic data-points and trained for less than five epochs. Furthermore, subjects 540, 575, and 591 improved significantly. Whereas, those trained for much longer epochs with 50, 19, and 100, respectively, and belong to the group of subjects with the largest dataset. Thus, it cannot be directly asserted that less data and less training disables individualized models and performance improvement. Data of subject 570 could be still representative enough, enabling better performance with fewer epochs. Nevertheless, the trend in the other subjects shows that larger datasets lead to better training and more stable results. In contrast, models trained with 6 classes either were not greatly impacted by individualized training or had slightly decreased performance for the first classes. Improvements could be only observed with minimal increases. In particular, subject 570 again showed a better performance for classes 0 and

1. Furthermore, subjects 563 and 596 increased in performance for more classes. All were trained for only one epoch which does not reflect the previously observed behavior. Moreover, it was concluded, that the subject-specific model using 6 classes profits mostly in class 4 in the population metrics considering all subjects. In contrast, the other classes are similar or slightly decreased compared to the population-based approach. One drawback of using transfer learning is that some subjects with fewer samples train for fewer epochs and possibly cannot impact the performance. In particular, this could be the case when training with only 6 classes, since the popular classes were removed resulting in fewer samples in total. It was also noticed that the performance of most subjects did not vary much. Therefore, it is not known if the performance would be better with more samples and more training. Most probably, individual models only trained with the data of the test subject can produce better results. Nevertheless, as could be identified, the sample sizes per subject are not sufficient to train and validate individual models without training on the data of other subjects.

The macro averages for the population-based model using 9 classes reveal that the model does not work that well and that the imbalance can cause biases. Even the most popular classes are not detected that well which can be caused by the applied weights. Also, it is asserted that the latter classes are not that distinctive. The population-based LSTM model using 9 classes was capable of classifying 50-68% of all hypoglycemic events 30-15 minutes before, while the performance of the individualized model increased to 50-72%. Contrariwise, 56-73%, and 56-71% instances were correctly classified 30-15 minutes before when using 6 classes, respectively for the population-based and subject-specific model. The macro average recall of all classes improved by 20% using 6 classes and was 59%. In detail, up to 30 minutes before the event could be prevented accurately while classes 3 and 4 were insufficiently classified. When using 6 classes, 70% of all hypoglycemic events can be foreseen before 15 minutes. Notably, in subject 552 even up to almost 80% of events are identified. For most of the subjects, 60-70% of all hypoglycemic cases could be predicted up to 30 minutes before. As a conclusion, most of the classes' performance increased when using fewer classes.

It was further observed that the performance of each model varied. While using 9 classes, LSTM was slightly better than ResNet and while using 6 classes, LSTM was significantly better. Additionally, the performance of each subject differed. Consequently, individually tuned and selected models for each subject are suggested which could be more sensible for a population-based approach rather than using the same model to test each subject. In general, subject 567 achieved the best results. In particular, for a population-based model

using 9 classes, subject 563 had its best metrics with a ResNet and subject 567 with an LSTM model. Furthermore, using individualized models, subject 570 achieved good results with an LSTM model while subject 575 had a better performance with a ResNet model. With 6 classes and individualized training, subjects 563 and 567 were better with the LSTM model, and 575 was still better using the ResNet and hybrid models. Lastly, for a population-based model with 6 classes, subject 567 obtained the best results for all models. Now, looking in more detail into the data of the best subjects, it can be noticed that subject 567 experienced the third most hypoglycemic data-points, and has the second most missing glucose values of a total of 11 days. Fewer instances of class 1 are available compared to the proportion in other subjects' data. This could indicate that most events are of longer duration and could also be severe hypoglycemia. Moreover, looking at the results of the correlation analysis no variables reach a moderate to strong relation, and the maximum score is obtained with 0.362 between glucose and basal insulin for class 4. In addition, subject 567 trained for 9 epochs in the population-based LSTM model and for 4 epochs in the individualized LSTM model using 9 classes, which were 9 and 100 epochs, respectively using 6 classes. The same can be seen in subject 575 who experienced the most hypoglycemic values, and in comparison has fewer samples in class 1. Thus, the model possibly can learn better without the data of the most popular persons. In other words, as foreseen before, the biased subjects possibly impact the model and disable a population-based classification.

There are still some challenges, especially with the low precision of classes, but it needs to be considered that the model was not tuned for training 6 classes. Thus, better results could be possible. Lastly, the size of the dataset is not sufficient to teach all patterns and behaviors observed in the patients. Nevertheless, since the training data can be classified very well, it can be assumed that it is possible to classify the onset of hypoglycemia up to 4 hours before.

7. Conclusion

This thesis described that hypoglycemia is a life-threatening condition mainly affecting patients with type 1 diabetes. The state can be prevented with glucose intake or an adapted lifestyle and good management of insulin dosages. Thus, AI methods can help in predicting the onset of hypoglycemia. It was demonstrated that physical activity could have an impact on insulin sensitivity and glucose values even 24 hours after the session, while the dosage of bolus insulin is still effective 12 hours after the injection. Basal insulin is fast-acting insulin and can result in an immediate decrease in glucose values. Consequently, actions up to 24 hours before can result in hypoglycemia. The literature review summarized that prediction models classify between short- and long-term prediction horizons, while each model only focused on one prediction time. Most studies forecast glucose values while classification models are mainly based on hypoglycemia, severe hypoglycemia, exercise-induced hypoglycemia, or nocturnal hypoglycemia identification. Thus, this thesis included nine prediction horizons into one classification model ranging from 0-24 hours before the occurrence of hypoglycemia. With this concept, multiple use cases are considered and the patient is supported in planning their daytime activities and meals, and in taking short-term preventive actions. To train the models, the OhioT1DM dataset was selected since it collected glucose, basal, and bolus insulin, and activity data. This work has further explored the OhioT1DM dataset and identified drawbacks such as gaps in the estimated glucose data resulting in information loss. Moreover, since data from two different cohorts was collected, the physical activity was estimated as two different parameters. Thus, pre-processing was required. Missing glucose values were imputed with linear interpolation while larger gaps of more than 2 hours were removed, and the step count was converted to the magnitude of acceleration to have uniform data. The pre-processed parameters were plotted, and the visualizations of the last 48 hours before a hypoglycemic event for each subject revealed that most events could be foreseen and prevented. The decrease in glucose seemed to be related to the administered insulin dosages and prior experienced (severe) hypoglycemia. This thesis further investigated the correlation between glucose, basal, and bolus insulin, and the magnitude of acceleration illustrating great variations between the subjects. The best coefficients were obtained between glucose and basal insulin in short-term prediction horizons, and between basal insulin and magnitude of acceleration in long-term prediction horizons. Thereafter, the time to the onset of hypoglycemia was classified up to 48 hours before utilizing 10 classes. The chosen deep learning architectures were based on CNN and RNN models. However,

the performance was not sufficient and many instances were misclassified as to why the last class was removed. Then, subsequent experiments were classified only up to 24 hours before the event. In particular, an LSTM model, a ResNet model, and a hybrid model of both architectures were compared for a population-based and a subject-specific approach. The comparison revealed that in general LSTM models are superior. Furthermore, the individualized models had significantly increased performance. It could be observed that patients with more data-samples could lead to biased training decreasing the capability of population-based classification. In general, those subjects obtained better performance with subject-specific models and also trained for more epochs than patients with fewer samples. Best classified classes were the first classes ranging from 0-30 minutes before hypoglycemia. The latter classes could not obtain a good performance with the proposed methods and data which led to many misclassifications. Thus, the same experiments were run with six classes classifying up to 4 hours before hypoglycemia. This approach improved the general performance, in all classes, and mostly for class 5. Furthermore, it showed that the misclassifications are mostly within the nearest neighbors. Most events could be predicted and prevented but some samples were classified with shifted time. Especially, classes 3 and 4, representing 1-2 hours before the event could not be identified with high confidence and were not very distinguishable. Classifying only 6 classes, it was also noticed that the subject-specific models did not result in performance improvement, and mostly the values of metrics among the classes were similar. In general, class 4 profited from the individualized models. Altogether, the training accuracy and training loss were very accurate and indicated a correct classification of almost all samples while the same behavior was not reflected in validating and testing. Nevertheless, the proposed architecture was seen as advantageous since it can better support patients and can help in completely preventing one event if the classification capability can reach sufficient performance. The best performance using 9 classes was 31%, 39%, and 32% improving to 55%, 59%, and 56% using 6 classes for precision, recall, and F1-measure, respectively. Hence, it is concluded that around 60% of all hypoglycemic states can be predicted with a short-term prediction model. In subject 563, even 70% of all states were detected considering the macro average metrics. Consequently, it is suggested to separate between short-term and long-term classifications such as in the literature review and to not include both PHs in the same model. Furthermore, the model should be tuned for the reduced classes. In addition, it was observed that proposed deep learning models lead to overfitting due to the small dataset size, as to why conventional machine learning models should be investigated.

8. Future Work

For future work, it is suggested to tune the model which was trained with 6 classes, and focus on short-term and long-term classification separately. Here, a layered classification system can be developed in which the first model decides if the time series sequence induces a short-term, long-term, or no risk. Based on the outcome, the time sequence is then either forwarded to the short-term classification model or the long-term classification model. In this context, the short-term prediction could classify up to 2 hours or up to 4 hours before while the long-term model should classify up to 24 hours before hypoglycemia. Furthermore, different machine learning models should be explored such as SVMs or other machine learning models usually applied for smaller datasets because those achieved better performance in the literature review. As depicted above, it is also suggested to explore other data imputation methods to compensate for the missing glucose data, since glucose patterns are generally not linear. A possible method could be cubic interpolation, linear regression for time series data, or random forest regression predicting the most possible values while considering the time. Thus, a prediction algorithm that works well for glucose data can be utilized to fill in missing values. A model considering multiple PHs and transforming the forecasting task into a classification task, which seems feasible for short-term classification, can be integrated into more complex systems. One use case could be digital twins since those consist of individual patient profiles, learn the patterns of each subject, and are individualized systems continuously learning and adapting to the patient's data and behavior. Thus, it could be possible to compensate for the variations within the subject. Furthermore, fault detection could be utilized to detect the cause of hypoglycemia, and improve decision support. An artificial pancreas might also be best integrated with digital twin technologies because those are used for predicting adverse events. Furthermore, the possible impact of the utilized insulin dosage with the estimated glucose trend considering multiple variables and features could be simulated. Thus, Laubenbacher et al. conclude that digital twins can advance the possibilities for health care as not only a small perspective is known but the system considers the whole patient profile, which is especially relevant with diabetes patients since the life quality is impacted by multiple components [84]. Lastly, to investigate exercise-induced hypoglycemia, not only the acceleration data but intensity and type of activity should be considered as in the work of Cho et al in [85]. Besides, even if collected in free-living conditions, the patients should be asked to do controlled sports sessions. Moreover, data from athletes or children could give more insights into the impact of physical activity on glucose.

References

[1] Metrics, The Institute For H. ; evaluation: *Global diabetes cases to soar from 529 million to 1.3 billion by 2050.* https://www.healthdata.org/news-events/newsroom/news-releases/global-diabetes-cases-soar-529-million-13-billion-2050. – last access: 17.12.2023

[2] Felizardo, Virginie ; Machado, Diogo ; Garcia, Nuno M. ; Pombo, Nuno ; Brandão, Pedro: Hypoglycaemia Prediction Models With Auto Explanation. In: *IEEE Access* 10 (2022), S. 57930–57941. http://dx.doi.org/10.1109/ACCESS.2021.3117340. – DOI 10.1109/ACCESS.2021.3117340

[3] Association, American D.: Diagnosis and Classification of Diabetes Mellitus. In: *Diabetes Care* 33 (2010), 01, Nr. Supplement 1, S62-S69. http://dx.doi.org/10.2337/dc10-S062. – DOI 10.2337/dc10–S062. – ISSN 0149–5992

[4] Glumčević, Sabina ; Mašetić, Zerina ; Viteškić, Benjamin: Closed-loop Artificial Pancreas Development: A Review. In: *2023 46th MIPRO ICT and Electronics Convention (MIPRO)*, 2023, S. 1173–1178

[5] Mobasseri, Majid ; Shirmohammadi, Masoud ; Amiri, Tarlan ; Vahed, Nafiseh ; Fard, Hossein H. ; Ghojazadeh, Morteza: Prevalence and incidence of type 1 diabetes in the world: a systematic review and meta-analysis. In: *Health Promotion Perspectives* 10 (2020), März, Nr. 2, 98–115. http://dx.doi.org/10.34172/hpp.2020.18. – DOI 10.34172/hpp.2020.18

[6] Piersanti, Agnese ; Salvatori, Benedetta ; Göbl, Christian ; Burattini, Laura ; Tura, Andrea ; Morettini, Micaela: A Machine-Learning Framework based on Continuous Glucose Monitoring to Prevent the Occurrence of Exercise-Induced Hypoglycemia in Children with Type 1 Diabetes. In: *2023 IEEE 36th International Symposium on Computer-Based Medical Systems (CBMS)*, IEEE, Juni 2023, ""

[7] Cockcroft, E. J. ; Narendran, P. ; Andrews, R. C.: Exercise-induced hypoglycaemia in type 1 diabetes. In: *Experimental Physiology* 105 (2020), Januar, Nr. 4, 590–599. http://dx.doi.org/10.1113/ep088219. – DOI 10.1113/ep088219

[8] Bhimireddy, Ananth R. ; Sinha, Priyanshu ; Oluwalade, Bolu ; Gichoya,

Judy W. ; Purkayastha, Saptarshi: Blood Glucose Level Prediction as Time-Series Modeling using Sequence-to-Sequence Neural Networks. In: *KDH@ECAI*, 2020, ""

[9] In: Goutham, Swapna ; Kp, Soman: *Diabetes Detection and Sensor-Based Continuous Glucose Monitoring – A Deep Learning Approach*. 2021. – ISBN 978–3–030–66632–3, S. 245–268

[10] Infante, Marco ; Baidal, David ; Rickels, Michael ; Fabbri, Andrea ; Skyler, Jay ; Alejandro, Rodolfo ; Ricordi, Camillo: Dual-hormone artificial pancreas for management of type 1 diabetes: Recent progress and future directions. In: *Artificial Organs* 45 (2021), 07. http://dx.doi.org/10.1111/aor.14023. – DOI 10.1111/aor.14023

[11] Felizardo, Virginie ; Garcia, Nuno M. ; Pombo, Nuno ; Megdiche, Imen: Data-based algorithms and models using diabetics real data for blood glucose and hypoglycaemia prediction – A systematic literature review. In: *Artificial Intelligence in Medicine* 118 (2021), aug, 102120. http://dx.doi.org/10.1016/j.artmed.2021.102120. – DOI 10.1016/j.artmed.2021.102120

[12] Diouri, Omar ; Cigler, Monika ; Vettoretti, Martina ; Mader, Julia K. ; Choudhary, Pratik ; and, Eric R.: Hypoglycaemia detection and prediction techniques: A systematic review on the latest developments. In: *Diabetes/Metabolism Research and Reviews* 37 (2021), mar, Nr. 7. http://dx.doi.org/10.1002/dmrr.3449. – DOI 10.1002/dmrr.3449

[13] Ometov, Aleksandr ; Shubina, Viktoriia ; Klus, Lucie ; Skibińska, Justyna ; Saafi, Salwa ; Pascacio, Pavel ; Flueratoru, Laura ; Gaibor, Darwin Q. ; Chukhno, Nadezhda ; Chukhno, Olga ; Ali, Asad ; Channa, Asma ; Svertoka, Ekaterina ; Qaim, Waleed B. ; Casanova-Marqués, Raúl ; Holcer, Sylvia ; Torres-Sospedra, Joaquín ; Casteleyn, Sven ; Ruggeri, Giuseppe ; Araniti, Giuseppe ; Burget, Radim ; Hosek, Jiri ; Lohan, Elena S.: A Survey on Wearable Technology: History, State-of-the-Art and Current Challenges. In: *Computer Networks* 193 (2021), Juli, 108074. http://dx.doi.org/10.1016/j.comnet.2021.108074. – DOI 10.1016/j.comnet.2021.108074. – ISSN 1389–1286

[14] Babatain, Wedyan ; Bhattacharjee, Sumana ; Hussain, Aftab M. ; Hussain, Muhammad M.: Acceleration Sensors: Sensing Mechanisms, Emerging Fabrication Strategies, Materials, and Applications. In: *ACS Applied Electronic Materials* 3

(2021), Januar, Nr. 2, 504–531. http://dx.doi.org/10.1021/acsaelm.0c00746. – DOI 10.1021/acsaelm.0c00746. – ISSN 2637–6113

[15] In: Nilam, Nilam ; M., Seyed ; N., Pappur: *Therapeutic Modelling of Type 1 Diabetes*. InTech, 2011

[16] Federation, International D.: *Diabetes around the world in 2021*. https://diabetesatlas.org/#:~:text=Diabetes%20around%20the%20world%20in%202021%3A,%2D%20and%20middle%2Dincome%20countries.. – last access: 17.12.2023

[17] Buchmann, Maike ; Tuncer, Oktay ; Auzanneau, Marie ; Eckert, Alexander J. ; Rosenbauer, Joachim ; Reitzle, Lukas ; Heidemann, Christin ; Holl, Reinhard W. ; Thamm, Roma: Incidence, prevalence and care of type 1 diabetes in children and adolescents in Germany: Time trends and regional socioeconomic situation. (2023). http://dx.doi.org/10.25646/11439.2. – DOI 10.25646/11439.2

[18] Soedamah-Muthu, S.S. ; Abbring, S. ; Toeller, M.: Diet, Lifestyle and Chronic Complications in Type 1 Diabetic Patients. Version: 2011. http://dx.doi.org/10.5772/20851. In: Liu, Chih-Pin (Hrsg.): *Type 1 Diabetes*. Rijeka : IntechOpen, 2011. – DOI 10.5772/20851, Kapitel 2

[19] Jessica Lucier, Ruth S. W.: *Type 1 Diabetes*. https://www.ncbi.nlm.nih.gov/books/NBK507713/. Version: January 2023. – last access: 03.11.2023

[20] Bolli, Geremia B. ; Porcellati, Francesca ; Lucidi, Paola ; Fanelli, Carmine G.: The physiological basis of insulin therapy in people with diabetes mellitus. In: *Diabetes Research and Clinical Practice* 175 (2021), Mai, 108839. http://dx.doi.org/10.1016/j.diabres.2021.108839. – DOI 10.1016/j.diabres.2021.108839. – ISSN 0168–8227

[21] Gillespie, K. M.: Type 1 diabetes: pathogenesis and prevention. In: *Canadian Medical Association Journal* 175 (2006), Juni, Nr. 2, 165–170. http://dx.doi.org/10.1503/cmaj.060244. – DOI 10.1503/cmaj.060244. – ISSN 1488–2329

[22] Scholten, Bernt J. ; Kreiner, Frederik F. ; Gough, Stephen C. L. ; Herrath, Matthias von: Current and future therapies for type 1 diabetes. In: *Diabetologia* 64 (2021), Februar, Nr. 5, 1037–1048. http://dx.doi.org/10.1007/s00125-021-05398-3. – DOI 10.1007/s00125–021–05398–3

[23] Soedamah-Muthu, Sabita S. ; Fuller, John H. ; Mulnier, Henrietta E. ; Raleigh, Veena S. ; Lawrenson, Ross A. ; Colhoun, Helen M.: High Risk of Cardiovascular Disease in Patients With Type 1 Diabetes in the U.K. In: *Diabetes Care* 29 (2006), April, Nr. 4, 798–804. http://dx.doi.org/10.2337/diacare.29.04.06.dc05-1433. – DOI 10.2337/diacare.29.04.06.dc05–1433. – ISSN 1935–5548

[24] Haak, Thomas ; Gölz, Stefan ; Fritsche, Andreas ; Füchtenbusch, Martin ; Siegmund, Thorsten ; Schnellbächer, Elisabeth ; Klein, Harald H. ; Uebel, Til ; Drossel, Diana: Therapy of Type 1 Diabetes. In: *Experimental and Clinical Endocrinology & Diabetes* 127 (2019), Dezember, Nr. S 01, S27–S38. http://dx.doi.org/10.1055/a-0984-5696. – DOI 10.1055/a–0984–5696. – ISSN 1439–3646

[25] Cederblad, Lars ; Eklund, Gustav ; Vedal, Amund ; Hill, Henrik ; Caballero-Corbalan, José ; Hellman, Jarl ; Abrahamsson, Niclas ; Wahlström-Johnsson, Inger ; Carlsson, Per-Ola ; Espes, Daniel: Classification of Hypoglycemic Events in Type 1 Diabetes Using Machine Learning Algorithms. In: *Diabetes Therapy* 14 (2023), April, Nr. 6, 953–965. http://dx.doi.org/10.1007/s13300-023-01403-7. – DOI 10.1007/s13300–023–01403–7. – ISSN 1869–6961

[26] Adolfsson, Peter ; Rentoul, Donald ; Klinkenbijl, Brigitte ; Parkin, Christopher G.: Hypoglycaemia Remains the Key Obstacle to Optimal Glycaemic Control – Continuous Glucose Monitoring is the Solution. In: *European Endocrinology* 14 (2018), Nr. 2, 50. http://dx.doi.org/10.17925/ee.2018.14.2.50. – DOI 10.17925/ee.2018.14.2.50. – ISSN 1758–3772

[27] Mujahid, Omer ; Contreras, Ivan ; Vehi, Josep: Machine Learning Techniques for Hypoglycemia Prediction: Trends and Challenges. In: *Sensors* 21 (2021), Januar, Nr. 2, 546. http://dx.doi.org/10.3390/s21020546. – DOI 10.3390/s21020546

[28] Parcerisas, Adrià ; Contreras, Ivan ; Delecourt, Alexia ; Bertachi, Arthur ; Beneyto, Aleix ; Conget, Ignacio ; Viñals, Clara ; Giménez, Marga ; Vehi, Josep: A Machine Learning Approach to Minimize Nocturnal Hypoglycemic Events in Type 1 Diabetic Patients under Multiple Doses of Insulin. In: *Sensors* 22 (2022), Februar, Nr. 4, 1665. http://dx.doi.org/10.3390/s22041665. – DOI 10.3390/s22041665. – ISSN 1424–8220

[29] Makroum, Mohammed A. ; Adda, Mehdi ; Bouzouane, Abdenour ; Ibrahim, Hussein: Machine Learning and Smart Devices for Diabetes Management: Systematic Review. In: *Sensors* 22 (2022), Februar, Nr. 5, 1843. http://dx.doi.org/10.3390/s22051843. – DOI 10.3390/s22051843

[30] Hwong, Harn H. ; Sivakumar, Saaveethya ; Lim, King H.: Machine Learning-Based Glucose Monitoring Techniques for Diabetes Management: A Review. In: *2023 International Conference on Digital Applications, Transformation & Economy (ICDATE)*, IEEE, Juli 2023, ""

[31] Home, Philip D. ; Mehta, Roopa: Insulin therapy development beyond 100 years. In: *The Lancet Diabetes & Endocrinology* 9 (2021), Oktober, Nr. 10, 695–707. http://dx.doi.org/10.1016/s2213-8587(21)00182-0. – DOI 10.1016/s2213–8587(21)00182–0. – ISSN 2213–8587

[32] Heise, Tim: The future of insulin therapy. In: *Diabetes Research and Clinical Practice* 175 (2021), Mai, 108820. http://dx.doi.org/10.1016/j.diabres.2021.108820. – DOI 10.1016/j.diabres.2021.108820. – ISSN 0168–8227

[33] Kelly, Dylan ; Hamilton, Jill K. ; Riddell, Michael C.: Blood Glucose Levels and Performance in a Sports Camp for Adolescents with Type 1 Diabetes Mellitus: A Field Study. In: *International Journal of Pediatrics* 2010 (2010), 1–8. http://dx.doi.org/10.1155/2010/216167. – DOI 10.1155/2010/216167. – ISSN 1687–9759

[34] Paldus, Barbora ; Morrison, Dale ; Lee, Melissa ; Zaharieva, Dessi P. ; Riddell, Michael C. ; O'Neal, David N.: Strengths and Challenges of Closed-Loop Insulin Delivery During Exercise in People With Type 1 Diabetes: Potential Future Directions. In: *Journal of Diabetes Science and Technology* 17 (2022), April, Nr. 4, 1077–1084. http://dx.doi.org/10.1177/19322968221088327. – DOI 10.1177/19322968221088327. – ISSN 1932–2968

[35] Gomez, Ana M. ; Gomez, Claudia ; Aschner, Pablo ; Veloza, Angelica ; Muñoz, Oscar ; Rubio, Claudia ; Vallejo, Santiago: Effects of Performing Morning Versus Afternoon Exercise on Glycemic Control and Hypoglycemia Frequency in Type 1 Diabetes Patients on Sensor-Augmented Insulin Pump Therapy. In: *Journal of Diabetes Science and Technology* 9 (2015), Januar, Nr. 3, 619–624. http://dx.

doi.org/10.1177/1932296814566233. – DOI 10.1177/1932296814566233. – ISSN 1932–2968

[36] Zhu, Taiyu ; Li, Kezhi ; Herrero, Pau ; Georgiou, Pantelis: Deep Learning for Diabetes: A Systematic Review. In: *IEEE Journal of Biomedical and Health Informatics* 25 (2021), Juli, Nr. 7, 2744–2757. http://dx.doi.org/10.1109/jbhi.2020.3040225. – DOI 10.1109/jbhi.2020.3040225. – ISSN 2168–2208

[37] In: Sindhu Meena, K. ; Suriya, S.: *A Survey on Supervised and Unsupervised Learning Techniques*. Springer International Publishing, 2020. – ISBN 9783030240516, 627–644

[38] Fawaz, Hassan I. ; Forestier, Germain ; Weber, Jonathan ; Idoumghar, Lhassane ; Muller, Pierre-Alain: Deep learning for time series classification: a review. In: *Data Mining and Knowledge Discovery* 33 (2019), mar, Nr. 4, 917–963. http://dx.doi.org/10.1007/s10618-019-00619-1. – DOI 10.1007/s10618–019–00619–1

[39] Huang, Shuzhan ; Tang, Jian ; Dai, Juying ; Wang, Yangyang: Signal Status Recognition Based on 1DCNN and Its Feature Extraction Mechanism Analysis. In: *Sensors* 19 (2019), April, Nr. 9, 2018. http://dx.doi.org/10.3390/s19092018. – DOI 10.3390/s19092018. – ISSN 1424–8220

[40] Zihao, Zhao ; Geng, Jie ; Jiang, Wen: A Time Series Classification Method Based on 1DCNN-FNN. In: *2021 33rd Chinese Control and Decision Conference (CCDC)*, IEEE, Mai 2021, ""

[41] Karim, Fazle ; Majumdar, Somshubra ; Darabi, Houshang ; Chen, Shun: LSTM Fully Convolutional Networks for Time Series Classification. In: *IEEE Access* 6 (2018), 1662–1669. http://dx.doi.org/10.1109/access.2017.2779939. – DOI 10.1109/access.2017.2779939. – ISSN 2169–3536

[42] Zhao, Bendong ; Lu, Huanzhang ; Chen, Shangfeng ; Liu, Junliang ; Wu, Dongya: Convolutional neural networks for time series classification. In: *Journal of Systems Engineering and Electronics* 28 (2017), Nr. 1, S. 162–169. http://dx.doi.org/10.21629/JSEE.2017.01.18. – DOI 10.21629/JSEE.2017.01.18

[43] Zhao, Bendong ; Lu, Huanzhang ; Chen, Shangfeng ; Liu, Junliang ; Wu, Dongya: Convolutional neural networks for time series classification. In: *Journal of Systems*

Engineering and Electronics 28 (2017), Februar, Nr. 1, 162–169. http://dx.doi.org/10.21629/jsee.2017.01.18. – DOI 10.21629/jsee.2017.01.18. – ISSN 1004–4132

[44] Chen, Zixuan ; Nie, Zedong ; Li, Jingzhen ; Wu, Youwen ; Zhou, Hongjun: Multi-parameter Blood Glucose Prediction Algorithm for Type 1 Diabetes Based on Hybrid Neural Network Deep Learning Technique. In: *""*, 2023, S. 474–479

[45] He, Kaiming ; Zhang, Xiangyu ; Ren, Shaoqing ; Sun, Jian: Deep Residual Learning for Image Recognition. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, S. 770–778

[46] Wang, Zhiguang ; Yan, Weizhong ; Oates, Tim: *Time Series Classification from Scratch with Deep Neural Networks: A Strong Baseline*. http://dx.doi.org/10.48550/ARXIV.1611.06455. Version: 2016

[47] Aiello, Eleonora M. ; Lisanti, Giuseppe ; Magni, Lalo ; Musci, Mirto ; Toffanin, Chiara: Therapy-driven Deep Glucose Forecasting. In: *Engineering Applications of Artificial Intelligence* 87 (2020), Januar, 103255. http://dx.doi.org/10.1016/j.engappai.2019.103255. – DOI 10.1016/j.engappai.2019.103255. – ISSN 0952–1976

[48] Siami-Namini, Sima ; Tavakoli, Neda ; Namin, Akbar S.: The Performance of LSTM and BiLSTM in Forecasting Time Series. In: *2019 IEEE International Conference on Big Data (Big Data)* (2019), 3285-3292. https://api.semanticscholar.org/CorpusID:211297310

[49] Dave, Darpit ; DeSalvo, Daniel J. ; Haridas, Balakrishna ; McKay, Siripoom ; Shenoy, Akhil ; Koh, Chester J. ; Lawley, Mark ; Erraguntla, Madhav: Feature-Based Machine Learning Model for Real-Time Hypoglycemia Prediction. In: *Journal of Diabetes Science and Technology* (2020), Juni, 193229682092262. http://dx.doi.org/10.1177/1932296820922622. – DOI 10.1177/1932296820922622. – ISSN 1932–2968

[50] Bremer, T ; Gough, D A.: Is blood glucose predictable from previous values? A solicitation for data. In: *Diabetes* 48 (1999), März, Nr. 3, 445–451. http://dx.doi.org/10.2337/diabetes.48.3.445. – DOI 10.2337/diabetes.48.3.445. – ISSN 1939–327X

[51] Tresp, V. ; Briegel, T. ; Moody, J.: Neural-network models for the blood glucose metabolism of a diabetic. In: *IEEE Transactions on Neural Networks* 10 (1999), Nr. 5, S. 1204–1213. http://dx.doi.org/10.1109/72.788659. – DOI 10.1109/72.788659

[52] Georga, E. I. ; Protopappas, V. C. ; Ardigo, Diego ; Marina, M. ; Zavaroni, I. ; Polyzos, D. ; Fotiadis, D. I.: Multivariate Prediction of Subcutaneous Glucose Concentration in Type 1 Diabetes Patients Based on Support Vector Regression. In: *IEEE Journal of Biomedical and Health Informatics* 17 (2013), Januar, Nr. 1, 71–81. http://dx.doi.org/10.1109/titb.2012.2219876. – DOI 10.1109/titb.2012.2219876. – ISSN 2168–2208

[53] Zarkogianni, K. ; Mitsis, K. ; Litsa, E. ; Arredondo, M.-T. ; Fico, G. ; Fioravanti, A. ; Nikita, K. S.: Comparative assessment of glucose prediction models for patients with type 1 diabetes mellitus applying sensors for glucose and physical activity monitoring. In: *Medical & Biological Engineering & Computing* 53 (2015), Juni, Nr. 12, 1333–1343. http://dx.doi.org/10.1007/s11517-015-1320-9. – DOI 10.1007/s11517–015–1320–9. – ISSN 1741–0444

[54] Munoz-Organero, Mario: Deep Physiological Model for Blood Glucose Prediction in T1DM Patients. In: *Sensors* 20 (2020), Juli, Nr. 14, 3896. http://dx.doi.org/10.3390/s20143896. – DOI 10.3390/s20143896. – ISSN 1424–8220

[55] Seo, Wonju ; Park, Sung-Woon ; Kim, Namho ; Jin, Sang-Man ; Park, Sung-Min: A personalized blood glucose level prediction model with a fine-tuning strategy: A proof-of-concept study. In: *Computer Methods and Programs in Biomedicine* 211 (2021), November, 106424. http://dx.doi.org/10.1016/j.cmpb.2021.106424. – DOI 10.1016/j.cmpb.2021.106424. – ISSN 0169–2607

[56] Nemat, Hoda ; Khadem, Heydar ; Eissa, Mohammad R. ; Elliott, Jackie ; Benaissa, Mohammed: Blood Glucose Level Prediction: Advanced Deep-Ensemble Learning Approach. In: *IEEE Journal of Biomedical and Health Informatics* 26 (2022), Juni, Nr. 6, 2758–2769. http://dx.doi.org/10.1109/jbhi.2022.3144870. – DOI 10.1109/jbhi.2022.3144870. – ISSN 2168–2208

[57] Jaloli, Mehrad ; Cescon, Marzia: Long-Term Prediction of Blood Glucose Levels in Type 1 Diabetes Using a CNN-LSTM-Based Deep Neural Network. In: *Journal of Diabetes Science and Technology* 17 (2022), April, Nr. 6, 1590–1601. http:

//dx.doi.org/10.1177/19322968221092785. – DOI 10.1177/19322968221092785. – ISSN 1932–2968

[58] In: Phadke, Rekha ; Nagaraj, H. C.: *Multivariate Long-Term Forecasting of T1DM: A Hybrid Econometric Model-Based Approach*. Springer Nature Singapore, 2022. – ISBN 9789811954825, 1013–1035

[59] Zhu, Taiyu ; Li, Kezhi ; Herrero, Pau ; Georgiou, Pantelis: Personalized Blood Glucose Prediction for Type 1 Diabetes Using Evidential Deep Learning and Meta-Learning. In: *IEEE Transactions on Biomedical Engineering* 70 (2023), Januar, Nr. 1, 193–204. http://dx.doi.org/10.1109/tbme.2022.3187703. – DOI 10.1109/tbme.2022.3187703. – ISSN 1558–2531

[60] Zhu, Taiyu ; Uduku, Chukwuma ; Li, Kezhi ; Herrero, Pau ; Oliver, Nick ; Georgiou, Pantelis: Enhancing self-management in type 1 diabetes with wearables and deep learning. In: *npj Digital Medicine* 5 (2022), Juni, Nr. 1. http://dx.doi.org/10.1038/s41746-022-00626-5. – DOI 10.1038/s41746–022–00626–5. – ISSN 2398–6352

[61] Chen, Zixuan ; Nie, Zedong ; Li, Jingzhen ; Wu, Youwen ; Zhou, Hongjun: Multi-parameter Blood Glucose Prediction Algorithm for Type 1 Diabetes Based on Hybrid Neural Network Deep Learning Technique. In: *2023 IEEE 9th International Conference on Cloud Computing and Intelligent Systems (CCIS)*, IEEE, August 2023, ""

[62] In: Swain, Abhijeet ; Ganatra, Vaibhav ; Saha, Snehanshu ; Mathur, Archana ; Phadke, Rekha: *P-LSTM: A Novel LSTM Architecture for Glucose Level Prediction Problem*. Springer Nature Singapore, 2023. – ISBN 9789819916481, 369–380

[63] Oviedo, Silvia ; Contreras, Ivan ; Quirós, Carmen ; Giménez, Marga ; Conget, Ignacio ; Vehi, Josep: Risk-based postprandial hypoglycemia forecasting using supervised learning. In: *International Journal of Medical Informatics* 126 (2019), Juni, 1–8. http://dx.doi.org/10.1016/j.ijmedinf.2019.03.008. – DOI 10.1016/j.ijmedinf.2019.03.008. – ISSN 1386–5056

[64] Bertachi, Arthur ; Viñals, Clara ; Biagi, Lyvia ; Contreras, Ivan ; Vehí, Josep ; Conget, Ignacio ; Giménez, Marga: Prediction of Nocturnal Hypoglycemia in Adults with Type 1 Diabetes under Multiple Daily Injections Using Continuous Glucose Monitoring and Physical Activity Monitor. In: *Sensors* 20 (2020), März,

Nr. 6, 1705. http://dx.doi.org/10.3390/s20061705. – DOI 10.3390/s20061705. – ISSN 1424–8220

[65] Vehí, Josep ; Contreras, Iván ; Oviedo, Silvia ; Biagi, Lyvia ; Bertachi, Arthur: Prediction and prevention of hypoglycaemic events in type-1 diabetic patients using machine learning. In: *Health Informatics Journal* 26 (2019), Juni, Nr. 1, 703–718. http://dx.doi.org/10.1177/1460458219850682. – DOI 10.1177/1460458219850682. – ISSN 1741–2811

[66] Tyler, Nichole S. ; Mosquera-Lopez, Clara ; Young, Gavin M. ; El Youssef, Joseph ; Castle, Jessica R. ; Jacobs, Peter G.: Quantifying the impact of physical activity on future glucose trends using machine learning. In: *iScience* 25 (2022), März, Nr. 3, 103888. http://dx.doi.org/10.1016/j.isci.2022.103888. – DOI 10.1016/j.isci.2022.103888. – ISSN 2589–0042

[67] Alvarado, J. ; Velasco, J. M. ; Chávez, F. ; Hidalgo, J. I. ; Vega, F. F.: *Patterns Detection in Glucose Time Series by Domain Transformations and Deep Learning*. http://dx.doi.org/10.48550/ARXIV.2303.17616. Version: 2023

[68] Felizardo, Virginie ; Garcia, Nuno M. ; Megdiche, Imen ; Pombo, Nuno ; Sousa, Miguel ; Babič, František: Hypoglycaemia Prediction Using Information Fusion and Classifiers Consensus. In: *Eng. Appl. Artif. Intell.* 123 (2023), aug, Nr. PA. http://dx.doi.org/10.1016/j.engappai.2023.106194. – DOI 10.1016/j.engappai.2023.106194. – ISSN 0952–1976

[69] Otten, Neri V.: *Regression Vs Classification — Understand How To Choose And Switch Between Them*. https://spotintelligence.com/2023/05/02/regression-vs-classification/#:~:text=Robustness%20to%20outliers%3A%20Classification%20models,than%20the%20exact%20numerical%20relationship. Version: 2023. – last access: 13.02.2024

[70] Marling, Cindy ; Bunescu, Razvan: The OhioT1DM Dataset for Blood Glucose Level Prediction: Update 2020. In: *CEUR workshop proceedings* 2675 (2020), 09, S. 71–74

[71] Nemat, Hoda ; Khadem, Heydar ; Elliott, Jackie ; Benaissa, Mohammed: Causality analysis in type 1 diabetes mellitus with application to blood glucose level prediction. In: *Computers in Biology and Medicine* 153 (2023),

Februar, 106535. http://dx.doi.org/10.1016/j.compbiomed.2022.106535. – DOI 10.1016/j.compbiomed.2022.106535. – ISSN 0010–4825

[72] Noor, M.N. ; Yahaya, A.S. ; Ramli, N.A. ; Al Bakri, Abdullah Mohd M.: Filling Missing Data Using Interpolation Methods: Study on the Effect of Fitting Distribution. In: *Key Engineering Materials* 594–595 (2013), Dezember, 889–895. http://dx.doi.org/10.4028/www.scientific.net/kem.594-595.889. – DOI 10.4028/www.scientific.net/kem.594–595.889. – ISSN 1662–9795

[73] Mucha, Mateusz ; Czernia, Dominik: *Velocity Calculator*. https://www.omnicalculator.com/physics/velocity. Version: 2024. – last access: 28.02.2024

[74] Sas, Wojciech: *Magnitude of Acceleration Calculator*. https://www.omnicalculator.com/physics/magnitude-of-acceleration. Version: 2024. – last access: 28.02.2024

[75] Janse, Roemer J. ; Hoekstra, Tiny ; Jager, Kitty J. ; Zoccali, Carmine ; Tripepi, Giovanni ; Dekker, Friedo W. ; Diepen, Merel van: Conducting correlation analysis: important limitations and pitfalls. In: *Clinical Kidney Journal* 14 (2021), Mai, Nr. 11, 2332–2337. http://dx.doi.org/10.1093/ckj/sfab085. – DOI 10.1093/ckj/sfab085. – ISSN 2048–8513

[76] Liu, Yaqing ; Mu, Yong ; Chen, Keyu ; Li, Yiming ; Guo, Jinghuan: Daily Activity Feature Selection in Smart Homes Based on Pearson Correlation Coefficient. In: *Neural Processing Letters* 51 (2020), Januar, Nr. 2, 1771–1787. http://dx.doi.org/10.1007/s11063-019-10185-8. – DOI 10.1007/s11063–019–10185–8. – ISSN 1573–773X

[77] Miot, Hélio A.: Correlation analysis in clinical and experimental studies. In: *Jornal Vascular Brasileiro* 17 (2018), 275 - 279. https://api.semanticscholar.org/CorpusID:67771605

[78] Mukaka, Mavuto: Statistics Corner: A guide to appropriate use of Correlation coefficient in medical research. In: *Malawi medical journal : the journal of Medical Association of Malawi* 24 (2012), 09, S. 69–71

[79] Xiao, Chengwei ; Ye, Jiaqi ; Esteves, Rui M. ; Rong, Chunming: Using Spearman's correlation coefficients for exploratory data analysis on big dataset. In: *Concurrency and Computation: Practice and Experience* 28 (2015), Dezember, Nr.

14, 3866–3878. http://dx.doi.org/10.1002/cpe.3745. – DOI 10.1002/cpe.3745. – ISSN 1532–0634

[80] TensorFlow: *Classification on imbalanced data*. https://www.tensorflow.org/tutorials/structured_data/imbalanced_data. Version: 2024. – last access: 05.03.2024

[81] Deng, Yixiang ; Lu, Lu ; Aponte, Laura ; Angelidi, Angeliki M. ; Novak, Vera ; Karniadakis, George E. ; Mantzoros, Christos S.: Deep transfer learning and data augmentation improve glucose levels prediction in type 2 diabetes patients. In: *npj Digital Medicine* 4 (2021), Juli, Nr. 1. http://dx.doi.org/10.1038/s41746-021-00480-x. – DOI 10.1038/s41746–021–00480–x. – ISSN 2398–6352

[82] Cha, Gi-Wook ; Moon, Hyeun ; Kim, Young-Min ; Hong, Won-Hwa ; Hwang, Jung-Ha ; Park, Won-Jun ; Kim, Young-Chan: Development of a Prediction Model for Demolition Waste Generation Using a Random Forest Algorithm Based on Small DataSets. In: *International journal of environmental research and public health* 17 (2020), 09. http://dx.doi.org/10.3390/ijerph17196997. – DOI 10.3390/ijerph17196997

[83] Grandini, Margherita ; Bagli, Enrico ; Visani, Giorgio: Metrics for Multi-Class Classification: an Overview. In: *ArXiv* abs/2008.05756 (2020). https://api.semanticscholar.org/CorpusID:221112671

[84] Laubenbacher, R. ; Niarakis, A. ; Helikar, T. ; An, G. ; Shapiro, B. ; Malik-Sheriff, R. S. ; Sego, T. J. ; Knapp, A. ; Macklin, P. ; Glazier, J. A.: Building digital twins of the human immune system: toward a roadmap. In: *npj Digital Medicine* 5 (2022), Mai, Nr. 1. http://dx.doi.org/10.1038/s41746-022-00610-z. – DOI 10.1038/s41746–022–00610–z. – ISSN 2398–6352

[85] Cho, Sunghyun ; Aiello, Eleonora M. ; Ozaslan, Basak ; Riddell, Michael C. ; Calhoun, Peter ; Gal, Robin L. ; Doyle, Francis J.: Design of a Real-Time Physical Activity Detection and Classification Framework for Individuals With Type 1 Diabetes. In: *Journal of Diabetes Science and Technology* (2023), Februar, 193229682311538. http://dx.doi.org/10.1177/19322968231153896. – DOI 10.1177/19322968231153896. – ISSN 1932–2968

A. Appendix

Table 26: Pearson correlation analysis for each class for each subject (1)

	Correlations	**Classes**									
		0	1	2	3	4	5	6	7	8	9
540	glucose/ basal	0.048	0.035	0.094	0.077	0.116	0.264	0.313	0.386	0.451	**0.475**
	glucose/ bolus	0.056	0.011	-0.055	-0.001	0.041	**-0.060**	0.030	0.047	0.018	0.020
	glucose/ macc	-0.226	0.016	-0.010	0.006	0.011	-0.170	-0.022	-0.075	-0.147	**-0.287**
	basal/ macc	0.006	**-0.198**	-0.044	-0.040	-0.078	-0.012	0.006	-0.031	-0.089	-0.151
	bolus/ macc	-0.033	0.105	0.120	-0.100	0.028	0.032	0.073	0.040	0.039	0.077
	basal/ bolus	-0.019	**-0.109**	-0.020	0.047	-0.035	0.103	-0.014	-0.030	0.038	-0.018
544	glucose/ basal	-0.110	-0.236	-0.430	**-0.611**	-0.361	-0.365	0.203	0.211	0.057	0.085
	glucose/ bolus	0.109	-0.124	**0.310**	NaN	NaN	-0.035	-0.075	-0.099	-0.097	-0.070
	glucose/ macc	0.210	**-0.222**	-0.173	-0.055	0.043	0.010	-0.193	-0.032	0.130	-0.143
	basal/ macc	-0.045	-0.181	-0.191	-0.275	-0.381	-0.528	-0.536	**-0.603**	-0.321	-0.216
	bolus/ macc	0.097	-0.100	-0.057	NaN	NaN	**0.280**	0.237	-0.034	-0.041	0.001
	basal/ bolus	-0.187	0.034	-0.083	NaN	NaN	**-0.274**	-0.143	-0.124	-0.035	-0.029
552	glucose/ basal	0.031	0.027	-0.044	0.055	0.130	**0.241**	0.034	0.001	0.112	0.127
	glucose/ bolus	-0.015	NaN	NaN	NaN	0.038	0.012	-0.018	0.020	0.025	**0.042**
	glucose/ macc	-0.011	-0.116	-0.154	-0.207	0.036	0.068	-0.075	**-0.315**	-0.094	-0.143
	basal/ macc	-0.277	-0.232	-0.147	-0.004	**0.358**	-0.117	-0.134	0.032	0.094	-0.057
	bolus/ macc	-0.003	NaN	NaN	NaN	**-0.150**	0.022	-0.073	-0.019	-0.022	-0.035
	basal/ bolus	0.053	NaN	NaN	NaN	**-0.165**	0.045	0.051	0.030	0.019	0.009

Table 27: Pearson correlation analysis for each class for each subject (2)

	Correlations	**Classes**									
		0	1	2	3	4	5	6	7	8	9
559	glucose/ basal	**0.495**	-0.061	0.120	0.070	0.102	0.293	0.086	0.024	0.064	0.143
	glucose/ bolus	0.042	NaN	NaN	NaN	-0.014	**0.051**	0.046	0.026	0.048	0.042
	glucose/ macc	-0.024	-0.018	**0.135**	0.092	-0.041	-0.018	0.012	0.029	0.003	-0.043
	basal/ macc	-0.057	**-0.085**	0.009	-0.056	0.015	-0.033	-0.031	-0.046	-0.026	-0.028
	bolus/ macc	-0.005	NaN	NaN	NaN	-0.031	**0.071**	0.003	0.018	0.002	-0.004
	basal/ bolus	**0.072**	NaN	NaN	NaN	-0.043	0.020	0.024	-0.024	-0.001	-0.012
563	glucose/ basal	-0.067	0.085	0.084	0.036	-0.016	0.096	0.274	**0.299**	0.082	0.032
	glucose/ bolus	-0.011	-0.019	NaN	-0.013	0.084	0.077	0.079	**0.103**	0.071	0.072
	glucose/ macc	-0.053	-0.124	**0.174**	0.005	-0.036	-0.028	-0.056	-0.099	-0.062	-0.019
	basal/ macc	0.005	0.097	0.110	**0.170**	-0.025	-0.076	0.003	-0.014	0.024	-0.028
	bolus/ macc	0.046	-0.034	NaN	**0.092**	-0.018	-0.063	-0.058	-0.003	-0.052	-0.026
	basal/ bolus	-0.024	-0.034	NaN	**0.327**	-0.004	-0.030	0.032	-0.008	-0.025	-0.011
567	glucose/ basal	-0.234	0.097	0.045	0.117	**0.362**	0.074	-0.293	0.002	0.180	-0.172
	glucose/ bolus	0.029	NaN	NaN	**0.155**	-0.058	0.052	0.096	0.123	0.069	0.078
	glucose/ macc	-0.098	-0.020	-0.127	-0.023	0.048	0.087	-0.145	0.067	**0.150**	0.015
	basal/ macc	-0.022	-0.036	-0.202	0.164	-0.175	-0.125	**0.188**	-0.178	0.006	0.022
	bolus/ macc	0.008	NaN	NaN	-0.070	**-0.072**	0.030	0.00	0.024	-0.029	-0.019
	basal/ bolus	0.006	NaN	NaN	-0.079	**-0.122**	0.018	0.008	0.023	0.004	-0.021

Table 28: Pearson correlation analysis for each class for each subject (3)

	Correlations	**Classes**									
		0	1	2	3	4	5	6	7	8	9
570	glucose/ basal	-0.143	0.311	**0.547**	0.151	-0.145	0.327	0.402	0.336	-0.022	0.236
	glucose/ bolus	0.085	-0.087	-0.039	0.067	0.067	0.016	-0.035	**0.140**	-0.045	0.039
	glucose/ macc	0.079	-0.186	**0.247**	0.078	0.051	0.066	0.107	-0.105	0.070	-0.081
	basal/ macc	0.006	**-0.124**	0.064	-0.037	0.005	0.008	0.037	-0.083	-0.063	0.00
	bolus/ macc	-0.045	**0.224**	-0.105	0.006	-0.071	0.006	0.017	-0.120	-0.015	-0.010
	basal/ bolus	0.040	-0.034	-0.015	-0.033	-0.056	0.369	0.055	**0.184**	-0.014	0.047
575	glucose/ basal	0.116	0.025	0.117	0.097	0.036	0.008	**-0.168**	-0.273	0.029	-0.144
	glucose/ bolus	-0.019	-0.010	NaN	-0.041	-0.027	0.021	0.002	0.014	**0.065**	0.043
	glucose/ macc	-0.019	0.019	**-0.068**	0.022	0.042	0.021	0.045	0.002	-0.002	-0.003
	basal/ macc	0.013	0.012	-0.003	0.005	0.022	0.031	-0.069	-0.127	-0.004	**-0.164**
	bolus/ macc	0.014	0.011	NaN	-0.022	-0.037	**0.047**	-0.026	0.023	-0.006	-0.005
	basal/ bolus	0.001	-0.024	NaN	0.021	0.033	0.026	**-0.040**	-0.035	-0.030	-0.038
584	glucose/ basal	0.528	-0.106	-0.216	-0.359	-0.059	**-0.628**	0.128	0.187	-0.156	0.105
	glucose/ bolus	NaN	NaN	NaN	**0.227**	-0.008	0.122	0.110	0.090	0.031	0.039
	glucose/ macc	-0.014	0.025	**0.330**	0.242	0.126	0.266	0.169	-0.041	-0.013	-0.113
	basal/ macc	-0.175	-0.097	-0.127	-0.166	-0.116	**-0.277**	-0.079	0.073	-0.060	0.008
	bolus/ macc	NaN	NaN	NaN	0.047	0.007	**0.085**	0.060	0.003	-0.026	-0.044
	basal/ bolus	NaN	NaN	NaN	**0.046**	0.026	-0.049	0.008	0.005	-0.005	0.017

Table 29: Pearson correlation analysis for each class for each subject (4)

	Correlations	**Classes**									
		0	1	2	3	4	5	6	7	8	9
588	glucose/ basal	**0.287**	-0.153	0.116	0.205	0.204	-0.193	0.041	0.052	0.085	-0.010
	glucose/ bolus	0.011	-0.093	NaN	**0.188**	-0.082	-0.027	-0.005	0.048	-0.014	-0.023
	glucose/ macc	0.056	-0.025	-0.025	0.053	-0.021	**0.066**	0.056	0.00	-0.036	0.019
	basal/ macc	-0.123	-0212	-0.105	-0.252	-0.317	-0.581	-0.258	-0.670	**-0.752**	-0.048
	bolus/ macc	-0.007	-0.013	NaN	-0.023	**-0.043**	-0.010	0.013	-0.014	0.004	-0.003
	basal/ bolus	**-0.078**	0.030	NaN	0.005	0.042	-0.006	-0.002	0.034	0.005	0.018
591	glucose/ basal	**0.201**	0.066	0.045	0.063	0.077	0.138	0.048	0.007	0.008	-0.074
	glucose/ bolus	0.005	-0.023	0.017	0.023	**0.067**	0.001	0.055	0.039	0.020	0.035
	glucose/ macc	**0.140**	-0.091	-0.120	-0.078	-0.071	-0.125	-0.030	-0.017	-0.061	-0.007
	basal/ macc	-0.113	-0.190	-0.296	-0.305	**-0.349**	-0.369	-0.031	-0.304	-0.250	-0.259
	bolus/ macc	0.027	0.027	**-0.049**	-0.025	-0.037	0.036	0.001	-0.020	-0.029	0.008
	basal/ bolus	**0.050**	0.024	0.029	0.031	0.024	-0.004	-0.036	0.015	0.003	-0.024
596	glucose/ basal	-0.094	0.151	0.119	0.137	0.154	0.026	-0.130	0.079	-0.086	**0.172**
	glucose/ bolus	0.040	**-0.083**	NaN	-0.048	0.027	-0.032	-0.019	0.004	-0.008	0.00
	glucose/ macc	-0.064	0.059	-0.033	**-0.257**	-0.044	0.201	0.063	0.080	0.047	-0.086
	basal/ macc	0.087	-0.011	0.003	0.014	0.062	0.119	0.053	-0.004	**0.189**	0.012
	bolus/ macc	0.052	**-0.115**	NaN	-0.169	-0.012	-0.087	0.00	0.028	0.003	-0.029
	basal/ bolus	0.052	-0.030	NaN	-0.014	-0.027	0.021	-0.024	**0.053**	0.026	0.002

Table 30: Population-based ResNet results for each subject using 9 classes

Subject	**Metric**	**Class**								
		0	1	2	3	4	5	6	7	8
540	Precision	0.73	0.32	**0.19**	0.13	0.15	0.07	0.29	0.16	0.32
	Recall	0.96	0.53	0.21	0.08	0.04	0.00	0.11	**0.70**	0.14
	F1-M	0.83	0.40	0.20	0.10	0.07	0.01	0.16	0.26	0.20
544	Precision	0.97	0.32	0.18	0.11	0.10	0.13	0.18	0.19	0.41
	Recall	0.80	0.51	**0.67**	**0.35**	0.33	0.11	0.29	0.12	0.21
	F1-M	0.88	0.39	**0.28**	0.16	0.15	0.12	0.22	0.15	0.28
552	Precision	0.98	0.37	**0.19**	**0.16**	0.13	0.21	0.23	0.19	0.45
	Recall	0.89	0.66	0.51	0.20	0.14	**0.24**	**0.51**	0.14	0.07
	F1-M	0.93	0.48	**0.28**	0.18	0.14	**0.22**	0.31	0.16	0.12
559	Precision	0.99	0.32	0.10	0.09	0.11	0.11	0.27	0.18	0.50
	Recall	0.95	**0.82**	0.47	0.17	0.11	0.07	0.15	0.42	0.18
	F1-M	**0.97**	0.46	0.17	0.12	0.11	0.09	0.10	0.25	0.26
563	Precision	**1.00**	0.30	**0.19**	**0.16**	0.16	0.25	0.28	0.17	0.45
	Recall	0.80	0.74	0.38	0.27	0.15	0.08	0.32	0.05	0.59
	F1-M	0.89	0.42	0.25	**0.20**	0.16	0.12	**0.30**	0.08	0.51
567	Precision	0.92	**0.53**	**0.16**	0.16	0.15	0.09	0.22	0.17	0.43
	Recall	**0.98**	0.28	0.09	0.09	0.06	0.04	0.50	0.13	0.30
	F1-M	0.95	0.36	0.12	0.12	0.09	0.06	**0.30**	0.15	0.35
570	Precision	0.97	0.39	0.15	0.05	0.05	0.05	0.14	0.15	0.53
	Recall	0.88	0.39	**0.19**	0.04	0.04	0.05	0.07	0.45	0.25
	F1-M	0.92	0.39	0.17	0.04	0.04	0.05	0.09	0.23	0.34
575	Precision	0.93	0.39	0.16	**0.16**	**0.24**	0.23	0.25	**0.22**	**1.00**
	Recall	0.93	0.57	0.41	0.15	0.15	0.13	0.34	0.58	0.01
	F1-M	0.95	0.47	0.23	0.15	0.19	0.16	0.29	**0.32**	0.01
584	Precision	0.97	0.39	**0.19**	0.10	0.06	0.12	0.14	0.16	0.78
	Recall	**0.98**	0.49	0.46	0.20	0.04	0.10	0.17	0.47	0.09
	F1-M	**0.97**	0.43	0.27	0.14	0.04	0.11	0.15	0.24	0.17
588	Precision	0.96	0.20	0.06	0.10	0.15	**0.28**	**0.30**	0.16	0.50
	Recall	0.92	0.76	0.24	0.25	0.18	0.18	0.05	0.10	**0.63**
	F1-M	0.94	0.32	0.09	0.14	0.17	**0.22**	0.08	0.12	**0.56**
591	Precision	0.97	0.47	**0.16**	0.13	0.09	0.11	0.26	0.17	0.22
	Recall	0.97	0.64	0.41	0.31	0.17	0.19	0.22	0.17	0.08
	F1-M	**0.97**	**0.54**	0.23	0.18	0.12	0.14	0.24	0.17	0.12
596	Precision	0.99	0.40	0.18	0.14	0.14	0.10	0.24	0.15	0.40
	Recall	0.90	0.74	0.49	0.31	**0.35**	0.12	0.36	0.08	0.22
	F1-M	0.94	0.52	0.26	0.19	**0.20**	0.11	0.28	0.10	0.29

Table 31: Population-based hybrid model results for each subject using 9 classes

Subject	**Metric**	**Class**								
		0	1	2	3	4	5	6	7	8
540	Precision	**1.00**	0.17	0.16	0.15	0.18	0.19	0.22	0.20	0.36
	Recall	0.63	0.28	0.35	0.34	0.31	0.16	0.11	0.22	0.32
	F1-M	0.77	0.21	0.22	0.21	0.23	0.18	0.14	0.21	0.34
544	Precision	0.97	0.36	0.17	0.09	0.05	0.06	0.24	0.29	0.44
	Recall	0.85	0.56	0.39	0.30	0.15	0.13	**0.24**	0.43	0.10
	F1-M	0.90	0.44	0.24	0.14	0.07	0.08	0.24	**0.35**	0.17
552	Precision	**1.00**	0.22	0.20	**0.22**	0.12	0.17	0.32	0.22	0.30
	Recall	0.70	0.32	0.43	0.19	0.18	**0.38**	0.07	0.43	0.13
	F1-M	0.82	0.26	0.27	0.20	0.14	**0.24**	0.11	0.29	0.18
559	Precision	0.99	0.27	0.13	0.11	0.12	0.18	0.45	0.22	0.41
	Recall	0.79	0.47	**0.53**	0.20	0.21	0.24	0.00	**0.67**	0.01
	F1-M	0.88	0.35	0.21	0.14	0.15	0.20	0.01	0.33	0.03
563	Precision	**1.00**	0.23	0.18	0.15	**0.19**	0.15	0.16	0.13	0.43
	Recall	0.48	**0.64**	0.37	0.35	0.25	0.05	0.07	0.01	0.71
	F1-M	0.65	0.34	0.24	0.21	**0.21**	0.07	0.10	0.02	0.53
567	Precision	0.97	**0.41**	**0.26**	0.21	0.12	0.17	**0.36**	**0.60**	0.48
	Recall	0.93	0.53	0.40	**0.43**	**0.46**	0.31	0.05	0.00	0.53
	F1-M	**0.95**	**0.46**	**0.32**	**0.28**	0.20	0.22	0.09	0.01	0.51
570	Precision	0.95	0.31	0.15	0.21	0.16	0.07	0.00	0.09	**0.50**
	Recall	0.91	0.56	0.41	0.30	0.14	0.05	0.00	0.01	**0.84**
	F1-M	0.93	0.40	0.22	0.25	0.15	0.06	0.00	0.02	**0.63**
575	Precision	0.88	0.25	0.14	0.15	0.18	**0.21**	0.30	0.19	0.35
	Recall	0.98	0.51	0.36	0.26	0.26	0.27	0.06	0.47	0.07
	F1-M	0.93	0.33	0.20	0.19	0.21	0.23	0.10	0.27	0.11
584	Precision	0.99	0.26	0.18	0.09	0.10	0.12	0.14	0.20	0.40
	Recall	0.78	0.33	0.42	0.18	0.22	0.21	0.07	0.32	0.14
	F1-M	0.87	0.29	0.26	0.12	0.13	0.15	0.10	0.25	0.21
588	Precision	**1.00**	0.25	0.11	0.09	0.13	0.19	0.15	0.19	0.40
	Recall	0.72	0.37	0.32	0.28	0.26	0.18	0.21	0.10	0.19
	F1-M	0.83	0.30	0.16	0.13	0.17	0.18	0.18	0.13	0.26
591	Precision	0.97	0.37	0.16	0.14	0.11	0.12	0.28	0.25	0.42
	Recall	0.89	0.61	0.40	0.38	0.24	0.19	0.22	0.36	0.05
	F1-M	0.93	**0.46**	0.22	0.20	0.15	0.15	**0.25**	0.30	0.08
596	Precision	0.83	0.39	0.15	0.14	0.10	0.14	0.10	0.24	0.36
	Recall	**1.00**	0.51	0.36	0.32	0.18	0.22	0.01	0.46	0.22
	F1-M	0.91	0.44	0.22	0.19	0.13	0.17	0.01	0.32	0.27

Table 32: Subject-specific ResNet results for each subject using 9 classes

Subject	**Metric**	**Class**								
		0	1	2	3	4	5	6	7	8
540	Precision	**1.00**	0.23	0.19	0.14	0.20	0.24	0.35	0.15	0.47
	Recall	0.73	0.67	0.54	0.32	0.40	0.12	0.24	0.05	**0.41**
	F1-M	0.84	0.35	0.28	0.20	0.26	0.16	0.28	0.08	**0.44**
544	Precision	0.86	0.25	0.16	0.07	0.15	0.00	0.22	0.18	0.43
	Recall	**1.00**	**0.81**	**0.69**	0.36	**0.53**	0.00	0.39	0.24	0.11
	F1-M	0.93	0.38	0.26	0.11	0.24	0.00	0.28	0.21	0.18
552	Precision	**1.00**	0.38	0.14	0.13	0.05	0.19	0.22	0.24	**0.54**
	Recall	0.81	0.48	0.67	**0.40**	0.15	0.30	0.30	0.12	0.23
	F1-M	0.89	0.42	0.23	0.19	0.08	0.23	0.26	0.16	0.33
559	Precision	**1.00**	0.33	0.13	0.12	0.06	0.07	0.38	0.24	0.48
	Recall	0.96	0.77	0.53	0.25	0.11	0.12	0.14	0.28	0.26
	F1-M	**0.98**	0.46	0.21	0.16	0.08	0.09	0.21	0.26	0.34
563	Precision	0.78	0.15	0.06	0.10	0.12	0.11	0.21	**0.34**	0.32
	Recall	**1.00**	0.45	0.28	0.31	0.23	0.07	0.18	0.20	0.16
	F1-M	0.87	0.22	0.10	0.15	0.16	0.08	0.20	0.25	0.22
567	Precision	0.94	0.35	**0.25**	**0.20**	0.16	0.18	0.45	0.00	0.49
	Recall	0.97	0.77	0.43	0.38	0.37	0.40	**0.46**	0.00	0.34
	F1-M	0.95	0.48	**0.32**	**0.26**	0.22	0.24	**0.45**	0.00	0.40
570	Precision	0.79	0.33	0.12	0.13	0.09	0.07	0.04	0.23	0.59
	Recall	0.95	0.54	0.29	0.25	0.10	0.09	0.03	**0.35**	0.18
	F1-M	0.86	0.41	0.17	0.17	0.10	0.08	0.03	**0.28**	0.27
575	Precision	0.99	0.21	0.16	0.17	**0.23**	0.25	0.28	0.16	0.48
	Recall	0.92	0.73	0.44	**0.40**	0.38	**0.45**	0.28	0.16	0.20
	F1-M	0.95	0.32	0.24	0.24	**0.29**	**0.32**	0.28	0.16	0.28
584	Precision	0.79	**0.75**	0.09	0.09	0.13	0.09	0.12	0.12	0.87
	Recall	**1.00**	0.43	0.33	0.28	0.11	0.12	0.18	0.31	0.12
	F1-M	0.88	**0.55**	0.14	0.14	0.12	0.10	0.15	0.17	0.22
588	Precision	0.88	0.21	0.09	0.10	0.09	0.24	0.40	0.14	0.24
	Recall	**1.00**	0.68	0.25	0.15	0.10	0.26	0.19	0.06	0.29
	F1-M	0.94	0.32	0.13	0.12	0.09	0.25	0.25	0.08	0.26
591	Precision	**1.00**	0.49	0.14	0.15	0.08	0.10	0.24	0.29	0.25
	Recall	0.95	0.54	0.20	0.21	0.11	0.12	0.24	0.31	0.11
	F1-M	0.95	0.54	0.20	0.21	0.11	0.12	0.24	0.21	0.11
596	Precision	0.88	0.38	0.12	0.11	0.12	**0.29**	**0.46**	0.18	0.41
	Recall	**1.00**	0.76	0.34	0.27	0.36	0.26	0.34	0.24	0.09
	F1-M	0.94	0.50	0.17	0.15	0.18	0.27	0.39	0.20	0.15

Table 33: Population-based ResNet results with less test data for each subject using 9 classes

Subject	**Metric**	**Class**								
		0	1	2	3	4	5	6	7	8
540	Precision	0.85	0.28	**0.26**	0.13	0.11	0.00	0.00	0.14	0.41
	Recall	**0.96**	0.65	0.28	0.08	0.04	0.00	0.00	**0.76**	0.17
	F1-M	0.90	0.39	0.27	0.10	0.06	0.00	0.00	0.23	0.24
544	Precision	0.98	0.25	0.20	0.19	0.06	0.10	0.00	0.03	0.43
	Recall	0.82	0.44	**0.77**	**0.68**	0.28	**0.28**	0.09	0.00	0.25
	F1-M	0.89	0.32	**0.31**	**0.29**	0.10	0.09	0.00	0.03	0.31
552	Precision	**1.00**	0.34	0.14	0.10	0.04	0.06	0.18	**0.28**	0.55
	Recall	0.82	0.48	0.50	0.21	0.06	0.07	0.39	0.20	0.15
	F1-M	0.90	0.40	0.22	0.13	0.04	0.06	0.25	0.23	0.23
559	Precision	**1.00**	0.32	0.12	0.08	0.06	0.06	0.33	0.15	0.42
	Recall	0.95	**0.79**	0.50	0.14	0.06	0.05	0.21	0.31	0.17
	F1-M	0.97	0.46	0.20	0.10	0.06	0.05	0.26	0.20	0.25
563	Precision	**1.00**	0.28	0.16	0.18	**0.17**	0.00	0.14	0.01	0.30
	Recall	0.72	0.78	0.39	0.30	0.16	0.00	0.18	0.00	0.32
	F1-M	0.84	0.41	0.22	0.22	0.17	0.00	0.16	0.01	0.31
567	Precision	0.92	0.29	0.14	**0.21**	0.14	0.14	0.18	0.16	0.52
	Recall	**0.96**	0.19	0.10	0.07	0.05	0.06	0.39	0.19	0.29
	F1-M	0.94	0.23	0.12	0.10	0.07	0.08	0.24	0.17	**0.37**
570	Precision	0.97	0.41	0.15	0.00	0.06	0.18	0.23	0.15	0.39
	Recall	0.76	0.38	0.13	0.00	0.05	0.11	0.18	0.38	0.19
	F1-M	0.85	0.39	0.14	0.00	0.05	0.14	0.20	0.21	0.26
575	Precision	0.98	0.36	0.12	0.10	0.11	0.03	0.14	**0.28**	**1.00**
	Recall	0.92	0.58	0.39	0.10	0.10	0.02	0.18	0.67	0.01
	F1-M	0.95	0.44	0.18	0.10	0.10	0.02	0.16	**0.39**	0.01
584	Precision	**1.00**	**0.83**	0.14	0.09	0.07	0.05	0.08	0.16	**0.69**
	Recall	**0.96**	0.71	0.44	0.33	0.06	0.05	0.12	0.44	0.10
	F1-M	**0.98**	**0.77**	0.22	0.14	0.06	0.05	0.09	0.23	0.17
588	Precision	0.93	0.17	0.07	0.09	0.11	**0.32**	0.00	0.15	0.31
	Recall	0.94	0.68	0.21	0.15	0.09	0.26	0.00	0.13	**0.45**
	F1-M	0.93	0.27	0.10	0.11	0.10	**0.29**	0.00	0.14	**0.37**
591	Precision	**1.00**	0.51	0.13	0.10	0.08	0.11	0.23	0.22	0.11
	Recall	0.95	0.64	0.31	0.25	0.15	0.16	0.18	0.21	0.05
	F1-M	0.97	0.57	0.18	0.14	0.10	0.13	0.20	0.21	0.06
596	Precision	0.99	0.51	0.20	0.13	**0.17**	0.19	**0.35**	0.13	0.34
	Recall	0.89	0.76	0.45	0.27	**0.36**	0.18	**0.44**	0.07	0.22
	F1-M	0.94	0.61	0.28	0.18	**0.23**	0.18	**0.39**	0.09	0.26

Table 34: Subject-specific hybrid model results for each subject using 9 classes

Subject	**Metric**	**Class**								
		0	1	2	3	4	5	6	7	8
540	Precision	0.99	0.26	0.15	**0.15**	0.15	0.28	0.27	0.17	0.40
	Recall	0.91	0.75	0.42	0.40	0.23	0.19	0.09	0.23	0.28
	F1-M	**0.95**	0.38	0.23	0.22	0.18	0.23	0.13	0.19	0.33
544	Precision	0.82	0.21	**0.18**	**0.15**	0.15	0.28	0.24	0.16	**0.59**
	Recall	**1.00**	0.62	0.54	**0.55**	**0.58**	**0.74**	0.14	0.34	0.08
	F1-M	0.90	0.32	0.27	0.23	**0.24**	**0.40**	0.17	0.22	0.14
552	Precision	**1.00**	0.30	0.15	0.10	0.07	0.15	0.16	0.27	0.37
	Recall	0.82	0.52	**0.67**	0.35	0.20	0.42	0.07	0.27	0.09
	F1-M	0.90	0.38	**0.24**	0.16	0.10	0.22	0.09	0.27	0.14
559	Precision	0.95	0.25	0.13	0.10	0.11	0.19	**0.91**	0.25	0.29
	Recall	0.91	0.60	0.40	0.22	0.21	0.16	0.02	**0.72**	0.07
	F1-M	0.93	0.35	0.20	0.14	0.14	0.18	0.04	**0.37**	0.11
563	Precision	0.86	0.21	0.06	0.10	0.12	0.23	0.12	0.18	0.32
	Recall	0.95	0.70	0.36	0.36	0.15	0.18	0.07	0.11	0.19
	F1-M	0.90	0.33	0.11	0.16	0.14	0.20	0.09	0.14	0.24
567	Precision	0.98	0.20	0.17	0.12	0.10	0.26	0.24	0.00	0.41
	Recall	0.85	0.68	0.27	0.27	0.33	0.42	0.07	0.00	0.48
	F1-M	0.91	0.31	0.21	0.16	0.15	0.32	0.10	0.00	**0.45**
570	Precision	0.82	0.23	0.12	0.13	0.16	**0.39**	0.05	0.01	0.36
	Recall	0.86	0.38	0.29	0.33	**0.26**	0.15	0.01	0.01	**0.55**
	F1-M	0.84	0.29	0.17	0.29	0.20	0.21	0.01	0.01	0.44
575	Precision	**1.00**	0.27	0.11	0.12	**0.17**	0.16	0.13	0.24	0.29
	Recall	0.87	0.60	0.44	0.32	0.40	0.32	0.02	0.45	0.05
	F1-M	0.93	0.37	0.17	0.18	**0.24**	0.21	0.04	0.31	0.09
584	Precision	0.83	**0.50**	0.09	0.09	0.07	0.08	0.23	0.08	0.00
	Recall	0.83	0.36	0.22	0.22	0.19	0.11	**0.49**	0.18	0.00
	F1-M	0.83	0.42	0.13	0.13	0.10	0.09	**0.32**	0.11	0.00
588	Precision	0.90	0.20	0.11	0.11	0.08	0.23	0.08	**0.48**	0.29
	Recall	0.85	0.44	0.29	0.27	0.13	0.15	0.00	0.27	0.44
	F1-M	0.90	0.28	0.16	0.16	0.10	0.18	0.00	0.35	0.35
591	Precision	0.99	0.36	0.11	0.08	0.07	0.12	0.31	0.25	0.37
	Recall	0.85	**0.79**	0.31	0.26	0.14	0.08	0.03	0.44	0.23
	F1-M	0.92	**0.52**	0.17	0.12	0.09	0.10	0.06	0.32	0.28
596	Precision	0.67	0.37	0.08	0.04	0.07	0.13	0.33	0.29	0.30
	Recall	**1.00**	0.33	0.24	0.08	0.13	0.18	0.17	0.41	0.14
	F1-M	0.80	0.35	0.12	0.05	0.09	0.15	0.23	0.34	0.19

Table 35: Population-based hybrid model results with less test data for each subject using 9 classes

Subject	**Metric**	**Class**								
		0	1	2	3	4	5	6	7	8
540	Precision	**1.00**	0.18	0.14	0.15	**0.19**	0.24	0.24	0.18	0.36
	Recall	0.68	0.35	0.35	0.33	0.30	0.16	0.04	0.27	0.36
	F1-M	0.81	0.24	0.20	0.21	**0.23**	0.19	0.07	0.21	0.36
544	Precision	0.96	0.27	0.13	0.14	0.06	0.08	0.16	0.18	**0.65**
	Recall	0.86	0.44	0.31	0.36	0.31	0.22	0.16	0.32	0.10
	F1-M	0.91	0.33	0.19	0.20	0.10	0.12	0.16	0.23	0.18
552	Precision	**1.00**	0.32	0.19	0.07	0.07	0.13	0.28	0.23	0.21
	Recall	0.77	0.32	**0.54**	0.09	0.15	**0.38**	0.07	0.33	0.09
	F1-M	0.87	0.32	0.28	0.08	0.10	0.20	0.11	0.27	0.12
559	Precision	0.97	0.25	0.15	0.12	0.09	0.17	**1.00**	0.24	0.30
	Recall	0.83	0.40	0.50	0.22	0.15	0.21	0.01	**0.78**	0.01
	F1-M	0.90	0.31	0.23	0.16	0.11	0.19	0.02	**0.36**	0.02
563	Precision	**1.00**	0.20	0.16	0.15	0.15	0.08	0.10	0.04	0.37
	Recall	0.36	**0.68**	0.47	**0.39**	0.21	0.03	0.08	0.00	0.47
	F1-M	0.53	0.31	0.24	0.21	0.17	0.04	0.09	0.01	0.41
567	Precision	0.99	0.35	0.25	**0.19**	0.09	0.22	0.58	**0.60**	0.43
	Recall	0.94	0.61	0.30	0.32	**0.38**	0.37	0.09	0.01	0.50
	F1-M	**0.96**	0.45	0.27	**0.23**	0.14	**0.28**	0.16	0.02	0.46
570	Precision	0.90	0.31	0.10	0.16	0.15	0.04	0.00	0.18	0.35
	Recall	0.93	0.49	0.23	0.27	0.12	0.03	0.00	0.02	**0.69**
	F1-M	0.91	0.38	0.14	0.20	0.13	0.04	0.00	0.03	**0.47**
575	Precision	0.88	0.16	0.16	0.12	0.15	0.18	0.24	0.23	0.36
	Recall	0.98	0.42	0.56	0.33	0.33	0.28	0.04	0.50	0.06
	F1-M	0.93	0.24	0.25	0.18	0.20	0.22	0.07	0.32	0.10
584	Precision	**1.00**	0.31	**0.33**	0.11	0.11	0.08	0.00	0.14	0.33
	Recall	0.52	0.29	0.33	0.28	0.33	0.23	0.00	0.26	0.09
	F1-M	0.69	0.30	**0.33**	0.16	0.16	0.12	0.00	0.18	0.14
588	Precision	**1.00**	0.13	0.07	0.11	0.11	**0.31**	0.24	0.30	0.29
	Recall	0.66	0.24	0.17	0.29	0.21	0.14	**0.22**	0.24	0.21
	F1-M	0.80	0.17	0.10	0.16	0.15	0.19	**0.23**	0.27	0.25
591	Precision	0.99	0.39	0.14	0.14	0.09	0.12	0.28	0.30	0.19
	Recall	0.84	0.66	0.36	0.38	0.22	0.16	0.19	0.40	0.02
	F1-M	0.91	0.49	0.20	0.20	0.13	0.14	**0.23**	0.34	0.04
596	Precision	0.85	**0.55**	0.16	0.13	0.07	0.16	0.00	0.25	0.25
	Recall	**1.00**	0.57	0.32	0.20	0.12	0.20	0.00	0.40	0.23
	F1-M	0.92	**0.56**	0.22	0.16	0.09	0.18	0.00	0.31	0.24

Table 36: Population-based ResNet results for each subject using 6 classes

Subject	**Metric**	**Class**					
		0	1	2	3	4	5
540	Precision	0.99	0.46	0.27	0.28	0.35	0.70
	Recall	0.92	**0.80**	**0.54**	0.34	0.27	0.55
	F1-M	0.96	0.59	0.36	0.31	0.30	0.62
544	Precision	0.98	0.39	0.31	0.24	0.41	0.74
	Recall	0.86	0.70	0.20	0.28	0.35	0.73
	F1-M	0.92	0.50	0.24	0.26	0.37	0.74
552	Precision	0.98	0.49	0.25	0.22	0.31	0.75
	Recall	0.98	0.78	0.52	0.28	0.24	0.61
	F1-M	0.98	0.60	0.34	0.25	0.27	0.67
559	Precision	0.85	0.40	0.26	0.32	0.35	0.71
	Recall	**1.00**	0.41	0.37	0.33	0.22	0.73
	F1-M	0.92	0.40	0.30	0.32	0.27	0.72
563	Precision	0.90	0.43	**0.42**	**0.35**	0.42	0.73
	Recall	0.98	0.61	0.29	0.43	0.33	0.73
	F1-M	0.94	0.51	0.34	**0.39**	0.37	0.73
567	Precision	0.98	0.61	**0.42**	**0.35**	0.45	0.74
	Recall	**1.00**	0.77	0.30	0.32	0.26	**0.90**
	F1-M	**0.99**	**0.68**	0.35	0.34	0.33	**0.81**
570	Precision	0.81	0.25	0.16	0.30	**0.54**	0.64
	Recall	**1.00**	0.39	0.25	0.20	0.15	0.77
	F1-M	0.90	0.31	0.20	0.24	0.23	0.70
575	Precision	0.94	0.57	0.28	**0.35**	0.49	0.65
	Recall	**1.00**	0.49	0.37	0.32	0.16	0.88
	F1-M	0.97	0.53	0.32	0.33	0.24	0.75
584	Precision	0.96	0.45	0.26	0.24	0.27	0.60
	Recall	0.92	0.37	0.29	0.18	0.18	0.78
	F1-M	0.94	0.40	0.28	0.20	0.22	0.68
588	Precision	0.95	**0.65**	0.23	0.23	0.34	0.59
	Recall	**1.00**	0.56	0.27	0.19	0.16	0.78
	F1-M	0.97	0.60	0.25	0.20	0.21	0.67
591	Precision	**1.00**	0.49	0.20	0.25	0.26	0.60
	Recall	0.90	0.51	0.29	0.21	0.15	0.76
	F1-M	0.95	0.50	0.24	0.23	0.19	0.67
596	Precision	**1.00**	0.45	0.39	0.32	0.41	**0.77**
	Recall	0.84	0.64	0.50	**0.44**	**0.36**	0.69
	F1-M	0.91	0.53	**0.44**	0.37	**0.38**	0.73

Table 37: Population-based hybrid model results for each subject using 6 classes

Subject	**Metric**	**Class**					
		0	1	2	3	4	5
540	Precision	0.98	0.40	0.26	0.29	0.33	0.67
	Recall	0.90	0.78	0.46	0.30	0.22	0.62
	F1-M	0.93	0.53	0.33	0.29	0.27	0.64
544	Precision	0.93	0.37	0.31	0.25	0.38	0.83
	Recall	**0.99**	0.76	0.20	0.41	0.24	0.68
	F1-M	0.96	0.50	0.25	0.31	0.29	0.75
552	Precision	0.98	0.45	0.24	0.26	0.31	0.71
	Recall	**0.99**	0.76	0.42	0.26	0.18	0.72
	F1-M	**0.98**	0.57	0.31	0.26	0.23	0.71
559	Precision	0.99	0.47	0.28	0.28	0.31	0.75
	Recall	0.91	0.66	0.55	0.33	0.29	0.63
	F1-M	0.95	0.55	0.37	0.30	0.30	0.68
563	Precision	0.98	0.43	0.32	**0.38**	**0.46**	0.70
	Recall	0.90	0.68	0.46	0.43	0.23	0.79
	F1-M	0.94	0.53	0.38	0.40	0.31	0.74
567	Precision	**1.00**	**0.56**	**0.40**	0.37	0.39	**0.83**
	Recall	0.97	0.76	**0.65**	**0.48**	0.37	0.70
	F1-M	**0.98**	**0.64**	**0.50**	**0.41**	0.38	**0.76**
570	Precision	0.90	0.34	0.16	0.24	0.39	0.70
	Recall	0.98	0.54	0.27	0.28	0.20	0.65
	F1-M	0.94	0.42	0.20	0.26	0.27	0.67
575	Precision	0.98	0.50	0.24	0.34	0.43	0.68
	Recall	0.97	0.73	0.39	0.17	0.27	**0.82**
	F1-M	0.97	0.59	0.30	0.22	0.33	0.74
584	Precision	**1.00**	0.38	0.15	0.16	0.35	0.60
	Recall	0.94	0.46	0.29	0.21	**0.39**	0.41
	F1-M	0.97	0.42	0.20	0.18	0.37	0.48
588	Precision	0.97	0.50	0.16	0.25	0.33	0.64
	Recall	0.95	0.63	0.20	0.29	0.27	0.63
	F1-M	0.96	0.56	0.18	0.27	0.30	0.63
591	Precision	0.93	0.54	0.30	0.30	0.25	0.66
	Recall	**0.99**	0.55	0.32	0.36	0.20	0.67
	F1-M	0.93	0.54	0.30	0.30	0.25	0.66
596	Precision	**1.00**	0.54	0.28	0.29	0.45	0.76
	Recall	0.96	**0.79**	0.43	0.27	**0.39**	0.70
	F1-M	**0.98**	**0.64**	0.34	0.28	**0.42**	0.73

Table 38: Subject-specific ResNet results for each subject using 6 classes

Subject	**Metric**	**Class**					
		0	1	2	3	4	5
540	Precision	**1.00**	0.39	0.28	0.30	0.40	0.71
	Recall	0.92	**0.91**	0.54	0.34	0.32	0.50
	F1-M	0.96	0.55	0.36	0.32	0.35	0.59
544	Precision	0.94	0.37	0.11	0.19	0.29	0.62
	Recall	0.94	0.68	0.08	0.21	0.08	0.74
	F1-M	0.94	0.48	0.09	0.20	0.13	0.67
552	Precision	**1.00**	0.53	**0.34**	0.28	0.28	0.70
	Recall	0.89	0.71	**0.61**	0.40	0.23	0.60
	F1-M	0.94	0.61	**0.44**	0.33	0.25	0.65
559	Precision	0.85	0.32	0.27	0.29	0.41	0.72
	Recall	**1.00**	0.40	0.50	0.30	**0.36**	0.56
	F1-M	0.92	0.35	0.35	0.29	**0.39**	0.63
563	Precision	0.84	0.33	0.30	**0.44**	0.50	0.77
	Recall	**1.00**	0.45	0.33	**0.59**	0.24	0.79
	F1-M	0.91	0.38	0.32	**0.50**	0.32	0.78
567	Precision	**1.00**	0.29	0.17	0.23	0.42	**0.82**
	Recall	0.80	0.74	0.17	0.32	0.34	0.80
	F1-M	0.89	0.42	0.17	0.26	0.38	**0.81**
570	Precision	0.77	0.40	0.22	0.11	0.28	0.59
	Recall	**1.00**	0.47	0.21	0.07	0.17	0.69
	F1-M	0.87	0.44	0.22	0.09	0.21	0.64
575	Precision	0.99	0.49	**0.34**	0.43	**0.61**	0.70
	Recall	0.91	0.54	0.58	0.39	0.22	**0.91**
	F1-M	0.95	0.52	0.43	0.41	0.32	0.79
584	Precision	**1.00**	**0.79**	0.32	0.27	0.23	0.58
	Recall	**1.00**	0.83	0.47	0.13	0.20	0.66
	F1-M	**1.00**	**0.81**	0.38	0.18	0.21	0.62
588	Precision	0.89	0.26	0.14	0.12	0.27	0.46
	Recall	0.95	0.53	0.29	0.06	0.14	0.50
	F1-M	0.92	0.35	0.19	0.08	0.18	0.48
591	Precision	**1.00**	0.38	0.23	0.38	0.29	0.64
	Recall	0.71	0.41	0.29	0.32	0.16	0.84
	F1-M	0.83	0.40	0.26	0.35	0.20	0.73
596	Precision	0.96	0.38	0.27	0.23	0.29	0.72
	Recall	0.94	0.82	0.43	0.28	0.25	0.52
	F1-M	0.95	0.51	0.33	0.25	0.27	0.61

Table 39: Population-based ResNet results with less test data for each subject using 6 classes

Subject	**Metric**	**Class**					
		0	1	2	3	4	5
540	Precision	**1.00**	0.43	0.29	0.35	0.35	0.67
	Recall	0.92	**0.84**	0.58	0.42	0.27	0.51
	F1-M	0.96	0.57	0.39	0.39	0.30	0.58
544	Precision	0.98	0.36	0.33	0.14	0.40	0.61
	Recall	0.90	0.68	0.23	0.12	0.17	0.74
	F1-M	0.93	0.47	0.27	0.13	0.24	0.67
552	Precision	**1.00**	0.63	0.34	0.28	0.29	0.73
	Recall	0.97	0.83	**0.69**	0.38	0.24	0.57
	F1-M	**0.99**	**0.72**	**0.45**	0.32	0.26	0.64
559	Precision	0.89	0.40	0.29	0.24	0.33	0.67
	Recall	**1.00**	0.44	0.47	0.30	0.27	0.64
	F1-M	0.94	0.42	0.36	0.32	0.29	0.66
563	Precision	0.87	0.46	**0.47**	**0.46**	0.56	**0.75**
	Recall	0.99	0.55	0.33	**0.61**	**0.33**	0.82
	F1-M	0.93	0.50	0.39	**0.52**	**0.42**	**0.78**
567	Precision	0.98	0.45	0.23	0.21	0.40	0.74
	Recall	**1.00**	0.74	0.17	0.20	0.28	0.81
	F1-M	**0.99**	0.56	0.19	0.21	0.33	**0.78**
570	Precision	0.70	0.35	0.21	0.17	0.38	0.51
	Recall	**1.00**	0.33	0.18	0.12	0.16	0.71
	F1-M	0.82	0.34	0.19	0.14	0.22	0.59
575	Precision	0.97	0.65	0.38	**0.49**	**0.50**	0.66
	Recall	**1.00**	0.48	0.53	0.35	0.14	**0.93**
	F1-M	0.98	0.55	0.44	0.41	0.22	0.77
584	Precision	0.84	**0.69**	0.40	0.46	0.23	0.57
	Recall	0.84	0.50	0.27	0.20	0.21	0.88
	F1-M	0.84	0.58	0.32	0.28	0.16	0.69
588	Precision	0.91	0.44	0.08	0.19	0.28	0.49
	Recall	1.00	0.42	0.12	0.17	0.19	0.55
	F1-M	0.95	0.43	0.10	0.18	0.23	0.52
591	Precision	**1.00**	0.51	0.20	0.32	0.26	0.65
	Recall	0.85	0.45	0.25	0.26	0.15	0.85
	F1-M	0.92	0.48	0.23	0.29	0.19	0.73
596	Precision	**1.00**	0.39	0.35	0.36	0.30	0.69
	Recall	0.89	0.64	0.47	0.44	0.31	0.55
	F1-M	0.94	0.48	0.40	0.40	0.31	0.61

Table 40: Subject-specific hybrid model results for each subject using 6 classes

Subject	**Metric**	**Class**					
		0	1	2	3	4	5
540	Precision	**1.00**	0.34	0.29	0.33	0.36	0.67
	Recall	0.80	0.82	**0.59**	0.36	0.28	0.51
	F1-M	0.89	0.48	0.39	0.34	0.31	0.58
544	Precision	0.91	0.36	0.50	0.32	0.43	0.76
	Recall	**1.00**	0.68	**0.46**	0.46	0.12	**0.78**
	F1-M	0.95	0.47	**0.48**	0.38	0.19	**0.77**
552	Precision	0.98	0.50	0.34	0.23	0.27	0.67
	Recall	0.84	0.66	0.61	0.24	0.19	0.68
	F1-M	0.91	0.57	0.44	0.23	0.22	0.67
559	Precision	0.98	0.51	0.29	0.27	0.31	0.69
	Recall	0.98	0.67	0.57	0.34	0.23	0.58
	F1-M	**0.98**	0.58	0.39	0.30	0.27	0.63
563	Precision	0.87	0.41	0.27	0.37	**0.59**	**0.77**
	Recall	**1.00**	0.55	0.48	0.48	0.23	**0.78**
	F1-M	0.93	0.47	0.35	0.42	0.33	**0.77**
567	Precision	**1.00**	0.24	0.25	0.24	0.36	**0.77**
	Recall	0.77	0.55	0.43	0.32	0.33	0.67
	F1-M	0.87	0.33	0.31	0.27	0.34	0.71
570	Precision	0.82	0.32	0.09	0.17	0.43	0.53
	Recall	0.95	0.39	0.11	0.20	0.17	0.58
	F1-M	0.88	0.35	0.10	0.18	0.25	0.55
575	Precision	0.99	0.47	0.32	**0.44**	0.51	0.76
	Recall	0.92	0.70	0.56	0.26	**0.45**	**0.78**
	F1-M	0.95	0.56	0.41	0.33	**0.48**	**0.77**
584	Precision	0.95	**0.92**	0.20	0.22	0.16	0.49
	Recall	**1.00**	0.67	0.20	0.17	0.18	0.50
	F1-M	0.97	**0.77**	0.20	0.19	0.17	0.49
588	Precision	0.80	0.24	0.05	0.15	0.44	0.59
	Recall	0.95	0.37	0.06	0.14	0.40	0.51
	F1-M	0.87	0.29	0.05	0.14	0.42	0.55
591	Precision	**1.00**	0.35	0.27	0.38	0.41	0.69
	Recall	0.68	0.45	0.37	**0.63**	0.17	**0.78**
	F1-M	0.81	0.39	0.31	**0.48**	0.24	0.73
596	Precision	0.95	0.48	0.29	0.28	0.36	0.74
	Recall	**1.00**	**0.91**	0.43	0.35	0.34	0.51
	F1-M	**0.98**	0.62	0.35	0.31	0.35	0.60

Table 41: Population-based hybrid model results with less test data for each subject using 6 classes

Subject	**Metric**	**Class**					
		0	1	2	3	4	5
540	Precision	**1.00**	0.35	0.29	0.38	0.35	0.66
	Recall	0.88	0.84	0.54	0.37	0.27	0.53
	F1-M	0.94	0.50	0.38	0.38	0.30	0.59
544	Precision	0.92	0.38	**0.50**	0.32	**0.47**	0.77
	Recall	**1.00**	0.74	0.46	0.42	0.19	0.78
	F1-M	0.96	0.50	**0.48**	0.36	0.27	**0.78**
552	Precision	0.97	0.57	0.40	0.32	0.37	0.70
	Recall	0.99	0.77	**0.61**	0.27	0.24	0.75
	F1-M	0.98	**0.66**	0.48	0.29	0.29	0.72
559	Precision	**1.00**	0.52	0.34	0.30	0.33	0.69
	Recall	0.92	0.65	0.65	0.33	0.32	0.58
	F1-M	0.96	0.58	0.45	0.32	0.32	0.63
563	Precision	0.98	0.45	0.36	0.35	**0.54**	0.73
	Recall	0.87	0.71	0.56	0.46	0.25	0.81
	F1-M	0.92	0.55	0.43	0.40	0.34	0.77
567	Precision	0.99	0.43	0.24	0.25	0.34	**0.82**
	Recall	0.96	0.77	0.33	0.35	**0.37**	0.60
	F1-M	0.98	0.55	0.28	0.29	0.35	0.69
570	Precision	0.81	0.38	0.11	0.21	0.42	0.60
	Recall	0.97	0.50	0.14	0.22	0.22	0.53
	F1-M	0.88	0.43	0.12	0.22	0.29	0.56
575	Precision	0.99	0.51	0.33	**0.48**	0.45	0.70
	Recall	0.97	0.71	0.44	0.24	0.26	0.87
	F1-M	0.98	0.60	0.37	0.32	0.33	0.77
584	Precision	**1.00**	0.75	0.19	0.23	0.22	0.47
	Recall	0.89	0.50	0.20	0.20	0.28	0.44
	F1-M	0.94	0.60	0.19	0.21	0.25	0.46
588	Precision	0.95	0.27	0.11	0.23	0.41	0.65
	Recall	0.95	0.37	0.18	0.28	0.33	0.56
	F1-M	0.95	0.31	0.14	0.25	0.37	0.60
591	Precision	0.91	0.64	0.38	0.36	0.41	0.69
	Recall	0.99	0.64	0.35	**0.52**	0.14	**0.83**
	F1-M	0.95	0.64	0.36	**0.42**	0.21	0.75
596	Precision	**1.00**	**0.99**	0.28	0.27	0.42	0.72
	Recall	0.99	**0.88**	0.43	0.31	**0.37**	0.57
	F1-M	**0.99**	0.63	0.34	0.29	**0.39**	0.64