HEALTH BEHAVIOR INFERENCE FROM CONTINUOUS BLOOD GLUCOSE DATA: A HIDDEN SEMI-MARKOV APPROACH FOR PATIENTS WITH DIABETES

A Thesis

by

MOHIT DEEPAK CHHAPARIA

Submitted to the Graduate and Professional School of
Texas A&M University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Chair of Committee,	Madhav Erraguntla
Co-Chairs of Committee,	Scott Bruce
Committee Members,	Mark Lawley
Head of Department,	Lewis Ntaimo

May 2023

Major Subject: Industrial Engineering

ABSTRACT

Background: Diabetes is a condition when the body doesn't produce enough insulin or fails to use it as efficiently as it should. As per American Diabetes Association, in 2019 about 12.84% Americans were children and adolescents who had type-I diabetes. For patients with diabetes, hypoglycemia is a condition in which blood sugar (glucose) is lower than the standard range whereas hyperglycemia is a condition in which blood sugar (glucose) is higher than the standard range. Hypoglycemic events can lead to serious life threatening consequences whereas hyperglycemic events can lead to slow and permanent damage to internal organs for patients with type I diabetes.

Objective: This research is aimed to develop a model to predict the probability of hypoglycemia in the next 1 hour at each 5 minute intervals. The model is expected to have accuracy comparable to the ML models, better interpretability, and ability to forecast events like hypoglycemia, hyperglycemia, glucose values, etc.

Methods: The research implements the Hidden semi-Markov model with the help of the `R` package `mhsmm`, and custom user defined distributions and applies a Monte Carlo approach for forecasting.

Results: Patient-specific and Population-level models are developed and the results are explained by comparing the predicted probability of hypoglycemia with the observed glucose values. For a specific threshold on the population-level model, the sensitivity, and specificity for 30 minute ahead forecast are 91.35% and 75.03% and for 60 minute ahead forecast are 89.76% and 65.27%, respectively. The 30 minute ahead forecast and 60 minute ahead forecast ROC-AUC for the population level model are 0.9035 and 0.8214, respectively. In literature the 30 minute ahead prediction sensitivity, specificity, and ROC-AUC are generally in the range 74%-95%, 79%-96%, and 0.73-0.93, respectively. The GitHub repository links for these models are provided in Appendix A.

Conclusions: For hypoglycemia prediction, the HSMM model provides better explainability of a patient's physiological latent states compared to the ML models and comparable sensitivity and specificity. The prediction accuracy can be further improved by introducing other parameters like carbohydrates and insulin as covariates directly into the model.

DEDICATION

To my mother, father, and brother.

ACKNOWLEDGMENTS

This study involves the use of secondary analysis of deidentified data that was not collected specifically for this project and is not human subject research (Texas AM IRB number 2019-0710). This research was based on de-identified CGM data obtained from 20 patients using Dexcom G6 CGM devices under normal living conditions. Corresponding insulin pump data for participants provided details on the amount of insulin administered, it's time of delivery, and the associated carbohydrate count. The author declares no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supported by a thesis committee consisting of Professor Madhav Erraguntla [advisor] and Professor Mark Lawley of the Department of Industrial and Systems Engineering and Professor Scott Bruce [co-advisor] of the Department of Statistics.

The data analyzed for this research was provided by Texas Children's Hospital. The entire thesis committee actively provided inputs on the research and development of the report.

All other work conducted for the thesis was completed by the student independently.

Funding Sources

Graduate study was supported by a Graduate Research Assistantship from Professor Mark Lawley during the Fall 2022 semester. No other outside source of funding was provided.

NOMENCLATURE

ANN	Artificial Neural Network
ARIMA	Autoregressive Integrated Moving Average
ARISES	Adaptive, Real-Time, and Intelligent System to Enhance Self-Care
AUC	Area Under the Curve
CGM	Continuous Glucose Monitoring
DL	Deep Learning
ELM	Extreme Learning Machines
FAR	False Alert Rate
FDA	Food and Drug Administration
FN	False Negative
FP	False Positive
FPR	False Positive Rate
HMM	Hidden Markov Model
HSMM	Hidden Semi-Markov Model
LASSO	Least Absolute Shrinkage and Selection Operator
LogRLasso	Logistic Linear Regression with LASSO Regularization
LR	Logistic Regression
MHSMM	Multiple Hidden Semi-Markov Model
ML	Machine Learning
NS	No Sampling
PH	Prediction Horizon
TN	True Negative

TP	True Positive
TPR	True Positive Rate
RELM	Regularized Extreme Learning Machines
RF	Random Forest
ROC	Receiver Operating Characteristic
SSRTSM	Subject-Specific Recursive Time Series Models
SVM	Support Vector Machine
VIP	Variable Importance Plot

TABLE OF CONTENTS

Page

LIST OF FIGURES

FIGURE Page

LIST OF TABLES

TABLE Page

1. INTRODUCTION AND LITERATURE REVIEW

Maintaining steady and appropriate blood glucose levels is essential to proper health and functioning. A hypoglycemic event occurs when blood glucose drops below typical levels and can be fatal. Alternatively, hyperglycemia can occur when blood glucose levels rise to unusually high levels, which can lead to organ damage if not controlled. Accurate predictions of hypoglycemic events will allow physicians to be aggressive with the insulin administration process. This will serve two purposes: (1) Higher insulin amounts will ensure that the patient spends less time with unusually high blood glucose. (2) Accurate hypoglycemic predictions will instill confidence in the physicians and the patients that the hypoglycemic events arising due to high insulin administration or any other reason won't prove to be fatal. Absence of hypoglycemic data is not only a challenge for predictive models but also an indication of a patient's blood glucose level control as absence of hypoglycemic data might suggest either very strong glycemic control or very high blood glucose levels. Handelsman and Turtle [1] concluded that the absence of any mild hypoglycemia is a strong pointer of poor glycemic control. A study by Ary et al. [2] showed that patients with Type II diabetes reported adhering about 53% the time to exercise prescriptions whereas patients with Type I diabetes adhered about 31% the time and Type II diabetes patients who were prescribed insulin reported adhering to the regimen 90.3% of the time compared to 78.3% adherence by the Type I diabetes patients. Such behaviors by patients with Type I diabetes makes hypoglycemia prediction more significant for this group of individuals.

Real-time continuous glucose monitoring is essential for diabetes management and is generally used to establish baselines to classify a event as hypoglycemic. The data set used in this project contained approximately 2.42% glucose readings in the hypoglycemic range. In this study glucose values below 70 mg/dL are defined as hypoglycemic, glucose values from 70 mg/dL to 180 mg/dL are defined as normal, and glucose values of 180 mg/dL and above are defined as hyperglycemic. The major problem in hypoglycemia prediction is the high False Alert Rate (FAR) of these models, this damages the confidence of a patient and a physician in the predictions and thus, makes their

real-time application inefficient. The underlying reason for high FAR of these models is generally the imbalanced nature of the data set containing continuous glucose values [3]. For best glycemic control and to avoid serious hypoglycemia, education of patient and advice not to miss meals and to cover periods of unusual exercise with additional carbohydrate intake usually suffice.

The first attempt to predict present and future glucose values using the recent past blood glucose history was made in 1999 by Bremer et al. [4]. Since then many time-series (like ARIMA-ARIMAX, state-space model, etc.), machine learning (like Logistic Regression, Support Vector Machine, Random Forests, Decision Trees, etc.), Artificial Neural Networks, and other kernel based models have been proposed to analyze, understand, and predict present & future glucose, insulin, hypoglycemia, and hyperglycemia events in adolescents and adults with type 1 as well as type 2 diabetes [5–20].

As described by ElMoaqet et al. [21], the fundamental problem in standard modeling and evaluation method used in analyzing engineering dynamic systems is to minimize the (mean) error between the real and predicted systems. These models are applied to multi-step ahead predictions of physiological signals, but clinically predicting relevant physiological events is as important as predicting the signals. Machine Learning models provide a high accuracy in predicting such events but lack in explaining the physiological state of the patient which lead to the specific signal. These models generally use multiple input variables like insulin intake, carbohydrate intake, glucose, heart rates, etc. while making predictions. All these variables are mostly not available from a single data source which makes real-time applications of these models difficult. As stated by Dave et al. [5], these models generally can be broken down as classification-based models for predicting future hypoglycemic events or regression-based models for predicting future glucose values. That is, these models generally have specific targeted outcomes and cannot be used for multiple simultaneous applications.

In order to appropriately characterize blood glucose dynamics, we propose the use of a hidden-semi Markov model using gamma distributions to flexibly model the emission and sojourn distributions for each of the latent states. This project applies Hidden semi-Markov model to develop

this probabilistic prediction approach with the help of the `mhsmm R` package explained in the journal paper by O'Connell and Højsgaard [22]. The `mhsmm` package allows custom distributions and uses EM algorithm for parameter estimation. O'Connell and Højsgaard [22] described two other software packages available for Hidden semi-Markov models. The first was `AMAPmod` software by Godin and Guédon [23] and the second was `hsmm` package by Bulla, Bulla, and Nenadić [24]. Compared to these two packages, the flexibility to estimate parameters and to create custom distributions make the `mhsmm` package a good choice for this project. The developed HSMM model is expected to have accuracy comparable to the Time-Series and Machine Learning models available in the literature for similar applications. In this project, the physiological state of a patient which result in specific glucose levels is of high clinical importance. It is also expected to better explain the physiological characteristics of the patients leading to specific glucose levels. Latent states are posited that characterize different physiological characteristics associated with different blood glucose levels. We also use a Monte Carlo approach to generate multiple latent state sequences for future time points that can be used to forecast various clinically-relevant outcomes, such as hypoglycemic events, hyperglycemic events, future glucose values, etc. The proposed model uses CGM values alone which is easily available and thus, makes real-time application simpler.

The following manuscript is organized as follows: Chapter 2 presents the theory of the models and the EM algorithm for HSMM. Chapter 3 explains the latent states, and their initialization process. Chapter 4 describes the dataset, the model parameters, some user-defined functions, and the types of models developed. Chapter 5 defines the prediction approach, the training and test datasets, and the model interpretation. Chapter 6 contains the evaluation process. Finally Chapter 7 contains the project summary and future work.

2.1 Discrete Markov Chains

A discrete Markov chain is a sequence of discrete random states of a system where the probability of entering a state at time $n+1$ depends only on the state at time n. Figure 2.1 shows a generic diagram of a 3 state discrete Markov chain. As stated by O'Connell and Højsgaard [22], mathematically it can be represented as,

$$P(S_{n+1} = s_{n+1}|S_0 = s_0, S_1 = s_1, S_2 = s_2, \ldots, S_n = s_n) = P(S_{n+1} = s_{n+1}|S_n = s_n) \quad (2.1)$$

where S_{n+1} is the state of the discrete Markov chain at time $n+1$.

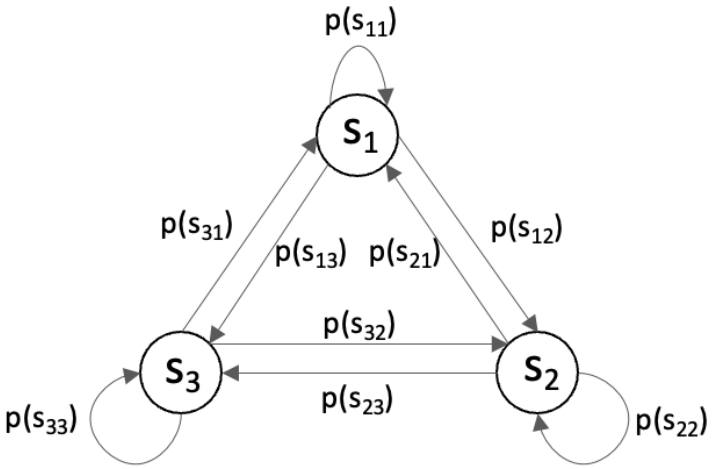

Figure 2.1: Generic representation of a 3 state discrete Markov chain. Here $p(s_{ij})$ represents the probability of a transition from state i to state j.

2.2 Hidden Markov Models

Discrete Markov chains as described above are used in many mathematical models as the driving mechanism for characterizing the dynamics of stochastic processes. For example, hidden Markov models (HMMs) use a discrete Markov chain is a latent, unobserved process that fully governs the characteristics of a stochastic process through modeling of state-specific stochastic dynamics. More specifically, Yu, Shun-Zheng [25] describes HMM as a doubly stochastic process where the underlying stochastic process is a discrete-time finite-state homogeneous Markov chain and the state sequence influences another stochastic process that produces a sequence of observations. In HMMs, the transition of state at time i to state at time $i+1$ is dependent only on the state at time i, hence HMMs are also regarded as a memoryless processes. Also, HMM has a non-zero self-transition property. Similar to the notation of O'Connell and Højsgaard [22], we have:

- $P(S_0)$ is the initial distribution and it is represented as a vector π. Initial Distribution represents the probability of being in a particular state at time 0.
- $P(S_t|S_{t-1})$ is the transition distribution given the state in the previous time point. This distribution can be represented as a collection of transition probabilities from state i to state j for $i, j = 1, 2, \ldots, M$ where M represents the number of latent states in matrix form.
- $P(X_t|S_t)$ is the emission distribution which is represented by b. Emission distribution represents the distribution of the observed stochastic process given the current state.

Here, S represents the latent discrete Markov chain and X corresponds to the stochastic process of interest that depends on S. Similarly, X_t represents the stochastic process of interest at time t and S_t represents the latent discrete Markov chain at time t. Figure 2.2 shows a generic diagram of a Hidden Markov model. O'Connell and Højsgaard [22] have specified HMM by a triple $\theta = (\pi, P, b)$ and presented its mathematical functional form as,

$$P(S, X) = P(S_0)\prod_{t=1}^{T} P(S_t|S_{t-1})\prod_{t=1}^{T} P(X_t|S_t) \tag{2.2}$$

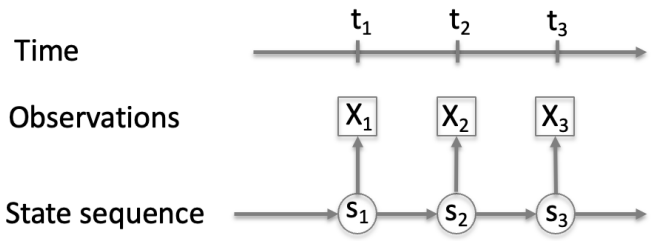

Figure 2.2: Generic representation of a hidden Markov model.

2.3 Hidden Semi-Markov Models

The limitation in using HMMs for this application is that the sojourn distribution for HMMs is generally geometrically distributed. A flexible approach to solve this problem is to use hidden semi-Markov models (HSMMs). In HSMMs, the distribution of the sojourn time varies both in shape and form (geometric, normal, gamma, etc.) depending upon the application. Zucchini, MacDonald and Roland [26] describe HMM as a special case of HSMM in which the sojourn distributions of the HSMM are geometrically distributed. The unobservable state sequence in HSMM is semi-Markov in nature, that is, the probability of transition to a new state is dependent on the previous state and a transition occurs when the system has spent the required time derived from the sojourn distribution associated with the previous state. Since HSMMs consider the sojourn time to determine when a transition will occur, they are no longer memoryless and the self-transition probabilities of all the states are zero. Along with the notations presented in the above section, O'Connell and Højsgaard [22] have used one additional notation in HSMM which is as follows:

- $d(u)$ is the sojourn distribution and it is represented by d. Sojourn distribution is a distribution of a set of values for individual states in the system representing the number of timepoints the system stays in a state once it enters the state.

Figure 2.3 shows a generic diagram of a Hidden semi-Markov model. Mathematically, in the `mhsmm` package, O'Connell and Højsgaard [22] have specified HSMM by a quadruple $\theta =$

(π, P, b, d) and stated the complete likelihood of a HSMM is,

$$P(X = x, S = s; \theta) = \pi_{s_1^*} d_{s_1^*}(u_1)\{\prod_{r=2}^{R} p_{s_{r-1}^*} s_r^* d_{s_r^*}(u_r)\} p_{s_{R-1}^*} s_R^* D_{s_R^*}(u_R) \prod_{t=1}^{T} b_{s_t}(x_t) \tag{2.3}$$

where s_r^* is the r^{th} visited state, u_r is the time spent in that state, and $D_i(u)$ is the survivor function.

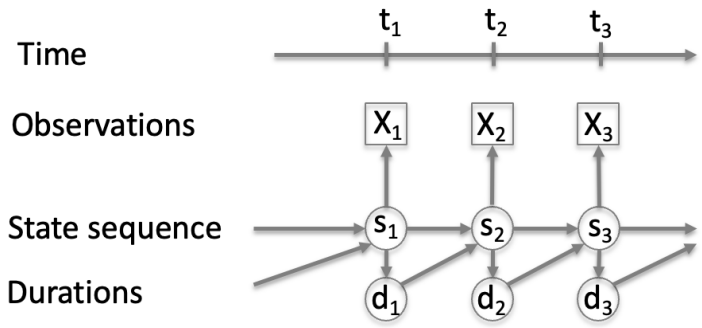

Figure 2.3: Generic representation of a hidden semi-Markov model.

2.4 The EM Algorithm for Hidden Semi-Markov Models

The EM algorithm for HSMM implemented in the `mhsmm` package by O'Connell and Højsgaard [22] is as follows:

2.4.1 E-step

The E-step implemented in the `mhsmm` package involves estimating three terms:

1. The probability of being in state i at time t given the observed sequence,

$$\gamma_t(i) = P(S_t = i | X = x; \theta) \tag{2.4}$$

2. The probability that the system left state i at time t and entered state j at time $t+1$ given the observed sequence,

$$\xi_t(i,j) = P(S_t = i, S_{t+1} = j | X = x; \theta) \tag{2.5}$$

3. The expected number of times a system spends u time steps in state j,

$$\begin{aligned}\eta_{iu} = P(S_u \neq i, S_{u-\upsilon} = i, \upsilon = 1, \ldots, u | X = x; \theta) \\ + \sum_{t=1}^{T} P(S_{t+u+1} \neq i, S_{t+u-\upsilon} = i, \upsilon = 0, \ldots, u-1, S_t \neq i | X = x; \theta)\end{aligned} \tag{2.6}$$

2.4.2 M-step

The M-step implemented in the `mhsmm` package involves estimating the following terms:

1. Estimating the initial and transition probabilities,

$$\hat{\pi}'_i = \gamma_0(i) \tag{2.7}$$

$$\hat{p}'_{ij} = \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \sum_{i \neq j} \xi_t(i,j)} \tag{2.8}$$

2. Estimating the parameter for the emission distribution: The `mhsmm` package offers multiple distributions and also the ability for users to define custom distributions. In this case, we use a custom defined gamma emission distribution which is further explained in section 4.2.

3. Estimating the state duration density: In this package, the author has implemented an ad-hoc solution for using parametric distributions with η_{iu} proposed by Guédon [27]. Just like emission distribution, the package offers multiple distributions for sojourn times and the

ability for users to define custom distributions. In this case, we use a custom defined gamma emission distribution which is further explained in section 4.2.

To better understand the development, implementation, and complexities of the EM algorithm for HMM and HSMM, one can refer to Rabiner [28] and Guédon [27], respectively.

Throughout this study instead of shape and scale, mean and standard deviation are used to describe gamma distribution (Emission and Sojourn Distributions). For calculating the shape and scale parameters of the distributions, following equations can be used,

$$shape = \Big(\frac{mean}{standard\ deviation}\Big)^2 \tag{2.9}$$

$$scale = \frac{(standard\ deviation)^2}{mean} \tag{2.10}$$

3. LATENT STATES

Latent states or hidden states are inferred based on the developed Hidden semi-Markov model and the observable sequence. In this project, we are interested in the probability of hypoglycemic event based on the predicted latent state by the trained Hidden semi-Markov model using the observed glucose values. Latent states are dependent on the physiological state of a patient. These physiological states are affected by factors like activity level, insulin in-take and individual absorption capacity, sleep cycle, carbohydrate consumption, stress levels, etc. These individual factors have different independent influence on CGM levels, for example, carbohydrates boosts blood glucose levels, insulin reduces CGM levels by helping glucose enter body's cells which is then converted into energy, exercise reduces CGM levels by burning blood glucose, etc. Although these factors have a significant effect on a patient's glucose levels, these data values can be less reliable as they're generally an approximate estimation as per the patient's input (which can be inaccurate) or are very noisy or are not easily available. On the other hand, CGM values are a set of continuous data streams recorded by a standardized device approved by the FDA. Thus, we seek to build a predictive model based on CGM alone as it is highly reliable, a long continuous data stream, and existing data all patients have for physicians.

As glucose levels change, physiological behavior also changes, so we are characterizing our latent states through potential physiological mechanisms associated with varying glucose levels. Based on this, we define the latent states as,

- Latent State 1: Physiological state that results in low glucose values. This physiological state can be a result of multiple factors like high insulin intake, low carbohydrate intake, high activity levels, etc. occurring together.

- Latent State 2: Physiological state that results in low-normal glucose values. This physiological state can be a result of occurrence of one or more factor like high to normal-high insulin intake, low to low-normal carbohydrate intake, high to normal-high activity levels,

etc. This is an intermediate state between normal and low glucose levels.

- Latent State 3: Physiological state that results in normal glucose values. This physiological state can be a result of a well-balanced and controlled combination of multiple factors.

- Latent State 4: Physiological state that results in normal-high glucose values. This physiological state can be a result of occurrence of one or more factor like low to low-normal insulin intake, high to normal-high carbohydrate intake, low to low-normal activity levels, etc. This is an intermediate state between normal and high glucose levels.

- Latent State 5: Physiological state that results in high glucose values. This physiological state can be a result of multiple factors like low insulin intake, high carbohydrate intake, low activity levels, etc. occurring together.

We have initialized the latent states using the glucose values from the data set which are described in Table 3.1. This initialization is further used to define emission distribution, sojourn distribution, and transition probabilities.

Table 3.1: Glucose ranges used for parameter initialization.

State	Glucose Range (G)
1	G < 15^{th} percentile
2	15^{th} percentile $\leq$ G < 37.5^{th} percentile
3	37.5^{th} percentile $\leq$ G < 62.5^{th} percentile
4	62.5^{th} percentile $\leq$ G < 85^{th} percentile
5	G $\geq$ 85^{th} percentile

4.1 Dataset

The data set used in this project consisted of glucose values recorded every 5 minutes for 20 patients with the help of Dexcom G6 CGM devices. The device measures glucose using the interstitial fluids. This data set is a part of the data set described and used by Dave et al. [5] where data was obtained for 112 patients over a range of 90 days under normal living conditions. This study comprised of 6 male and 14 female patients. Baseline characteristics for the patients included in this study are described in Table 4.1. Table 4.2 gives descriptive statistics of the glycemic values in data sets.

Table 4.1: Patient baseline characteristics when the data was recorded.

Metric	Mean $\pm$ Standard Deviation	Median	Range
Size of Data set (Days)	27.55 ± 4.40	28.65	14.87 - 34.74
Age of Patient (Years)	10.95 ± 5.65	11.00	1.00 - 19.00
Duration of Diabetes (Years)	3.99 ± 3.50	2.22	0.32 - 14.53
HbA1c (%)	7.9 ± 1.5	7.4	5.6 - 10.7

Table 4.2: CGM metrics.

Metric	Mean $\pm$ Standard Deviation	Median	Range
Hypoglycemic values/day[1]	6.86 ± 6.01	5.90	0.14 - 23.55
% hypoglycemic values[2]	2.38 ± 2.09	2.05	0.05 - 8.18
Hyperglycemic values/day[3]	117.84 ± 68.02	114.31	0.00 - 244.59
% hyperglycemic values[4]	40.92 ± 23.62	39.69	0.00 - 84.93

[1]Mean number of a patient's glycemic values below 70 mg/dL in one day.
[2]Percent of a patient's glycemic values below 70 mg/dL.
[3]Mean number of a patient's glycemic values above 180 mg/dL.
[4]Percent of a patient's glycemic values above 180 mg/dL.

4.2 Parameters of Hidden Semi-Markov Models

4.2.1 Initial Distribution

In HSMM, the initial distribution provides the probability of being in a state when we start the model training process. In this project, the initial probability of all states is set to be equal. We also tried another approach to set the initial distribution as per the quantiles used for state parameter initialization (that is, 0.15, 0.225, 0.25, 0.225, 0.15 for states 1 to 5, respectively). A change in the initial probability did not give any notable change in the predictions or the trained model parameters, indicating that the model is not sensitive to initial distribution (π).

4.2.2 Emission Distribution

This is the distribution of the observed values for each state in the system. The threshold for each state in the emission distribution is determined using 15^{th}, 37.5^{th}, 62.5^{th}, and 85^{th} percentile of Glucose values. Gamma distribution showed the best fit for each individual state within the set of emission values (Glucose values). Figure 4.1 shows the gamma density distribution plot of all the states. The plot was developed using the trained emission distribution parameters from the population-level model. Table 4.3 shows the before and after training (population-level model) mean and standard deviation values for the gamma emission distribution parameter for all the states. Table 4.5 shows the mean and standard deviation values for the gamma emission distribution parameter for states 1 and 5. The table helps to describe the variation in state distributions in different models, thus explaining the physiological state of a patient. This point is further explained in section 5.3.

4.2.3 Sojourn Distribution

Sojourn distribution is a distribution of the time the system spends in individual states once it enters the state. Using the initialization thresholds for each state (described in Table 3.1), the set of time values the system spends in a state is determined. Figure 4.4 shows the fit of a Gamma Distribution using the trained model shape and scale sojourn distribution parameters of the population-

Table 4.3: Gamma emission distribution parameters for the population-level model.

State	Initialized Values (mg/dL)	Trained Values (mg/dL)
	Mean ± Standard Deviation	Mean ± Standard Deviation
1	80.12 ± 11.17	85.96 ± 13.24
2	113.18 ± 11.22	121.93 ± 11.62
3	159.52 ± 15.78	163.18 ± 14.22
4	221.22 ± 20.03	214.58 ± 17.45
5	305.34 ± 38.12	294.20 ± 40.83

level model. Figure C.1 shows the sojourn distribution for all the latent states after the final iteration while training the population level model for 288 timepoints. The plot shown is a replication of the plot generated with the help of the graphical argument in the hsmmfit function present in the `mhsmm R` package. Table 4.4 shows the before and after training (population-level model) mean and standard deviation values for the gamma sojourn distribution parameter for all the states.

Table 4.4: Gamma sojourn distribution parameters for the population-level model.

State	Initialized Values	Trained Values
	Mean ± Standard Deviation	Mean ± Standard Deviation
1	14.26 ± 20.76	25.64 ± 29.45
2	10.26 ± 12.62	11.87 ± 11.91
3	11.80 ± 12.62	12.39 ± 11.66
4	14.21 ± 15.59	15.26 ± 13.87
5	24.34 ± 31.57	38.76 ± 35.56

4.2.4 Transition Matrix

Transition matrix $P = (p_{ij})$ represents the probability matrix of having a transition from state s_i at any time t_n to state s_j at time t_{n+1}. Equation 4.1 represents the generic form of such a transition matrix. The initial transition matrix was estimated by calculating the fractions of transition from each state available in the data set in consideration with the help of the thresholds defined in Table 3.1. Glucose values of a patient increase or decrease gradually and hence, transitions only

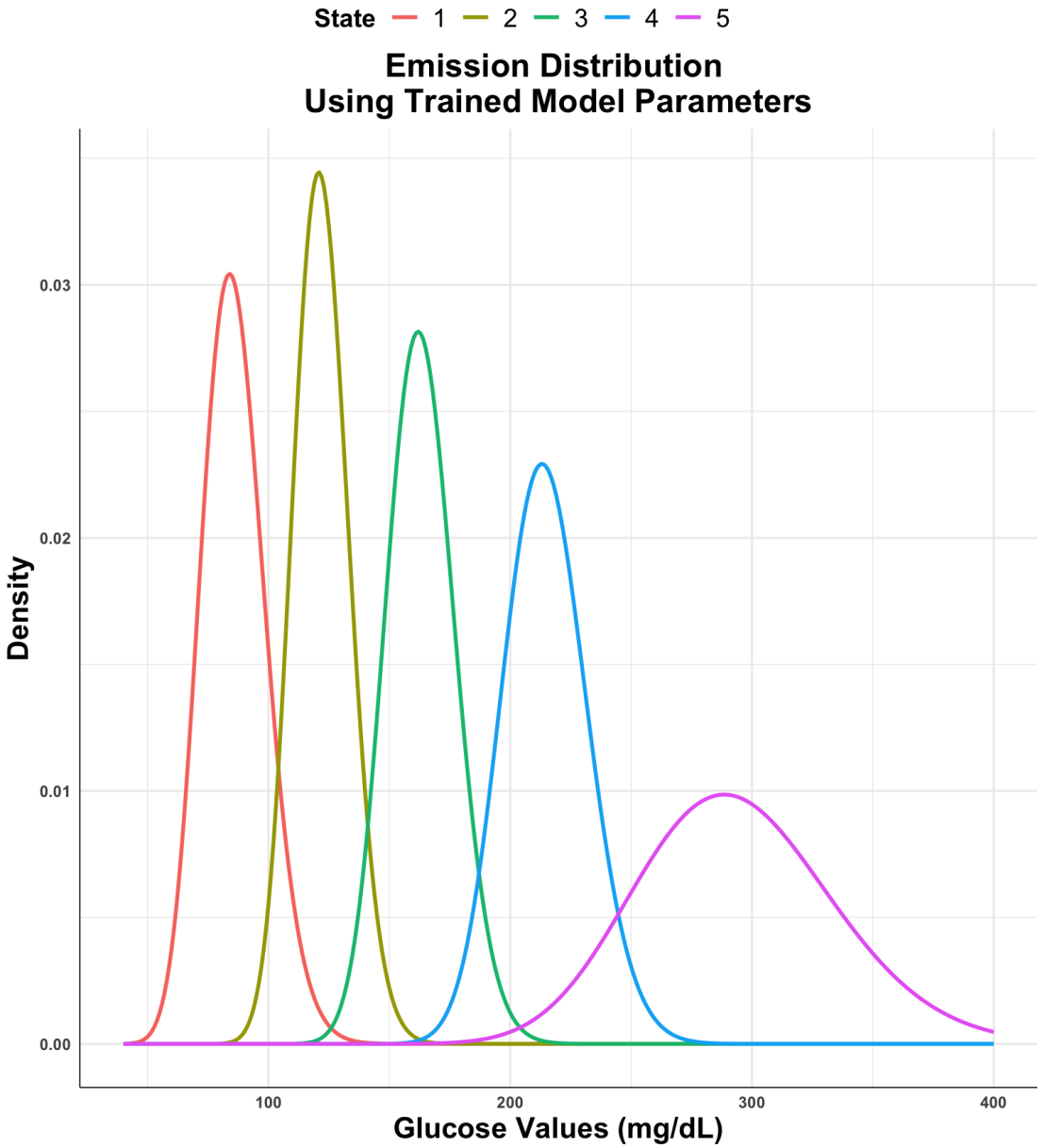

Figure 4.1: Density plot of glucose values based on trained model (population-level) emission distribution parameters for all the latent states.

occur when $i - j = \pm 1$. In HSMM, the transition probability within the same state (self-transition probability) is zero ($p(s_{ij}) = 0$ when $i = j$). Figure 4.3 shows all possible state transitions in this project. Equation 4.2 and Equation 4.3 show the initialized and trained transition matrix for the population-level model. For presentation purposes, the numbers in these two matrices have been rounded upto 3 digits after the decimal point.

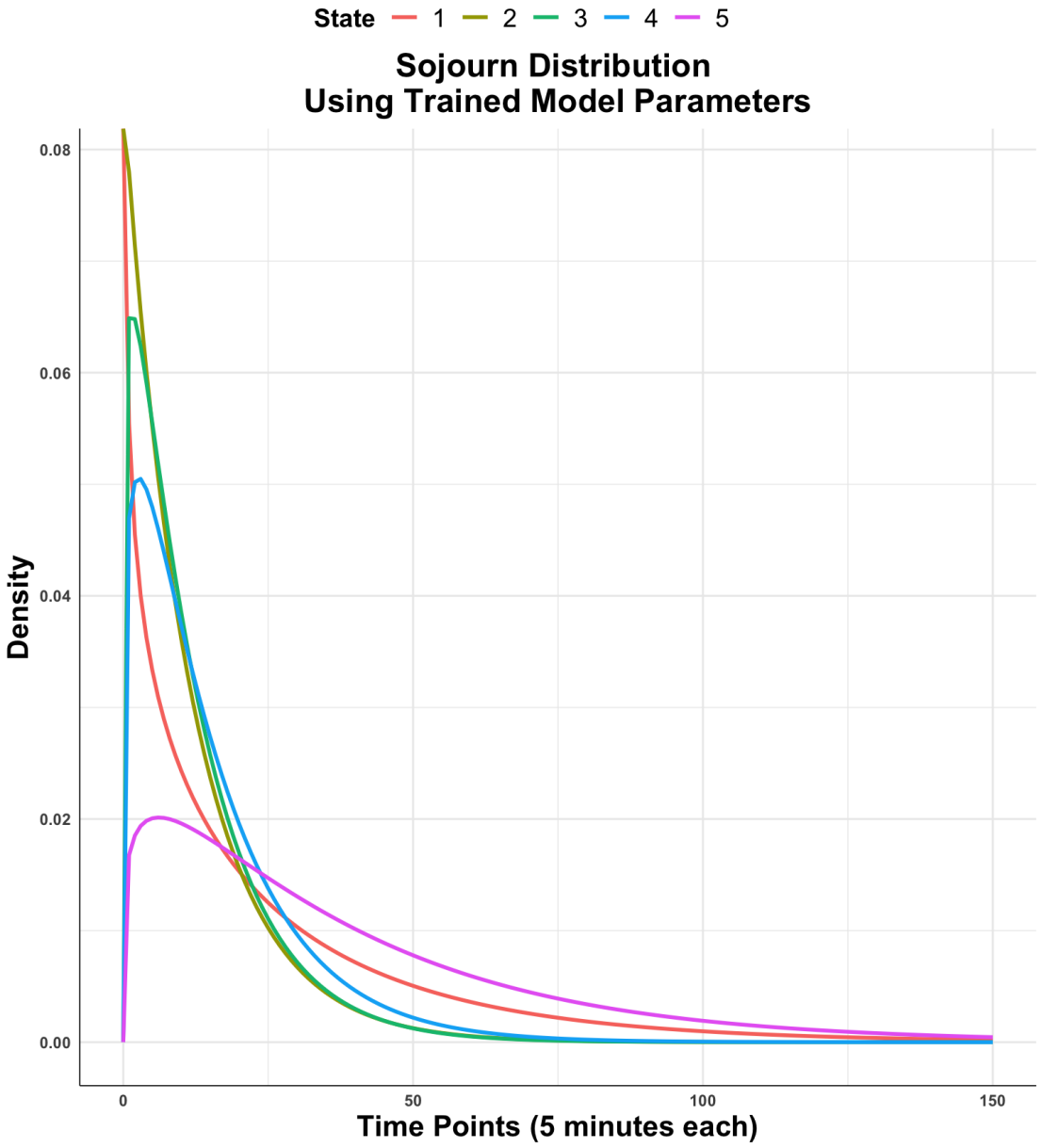

Figure 4.2: Density plot of time points based on trained model (population-level) sojourn distribution parameters for all the latent states.

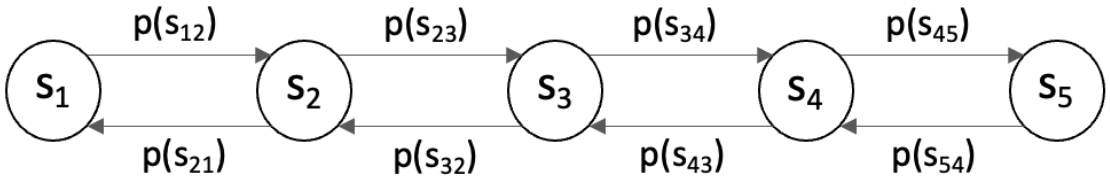

Figure 4.3: All possible transitions in this model.

$$Generic\,Transition\,Matrix\;=\;\begin{array}{c} State \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array}\begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{pmatrix} p(s_{11}) & p(s_{12}) & p(s_{13}) & p(s_{14}) & p(s_{15}) \\ p(s_{21}) & p(s_{22}) & p(s_{23}) & p(s_{24}) & p(s_{25}) \\ p(s_{31}) & p(s_{32}) & p(s_{33}) & p(s_{34}) & p(s_{35}) \\ p(s_{41}) & p(s_{42}) & p(s_{43}) & p(s_{44}) & p(s_{45}) \\ p(s_{51}) & p(s_{52}) & p(s_{53}) & p(s_{54}) & p(s_{55}) \end{pmatrix} \end{array} \tag{4.1}$$

$$Initialized\,Transition\,Matrix\;=\;\begin{array}{c} State \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array}\begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 0.463 & 0 & 0.537 & 0 & 0 \\ 0 & 0.553 & 0 & 0.447 & 0 \\ 0 & 0 & 0.612 & 0 & 0.388 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix} \end{array} \tag{4.2}$$

$$
Trained\,Transition\,Matrix\;=\;
\begin{array}{c}
State \\
1 \\
2 \\
3 \\
4 \\
5
\end{array}
\begin{array}{c}
\begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\
\left(\begin{array}{ccccc}
0 & 1 & 0 & 0 & 0 \\
0.467 & 0 & 0.533 & 0 & 0 \\
0 & 0.546 & 0 & 0.454 & 0 \\
0 & 0 & 0.613 & 0 & 0.387 \\
0 & 0 & 0 & 1 & 0
\end{array}\right)
\end{array}
\tag{4.3}
$$

4.3 User-Defined Functions

The following are the user defined functions which are used during the prediction and the model development process.

4.3.1 Time to State

A function to determine the time a patient has already spent in a state as per the observed glucose sequence. With the help of this function and sojourn distribution, we determine when a transition will occur.

4.3.2 Re-estimation of Parameters (mstep.gamma)

This function is used to re-estimate the parameters for the emission distribution as part of the M-step of the EM algorithm defined in Section 2.4. In this case, we've developed a function to generate the shape and scale parameters of the gamma distribution for all the states.

4.3.3 Generate Random Deviates (rgamma)

This function is used to generate random deviates of the emission distribution. In this case, we've used the in-built `rgamma()` function of `R`.

4.3.4 Density Calculation (dgamma)

This function is used to calculate the density of the emission distribution observations. In this case, we've used the in-built `dgamma()` function of `R`.

4.4 Model Types

Two types of models were developed in this study.

1. Population-Level Model: This model uses parts of data set from all the patients and joins them together in series for the model training and initial parameter estimation process. Refer to Figure 5.2 and section 5.1 to understand how the data from different patients is sliced and used for training the population level model. This model gives better predictions as it is generated on the basis of diverse learning from multiple patients. The parameters generated here are a great representation of the group of patients and the general behavior of a patient with type 1 diabetes.

2. Patient-Level Model: This model uses part of data set from a single patient for model training and initial parameter estimation process. The data slicing process here is the same as it is for the population-level model. The only difference in the slicing process is that we are using only 1 patient's data set. This model gives better explainability of the physiological latent states as the parameters generated from the model are based only on the data set of the patient under consideration. This is further explained in Section 5.3. The parameters generated with the help of this model are a good representation of the glycemic behavior of individual patients and how their body reacts to carbohydrates, insulin, and different activities.

Table 4.5 shows the variation in gamma emission distribution parameters of latent state 1 and latent state 5 for the population-level model and the patient-level model of each patient.

Table 4.5: Variation in gamma emission distribution: trained model parameters.

Model	Latent State 1	Latent State 5
	Mean $\pm$ Standard Deviation	Mean $\pm$ Standard Deviation
Population	85.96 ± 13.24	294.20 ± 40.83
Patient 1	146.11 ± 31.99	312.11 ± 21.05
Patient 2	143.04 ± 27.41	355.88 ± 25.78
Patient 3	82.05 ± 12.35	233.01 ± 27.41
Patient 4	80.54 ± 14.40	299.21 ± 38.63
Patient 5	75.98 ± 9.95	250.41 ± 41.03
Patient 6	82.77 ± 15.05	279.29 ± 37.51
Patient 7	67.86 ± 9.99	217.76 ± 39.04
Patient 8	103.26 ± 17.02	313.75 ± 44.51
Patient 9	92.25 ± 16.03	293.38 ± 37.49
Patient 10	74.95 ± 13.80	290.95 ± 42.74
Patient 11	105.53 ± 25.39	338.60 ± 32.90
Patient 12	90.51 ± 14.98	299.24 ± 40.32
Patient 13	83.93 ± 14.21	257.35 ± 35.44
Patient 14	84.00 ± 15.10	262.13 ± 33.99
Patient 15	75.37 ± 4.75	109.04 ± 4.99
Patient 16	90.90 ± 11.91	237.50 ± 31.15
Patient 17	159.06 ± 35.17	386.07 ± 14.27
Patient 18	90.94 ± 14.64	264.34 ± 34.54
Patient 19	93.31 ± 13.17	322.98 ± 47.66
Patient 20	82.12 ± 8.20	167.43 ± 19.49

5. PREDICTION

The user-defined predict function developed for this project uses the trained Hidden semi-Markov model, the glucose sequence, and the number of time-points for which predictions are to be made (usually 30 minutes (6 time-points) or 60 minutes (12 time-points)). To replicate real-time conditions, only 1 prediction of 12 time-points is made based on the glucose sequence immediately following the last value on the training data set. This process is repeated 288 times to generate a prediction data set, representing a day worth of predictions, to evaluate the prediction accuracy. The user-defined predict function uses the in-built predict function from the `mhsmm` package to generate the underlying latent state sequence for a given series of CGM values up to time t. This in-built predict function uses Viterbi algorithm to generate the above stated latent state sequence which serves the following purposes in this study:

- To identify the latent state for the last CGM observation of the patient's training dataset.
- To identify how long the patient has been in the latent state identified in the above point. This is done by passing the generated latent state sequence for the training data to the user-defined Time to State function described in Section 4.3.1.

The time already spent in the current latent state and the sojourn distribution for that latent state help us determine how many additional time points the system will stay in the current latent state or when will a transition occur. Once we reach the point of transition a random draw is made from a 0 to 1 uniform distribution and random value is compared with the outgoing transition probabilities from the current latent state. Figure 4.3 shows all the possible transitions in this study. Once a transition is made, the sojourn distribution of the new latent state helps us determine how long we'll stay in the new latent state. This process is repeated till we've a prediction for the next 12 timepoints (1 hour).

A Monte Carlo approach is used where this simulation of generating a prediction for the next 12 timepoints is repeated 10000 times. Through this approach we generate a 5 x 12 matrix which

gives us the probability of being in latent states (1 to 5) at future timepoints (1 to 12). Additionally, a 3 x 5 matrix is developed with the help of the in-built `pgamma R` function and the trained model emission distribution parameters. This new matrix gives us the probability of glucose values being less than 70 mg/dL (hypoglycemic state), between 70 mg/dL to 180 mg/dL (normal state), and more than 180 mg/dL (hyperglycemic state) given that the system is in a particular latent state (1 to 5). These two matrices are then multiplied to generate a total probability (a 3 x 12 matrix) which gives us the probability of being in hypoglycemic, normal, and hyperglycemic states for each of the 12 future timepoints. Mathematically, this multiplication can be represented as follows:

$$\begin{aligned} P(Hypoglycemia \mid A\ timepoint) = P(Hypoglycemia \mid A\ latent\ state) \\ * P(A\ latent\ state \mid A\ timepoint) \end{aligned} \tag{5.1}$$

Similarly, we can mathematically represent the probability of being in a normal state and a hyperglycemic state. The calculated probability is then compared with a selected threshold to classify the prediction as a state of hypoglycemia or not. The threshold values and prediction evaluation process is described in Section 6.1. This method can be extended to classify the prediction as a state of hyperglycemia or not and to predict the glucose values. Figure 5.1 shows the predicted probability for the latent state of hypoglycemia against the observed glucose values for 30 minute ahead prediction generated using the population-level model.

5.1 Training Dataset

A moving window approach was implemented to develop 288 models. Each model was trained on a constant size of data set. The length of the training data set for model 1 is given by Equation 5.2. Similarly, the length of the training data set for the remaining 287 models can be shown. Through this process, we've a constant length of the training data set. Unequal sizes of the data from individual patient used for training purpose did not affect the model parameters or make the fit bias for any one patient as there was a small difference in the sizes of individual data sets (as described in section 4.1) and the glycemic levels of all the patients was spread across the

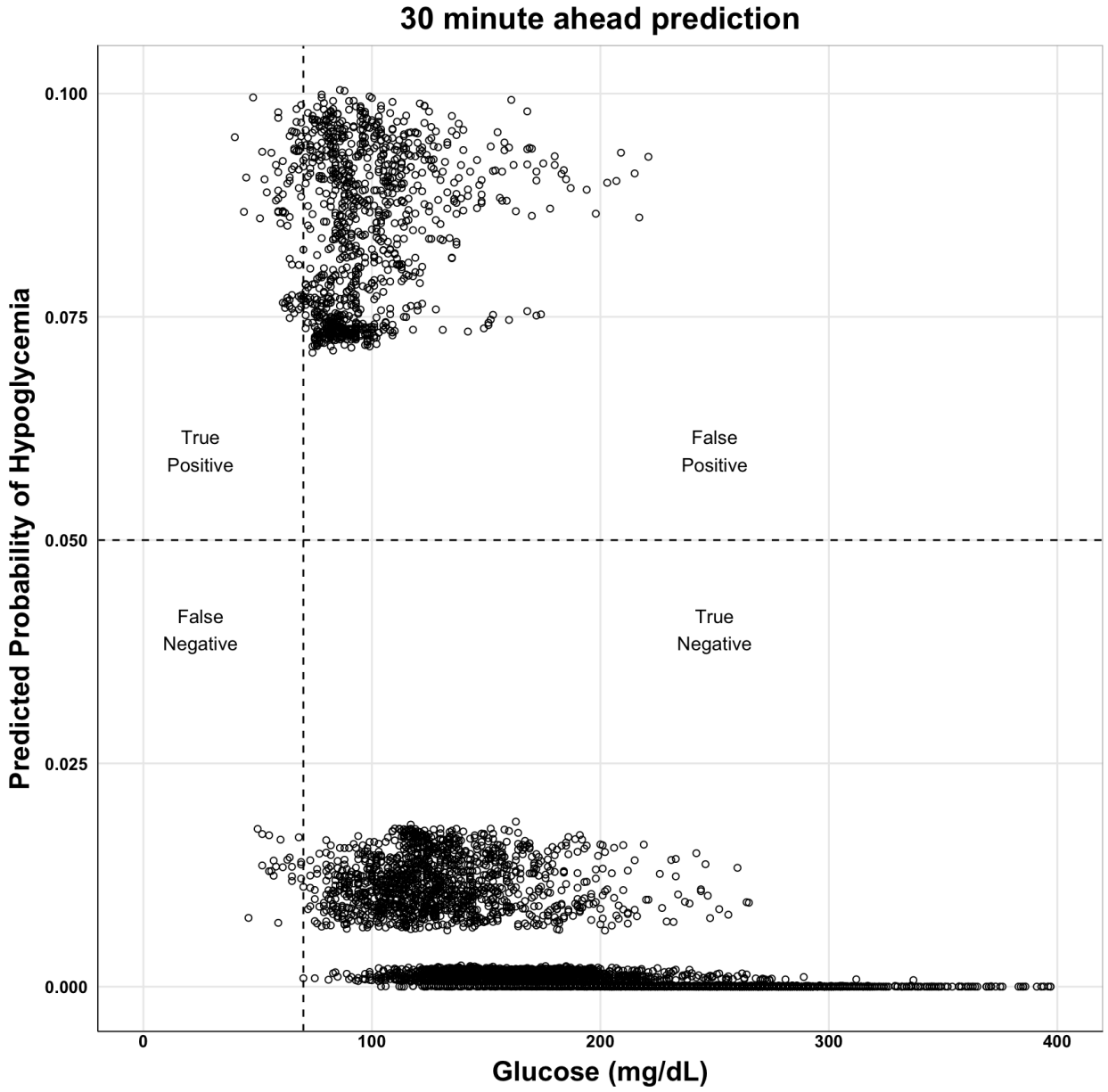

Figure 5.1: Predicted probability of hypoglycemia (from the population-level model) versus the observed glucose values (mg/dL).

entire range of values (as shown in Tables 4.3 and 4.5). The sliced data set from each patient was combined in series and passed as a single glucose sequence for training the population-level model. Figure 5.2 shows how the data set of individual patient was divided for training 288 models.

$$\begin{aligned} Total\ length\ of\ dataset\ for\ training\ model\ 1 &= (N_1 - 299) + (N_2 - 299) \\ &+ \ldots + (N_{19} - 299) + (N_{20} - 299) \\ &= N_1 + N_2 + \ldots + N_{20} - 20 * 299 \end{aligned} \tag{5.2}$$

where N_1 is the total number of records in the data set for patient 1, N_2 is the total number of records in the data set for patient 2, and so on.

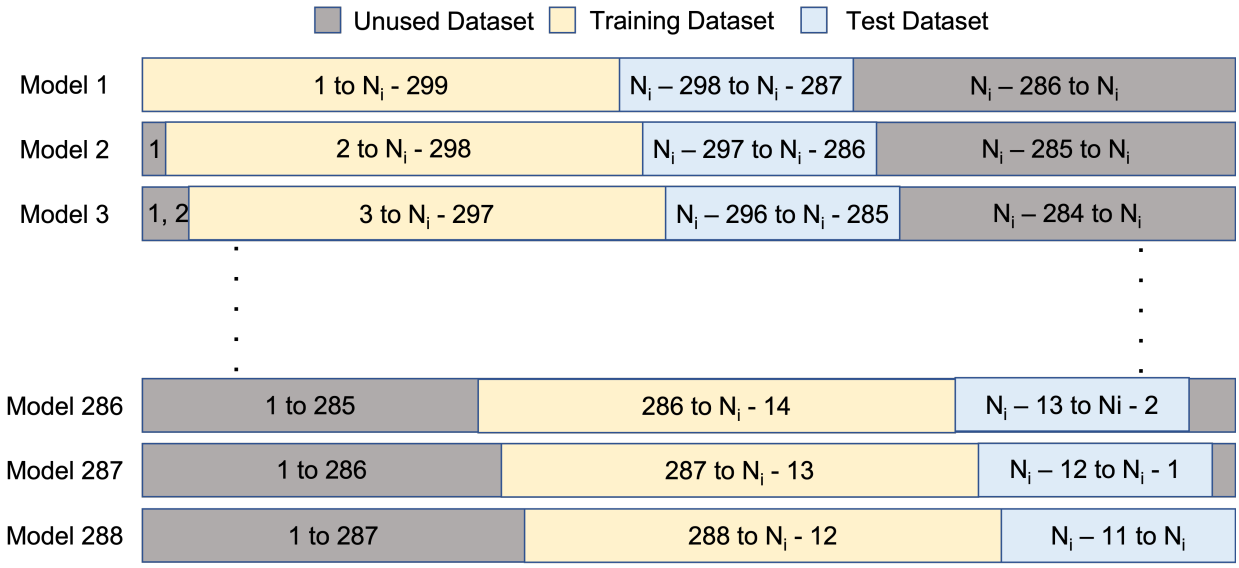

Figure 5.2: The division of data set for training and testing each model for a patient i (with N_i records).

5.2 Test Dataset

The same moving window approach was implemented to generate the predictions. From each model, 1 set of prediction was made for each patient for the 12 records immediately next to the last record used for training from the patient's data set. Thus, a total of 288 sets of predictions of 12 time-points each was made for every patient. To do this the training glucose sequence for individual patient, as shown in Figure 5.2, was passed to the user-defined predict function and a prediction for next 12 time-points was generated. The size of predictions from every population-level model has been shown in Equation 5.3.

$$Total\ length\ of\ predictions\ from\ every\ model\ =\ 20\ *\ 12 \tag{5.3}$$

5.3 Model Interpretation

As it can be seen in Table 4.5 the patient specific models better captures the true nature of distributions instead of approximating it for a group. Hence, when predicting a latent state the patient specific model is better able to explain the physiological state of a patient. The sensitivity, specificity, and AUC values described in Table 6.2 and Table 6.3 compared to the values described in Table 5.1 show us that the population-level model has a better predictability for hypoglycemia compared to the patient-specific model. At the same time, the high false alert rates (as shown in Figure 5.1) is a result of the unbalanced nature of the data set. Out of the total test data set of the 20 patients, only 1.77% glucose observations were below the hypoglycemic threshold of 70 mg/dL. One patient has been left out in the calculation of the data represented in Table 5.1, as the patient's test data set had no hypoglycemic instances and out of 288 predictions 2 were False Positives and 286 were True Negatives.

[1]Prediction Threshold = 0.01

[2]Number of values under 70 mg/dL in the test data (1 day for each patient)

Table 5.1: Analysis of output for the 19 patient specific models for 30 minute ahead prediction.

Metric	Mean $\pm$ Standard Deviation	Median	Range
Sensitivity(%)[1]	67.15 ± 35.22	80.95	0.00 - 100.00
Specificity(%)[1]	80.25 ± 16.16	83.57	40.30 - 100.00
AUC	0.8569 ± 0.0818	0.8504	0.7112 - 0.9926
Hypoglycemia Count[2]	5.37 ± 9.29	0.00	0.00 - 36.00

6. ANALYSIS OF THE OUTPUT

6.1 Evaluation Metrics

6.1.1 Threshold Values

If the predicted probability of hypoglycemia was greater than or equal to the selected threshold value then the patient is predicted to be in a state of hypoglycemia. The focus of this effort is identification of hypoglycemia. But due to class imbalance (typically 2.42% hypoglycemia values), the threshold value for identifying hypoglycemic event will need to be adjusted for better sensitivity. Due to these reasons, the use of a threshold to predict a hypoglycemic event gives better and accurate predictions.

6.1.2 Confusion Matrix

Using the above described threshold values, confusion matrix was defined for a state of hypoglycemia. Table 6.1 gives the generic view of a confusion matrix.

Table 6.1: Generic view of the confusion matrix in tabular form.

		True state		
		Positive	Negative	Total
Predicted state	Positive	TP	FP	$TP + FP$
	Negative	FN	TN	$FN + TN$
	Total	$TP + FN$	$FP + TN$	

The confusion matrix was used to calculate the true positive rates and the false positive rates for various threshold values which were then used to develop the ROC curves at each time point (or every 5 minutes). Mathematically, true positive rate and false positive rates are defined as follows:

$$\begin{aligned} True\ Positive\ Rate &= \frac{TP}{TP + FN} \\ &= Sensitivity \end{aligned} \tag{6.1}$$

$$
\begin{aligned}
False\ Positive\ Rate &= \frac{FP}{TN + FP} \\
&= 1 - Specificity
\end{aligned}
\tag{6.2}
$$

With the help of Table 6.1 and Equations 6.1 and 6.2, we determined the sensitivity and specificity values for specific thresholds, one of which for the population-level model is shown in Table 6.2 along with the values available in the literature for various models used for the same application for 30 minute prediction horizon. A extended version of this table showing sensitivity and specificity at various prediction horizons is shown in Table B.1. Table 5.1 shows the approximate distribution and range of the sensitivity and specificity of various patient-specific models.

Table 6.2: Sensitivity and specificity of various models presented in the literature for a 30 minute prediction horizon.

Source	Model	Sensitivity (%)	Specificity (%)
	HSMM [1,2]	91.35	75.03
	HSMM - Day Time [1,2]	89.11	73.56
	HSMM - Night Time [1,2]	95.34	78.19
Dave et al. [5]	LASSO Optimized LR [2]	73.75	94.87
	VIP Optimized RF [2]	90.93	93.65
	RF - Day [2]	88.43	92.9
	RF - Night [2]	94.92	95.85
Palerm et al. [29]	Kalman Filter-Based Approach [2]	90	79
Berikov et al. [30]	RF (NS) [2]	87.1	87.1
	LogRLasso (NS) [2]	87.1	90.8
	ANN (NS) [2]	86.6	88.7

[1] Prediction Threshold = 0.01
[2] At Prediction Horizon

6.1.3 ROC Curve

The receiver operating characteristic curve (or ROC curve) helps us to identify a threshold which helps in improving the prediction of a model and it also helps define the classification efficiency of a model. For the population-level model, Figure 6.2 gives us the area under the ROC curves (along with their 95% confidence interval) for each time point (up to 60 minutes) and Figure 6.1 gives us the ROC curve for 30 minute ahead prediction along with it's 95% confidence interval. Similarly, for day time predictions (06:00 to 21:59) ROC and area under the ROC curve plots are shown in Figures C.2 and C.3, respectively and for night time predictions (22:00 to 05:59) ROC and area under the ROC curve plots are shown in Figures C.4 and C.5, respectively. Table 6.3 gives area under the curve (overall, day time, and night time) for the 30 minute ahead prediction ROC curve along with the values available in the literature for various models used for the same application for 30 minute prediction horizon. A extended version of this table showing the area under the curve for ROC curves at various prediction horizons is shown in Table B.2.

Table 6.3: AUC of ROC curves of various models presented in the literature for a 30 minute prediction horizon.

Source	Model	AUC
	HSMM [1,2]	0.9035
	HSMM - Day Time [1,2]	0.8890
	HSMM - Night Time [1,2]	0.9277
Mo et al. [31]	RELM [2]	0.74
	ELM [2]	0.731
Berikov et al. [30]	RF (NS) [2]	0.92
	LogRLasso (NS) [2]	0.928
	ANN (NS) [2]	0.924

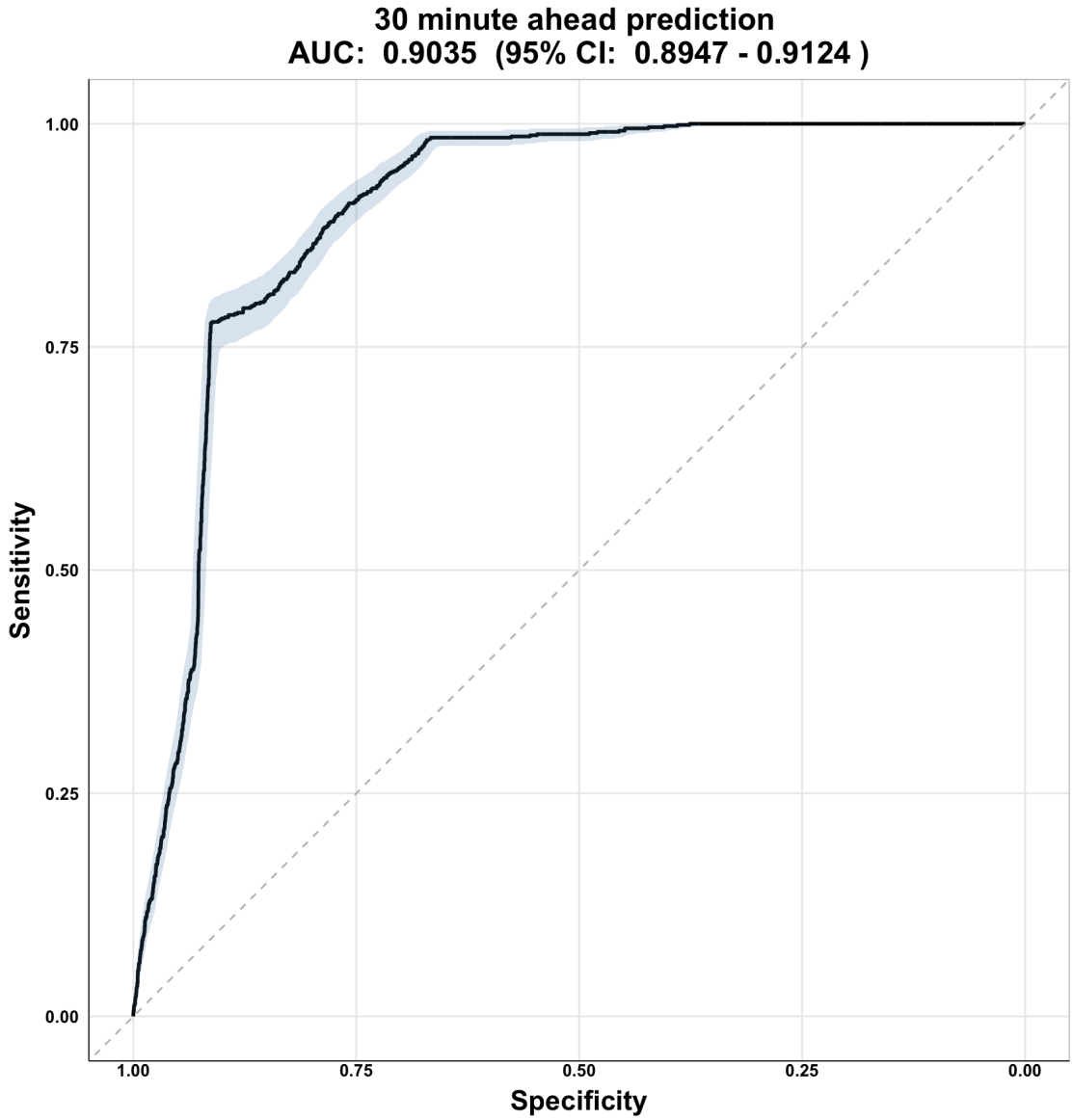

Figure 6.1: The ROC curve of the population-level model at time point 6 (30 minutes ahead prediction).

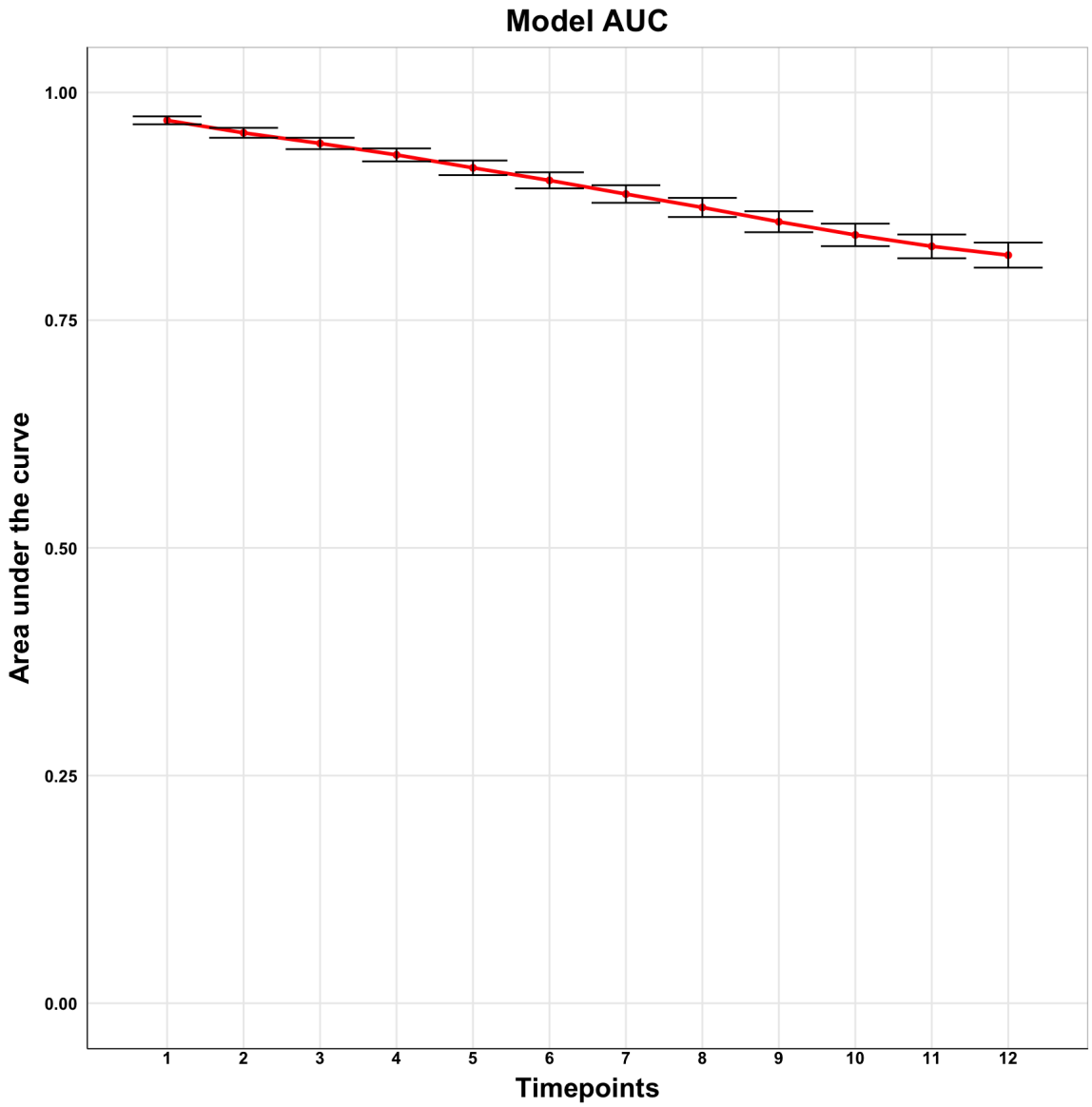

Figure 6.2: The AUC of the population-level model at all time points for a prediction threshold of 0.05 along with its 95% confidence intervals.

7. SUMMARY

The developed Hidden semi-Markov model showed comparable performance and better explanability compared to the other Machine Learning models available in the literature for hypoglycemia prediction. The population-level model showed better predictability whereas the patient-level model showed better explanability of the latent states. The population-level model better showed the trend of a generic patient with type 1 diabetes whereas patient-level model better showed the trends of individual patients and the impact of their circadian cycles. Carbohydrate to insulin ratios are another group of important variables that help explain the predicted outputs and have a significant impact on the glycemic thrends and transitions of the latent states. The population-level model sensitivity and specificity for 30 minute ahead forecast using a specific threshold was 93.03% and 72.50% whereas the same for 60 minute ahead forecast using the same threshold was 89.76% and 65.27%, respectively. The ROC-AUC values greater than 0.9 for 30 minute ahead forecast and greater than 0.82 for 60 minute ahead forecast show the efficiency of the model for separating positive hypoglycemic and negative hypoglycemic events.

7.1 Future Work

A larger group of patients and larger individual datasets will provide better opportunities to understand the patient-level model which is now restricted due to fewer individual hypoglycemic events. This study only uses each patient's glucose sequence while developing the model, including carbohydrate, basal insulin, and bolus insulin information as covariates is expected to give better results. We plan to incorporate carbohydrate to insulin ratios by using the explanations and calculation methodology explained by Oerum [32], Oerum [33], and on OpenAPS [34] to explain state transitions and in the model validation process. The analysis of these ratios and their applicability as per the model output was carried out by another member of the team. Similarly, incorporating type of activity, sleep cycle, and time of day information will help to better predict special cases that arise majorly due to circadian cycles and individual patterns.

REFERENCES

[1] D. J. Handelsman and J. R. Turtle, "Diabetes Control: A Changing Scene," *The Medical Journal of Australia*, vol. 1, no. 13, pp. 607–611, 1979.

[2] D. V. Ary, D. Toobert, W. Wilson, and R. E. Glasgow, "Patient Perspective on Factors Contributing to Nonadherence to Diabetes Regimen," *Diabetes Care*, vol. 9, no. 2, pp. 168–172, 1986.

[3] D. Dave, M. Erraguntla, M. Lawley, D. DeSalvo, B. Haridas, S. McKay, C. Koh, *et al.*, "Improved Low-Glucose Predictive Alerts Based on Sustained Hypoglycemia: Model Development and Validation Study," *JMIR Diabetes*, vol. 6, no. 2, p. e26909, 2021.

[4] T. Bremer and D. A. Gough, "Is blood glucose predictable from previous values? A solicitation for data.," *Diabetes*, vol. 48, no. 3, pp. 445–451, 1999.

[5] D. Dave, D. J. DeSalvo, B. Haridas, S. McKay, A. Shenoy, C. J. Koh, M. Lawley, and M. Erraguntla, "Feature-Based Machine Learning Model for Real-Time Hypoglycemia Prediction," *Journal of Diabetes Science and Technology*, vol. 15, no. 4, pp. 842–855, 2021.

[6] V. Naumova, S. V. Pereverzyev, and S. Sivananthan, "A meta-learning approach to the regularized learning—Case study: Blood glucose prediction," *Neural Networks*, vol. 33, pp. 181–193, 2012.

[7] M. Eren-Oruklu, A. Cinar, and L. Quinn, "Hypoglycemia Prediction with Subject-Specific Recursive Time-Series Models," *Journal of Diabetes Science and Technology*, vol. 4, no. 1, pp. 25–33, 2010.

[8] M. Eren-Oruklu, A. Cinar, L. Quinn, and D. Smith, "Estimation of Future Glucose Concentrations with Subject-Specific Recursive Linear Models," *Diabetes Technology & Therapeutics*, vol. 11, no. 4, pp. 243–253, 2009.

[9] G. Sparacino, F. Zanderigo, S. Corazza, A. Maran, A. Facchinetti, and C. Cobelli, "Glucose Concentration can be Predicted Ahead in Time From Continuous Glucose Monitoring Sensor Time-Series," *IEEE Transactions on Biomedical Engineering*, vol. 54, no. 5, pp. 931–937, 2007.

[10] H. T. Abbas, L. Alic, M. Erraguntla, J. X. Ji, M. Abdul-Ghani, Q. H. Abbasi, and M. K. Qaraqe, "Predicting long-term type 2 diabetes with support vector machine using oral glucose tolerance test," *PLoS ONE*, vol. 14, no. 12, p. e0219636, 2019.

[11] T. Hamdi, J. B. Ali, V. Di Costanzo, F. Fnaiech, E. Moreau, and J.-M. Ginoux, "Accurate prediction of continuous blood glucose based on support vector regression and differential evolution algorithm," *Biocybernetics and Biomedical Engineering*, vol. 38, no. 2, pp. 362–372, 2018.

[12] B. W. Bequette, "Continuous Glucose Monitoring: Real-Time Algorithms for Calibration, Filtering, and Alarms," *Journal of Diabetes Science and Technology*, vol. 4, no. 2, pp. 404–418, 2010.

[13] K. Li, C. Liu, T. Zhu, P. Herrero, and P. Georgiou, "GluNet: A Deep Learning Framework for Accurate Glucose Forecasting," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 2, pp. 414–423, 2019.

[14] C. Pérez-Gandía, A. Facchinetti, G. Sparacino, C. Cobelli, E. Gómez, M. Rigla, A. de Leiva, and M. Hernando, "Artificial Neural Network Algorithm for Online Glucose Prediction from Continuous Glucose Monitoring," *Diabetes Technology & Therapeutics*, vol. 12, no. 1, pp. 81–88, 2010.

[15] K. Turksoy, E. S. Bayrak, L. Quinn, E. Littlejohn, D. Rollins, and A. Cinar, "Hypoglycemia Early Alarm Systems Based on Multivariable Models," *Industrial & Engineering Chemistry Research*, vol. 52, no. 35, pp. 12329–12336, 2013.

[16] G. Cappon, M. Vettoretti, F. Marturano, A. Facchinetti, and G. Sparacino, "A Neural-Network-Based Approach to Personalize Insulin Bolus Calculation Using Continuous Glu-

cose Monitoring," *Journal of Diabetes Science and Technology*, vol. 12, no. 2, pp. 265–272, 2018.

[17] L. H. Messer, P. Calhoun, B. Buckingham, D. M. Wilson, I. Hramiak, T. T. Ly, M. Driscoll, P. Clinton, D. M. Maahs, and I. H. C. L. S. Group, "In-home nighttime predictive low glucose suspend experience in children and adults with type 1 diabetes," *Pediatric Diabetes*, vol. 18, no. 5, pp. 332–339, 2017.

[18] C. Zecchin, A. Facchinetti, G. Sparacino, G. De Nicolao, and C. Cobelli, "Neural Network Incorporating Meal Information Improves Accuracy of Short-Time Prediction of Glucose Concentration," *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 6, pp. 1550–1560, 2012.

[19] R. Hovorka, V. Canonico, L. J. Chassin, U. Haueter, M. Massi-Benedetti, M. O. Federici, T. R. Pieber, H. C. Schaller, L. Schaupp, T. Vering, *et al.*, "Nonlinear model predictive control of glucose concentration in subjects with type 1 diabetes," *Physiological Measurement*, vol. 25, no. 4, pp. 905–920, 2004.

[20] C. Zecchin, A. Facchinetti, G. Sparacino, and C. Cobelli, "Jump neural network for online short-time prediction of blood glucose from continuous monitoring sensors and meal information," *Computer Methods and Programs in Biomedicine*, vol. 113, no. 1, pp. 144–152, 2014.

[21] H. ElMoaqet, D. M. Tilbury, and S. K. Ramachandran, "Multi-Step Ahead Predictions for Critical Levels in Physiological Time Series," *IEEE Transactions on Cybernetics*, vol. 46, no. 7, pp. 1704–1714, 2016.

[22] J. O'Connell and S. Højsgaard, "Hidden Semi Markov Models for Multiple Observation Sequences: The mhsmm Package for R," *Journal of Statistical Software*, vol. 39, no. 4, pp. 1–22, 2011.

[23] C. Godin and Y. Guédon, "AMAPmod Introduction and Reference Manual Version 1.8." `https://hal.inria.fr/hal-00827487/document`. Accessed: 2023-02-06.

[24] J. Bulla, I. Bulla, and O. Nenadić, "hsmm—An R package for analyzing hidden semi-Markov models," *Computational Statistics & Data Analysis*, vol. 54, no. 3, pp. 611–619, 2010.

[25] Yu, Shun-Zheng, *Hidden Semi-Markov Models: Theory, Algorithms and Applications*. Morgan Kaufmann, 2015.

[26] W. Zucchini, I. L. MacDonald, and R. Langrock, *Hidden Markov Models for Time Series: An Introduction Using R*. CRC Press, 2017.

[27] Y. Guédon, "Estimating Hidden Semi-Markov Chains From Discrete Sequences," *Journal of Computational and Graphical Statistics*, vol. 12, no. 3, pp. 604–639, 2003.

[28] L. R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[29] C. C. Palerm and B. W. Bequette, "Hypoglycemia Detection and Prediction Using Continuous Glucose Monitoring—A Study on Hypoglycemic Clamp Data," *Journal of Diabetes Science and Technology*, vol. 1, no. 5, pp. 624–629, 2007.

[30] V. B. Berikov, O. A. Kutnenko, J. F. Semenova, and V. V. Klimontov, "Machine Learning Models for Nocturnal Hypoglycemia Prediction in Hospitalized Patients with Type 1 Diabetes," *Journal of Personalized Medicine*, vol. 12, no. 8, p. 1262, 2022.

[31] X. Mo, Y. Wang, and X. Wu, "Hypoglycemia Prediction Using Extreme Learning Machine (ELM) and Regularized ELM," in *2013 25th Chinese Control and Decision Conference (CCDC)*, pp. 4405–4409, IEEE, 2013.

[32] C. Oerum, "Insulin Types: Their Peak Times and Durations." `https://diabetesstrong.com/insulin-types/`. Accessed: 2023-02-23.

[33] C. Oerum, "How to Calculate Active Insulin On Board (IOB)." `https://diabetesstrong.com/how-to-calculate-active-insulin-on-board/`. Accessed: 2023-02-23.

[34] Unknown, "Understanding Insulin on Board (IOB) Calculations." https://openaps.readthedocs.io/en/latest/docs/While%20You%20Wait%20For%20Gear/understanding-insulin-on-board-calculations.html. Accessed: 2023-02-23.

[35] T. Zhu, C. Uduku, K. Li, P. Herrero, N. Oliver, and P. Georgiou, "Enhancing self-management in type 1 diabetes with wearables and deep learning," *npj | Digital Medicine*, vol. 5, no. 1, p. 78, 2022.

[36] S. Oviedo, I. Contreras, C. Quiros, M. Giménez, I. Conget, and J. Vehi, "Risk-based postprandial hypoglycemia forecasting using supervised learning," *International Journal of Medical Informatics*, vol. 126, pp. 1–8, 2019.

[37] J. Yang, L. Li, Y. Shi, and X. Xie, "An ARIMA Model With Adaptive Orders for Predicting Blood Glucose Concentrations and Hypoglycemia," *IEEE Journal of Biomedical and Health Informatics*, vol. 23, no. 3, pp. 1251–1260, 2018.

[38] M. H. Jensen, C. Dethlefsen, P. Vestergaard, and O. Hejlesen, "Prediction of Nocturnal Hypoglycemia From Continuous Glucose Monitoring Data in People With Type 1 Diabetes: A Proof-of-Concept Study," *Journal of Diabetes Science and Technology*, vol. 14, no. 2, pp. 250–256, 2020.

[39] J. Vehí, I. Contreras, S. Oviedo, L. Biagi, and A. Bertachi, "Prediction and prevention of hypoglycaemic events in type-1 diabetic patients using machine learning," *Health Informatics Journal*, vol. 26, no. 1, pp. 703–718, 2020.

APPENDIX A

URL FOR CODE

A.1 GitHub Repository

https://github.com/mohit-chhaparia/Health-Behavior-Inference-from-Continuous-Blood-Glucose-Data-A-Hidden-Semi-Markov-Approach

A.2 Population - Level Model

https://github.com/mohit-chhaparia/Health-Behavior-Inference-from-Continuous-Blood-Glucose-Data-A-Hidden-Semi-Markov-Approach/blob/main/Population%20Level%20Model.R

A.3 Patient - Specific Model

https://github.com/mohit-chhaparia/Health-Behavior-Inference-from-Continuous-Blood-Glucose-Data-A-Hidden-Semi-Markov-Approach/blob/main/Patient%20Specific%20Model.R

APPENDIX B

TABLES

Table B.1: Sensitivity and specificity of various models presented in the literature for various prediction horizons.

Source	Model	PH (min)	Sensitivity (%)	Specificity (%)
Dave et al. [5]	HSMM [1,2]	15 / 30 / 45 / 60	89.66 / 91.35 / 94.22 / 89.76	91.20 / 75.03 / 65.99 / 65.27
	HSMM - Day Time [1,2]	15 / 30 / 45 / 60	86.69 / 89.11 / 91.94 / 86.41	90.22 / 73.56 / 63.55 / 62.61
	HSMM - Night Time [1,2]	15 / 30 / 45 / 60	94.96 / 95.34 / 98.23 / 95.49	93.32 / 78.19 / 71.25 / 71.01
	LASSO Optimized LR [2]	15 / 30 / 45 / 60	91.85 / 73.75 / 55.06 / 43.28	96.25 / 94.87 / 95.5 / 95.25
	VIP Optimized RF [2]	15 / 30 / 45 / 60	94.2 / 90.93 / 88.04 / 86.28	96.67 / 93.65 / 92.68 / 93.07
	RF - Day [2]	15 / 30 / 45 / 60	93.08 / 88.43 / 84.1 / 82.92	96.25 / 92.9 / 91.96 / 92.97
	RF - Night [2]	15 / 30 / 45 / 60	96.18 / 94.92 / 94.77 / 93.85	97.57 / 95.85 / 94.44 / 93.97
Eren-Oruklu et al. [7]	SSRTSM: Absolute Predicted Glucose Values [3]	30 ± 5.51	89	67
	SSRTSM: Cumulative-Sum Control Chart [3]	25.8 ± 6.46	87.5	74

[1]Prediction Threshold = 0.01
[2]At Prediction Horizon
[3]Mean value for time to detection

Continuation of Table B.1				
Source	Model	PH (min)	Sensitivity (%)	Specificity (%)
	SSRTSM: Exponentially Weighted Moving-Average Control Chart [3]	27.7 ± 5.32	89	78
Palerm et al. [29]	Kalman Filter-Based Approach [2]	30	90	79
Berikov et al. [30]	RF (NS) [2]	15 / 30	91.8 / 87.1	91.1 / 87.1
	LogRLasso (NS) [2]	15 / 30	93.6 / 87.1	91.2 / 90.8
	ANN (NS) [2]	15 / 30	88.6 / 86.6	92.6 / 88.7
Zhu et al. [35]	ARISES with an embedded DL algorithm [4]	60	70.3	
Oviedo et al. [36]	SVM [4]	240	71	79
Yang et al. [37]	ARIMA with Adaptive Identification Algorithm [5]	24.8	100	FAR - 10.7
Jensen et al. [38]	Linear Discriminant Analysis [6]	195	75	70
Vehí et al. [39]	SVM [4]	240	69	80
	ANN [4]	360	44	85.9
End of Table B.1				

[4] Over / Within Prediction Horizon
[5] Average time for action
[6] At Prediction Horizon (average)

Table B.2: AUC of various models presented in the literature at various prediction horizons.

Source	Model	PH (min)	AUC
	HSMM [1,2]	15 / 30 / 45 / 60	0.9441 / 0.9035 / 0.8581 / 0.8214
	HSMM - Day Time [1,2]	15 / 30 / 45 / 60	0.9389 / 0.8890 / 0.8321 / 0.7867
	HSMM - Night Time [1,2]	15 / 30 / 45 / 60	0.9521 / 0.9277 / 0.9025 / 0.8775
Mo et al. [31]	RELM [2]	10 / 20 / 30	0.932 / 0.838 / 0.74
	ELM [2]	10 / 20 / 30	0.935 / 0.817 / 0.731
Berikov et al. [30]	RF (NS) [2]	15 / 30	0.959 / 0.92
	LogRLasso (NS) [2]	15 / 30	0.957 / 0.928
	ANN (NS) [2]	15 / 30	0.934 / 0.924
End of Table B.2			

APPENDIX C

FIGURES

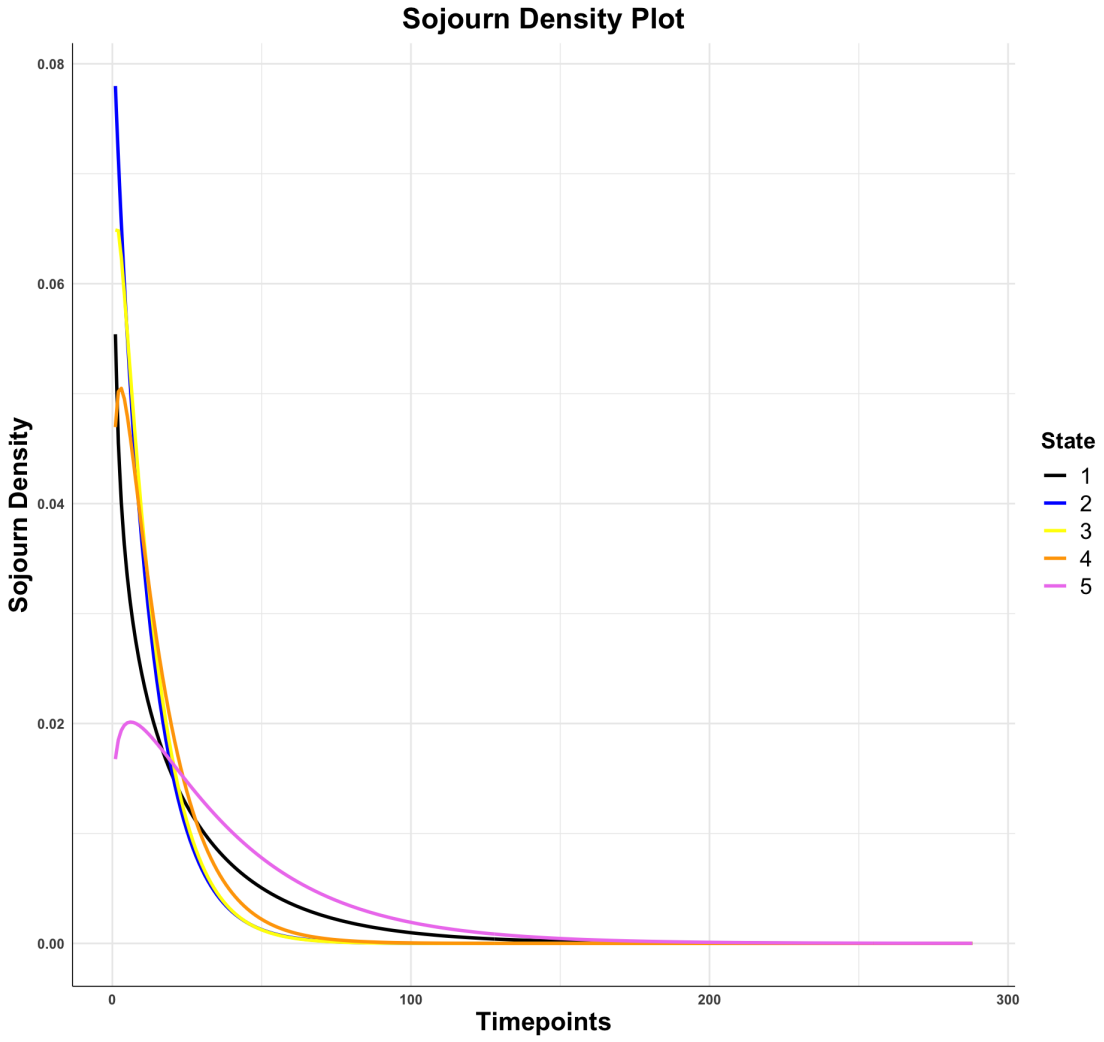

Figure C.1: Sojourn density plot of time points based on the trained population-level model for all the latent states.

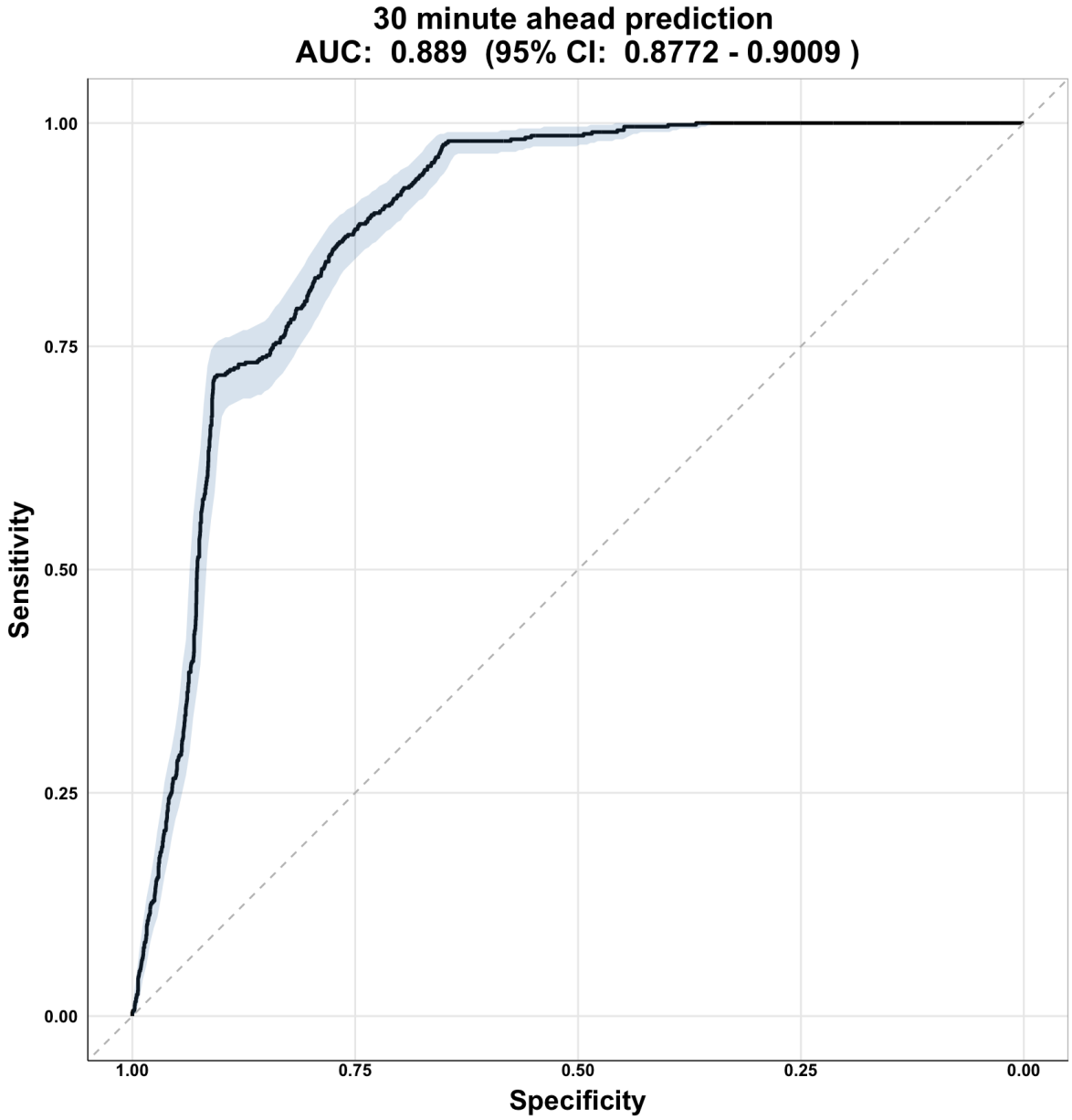

Figure C.2: The ROC curve of the population-level model at time point 6 (30 minutes ahead prediction) for day time predictions (06:00 to 21:59).

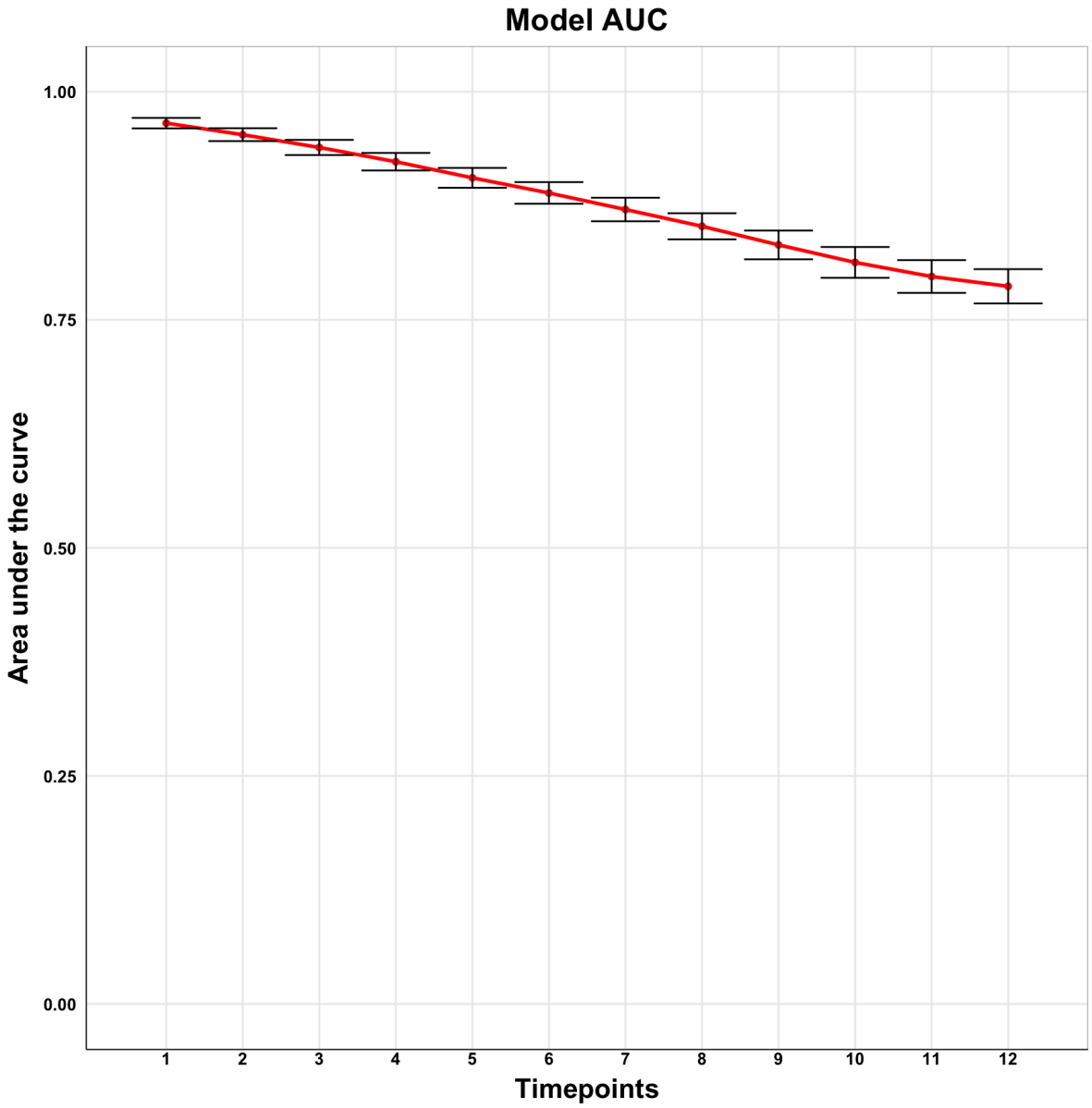

Figure C.3: The AUC of the population-level model at day time (06:00 to 21:59) for a prediction threshold of 0.05 along with its 95% confidence intervals.

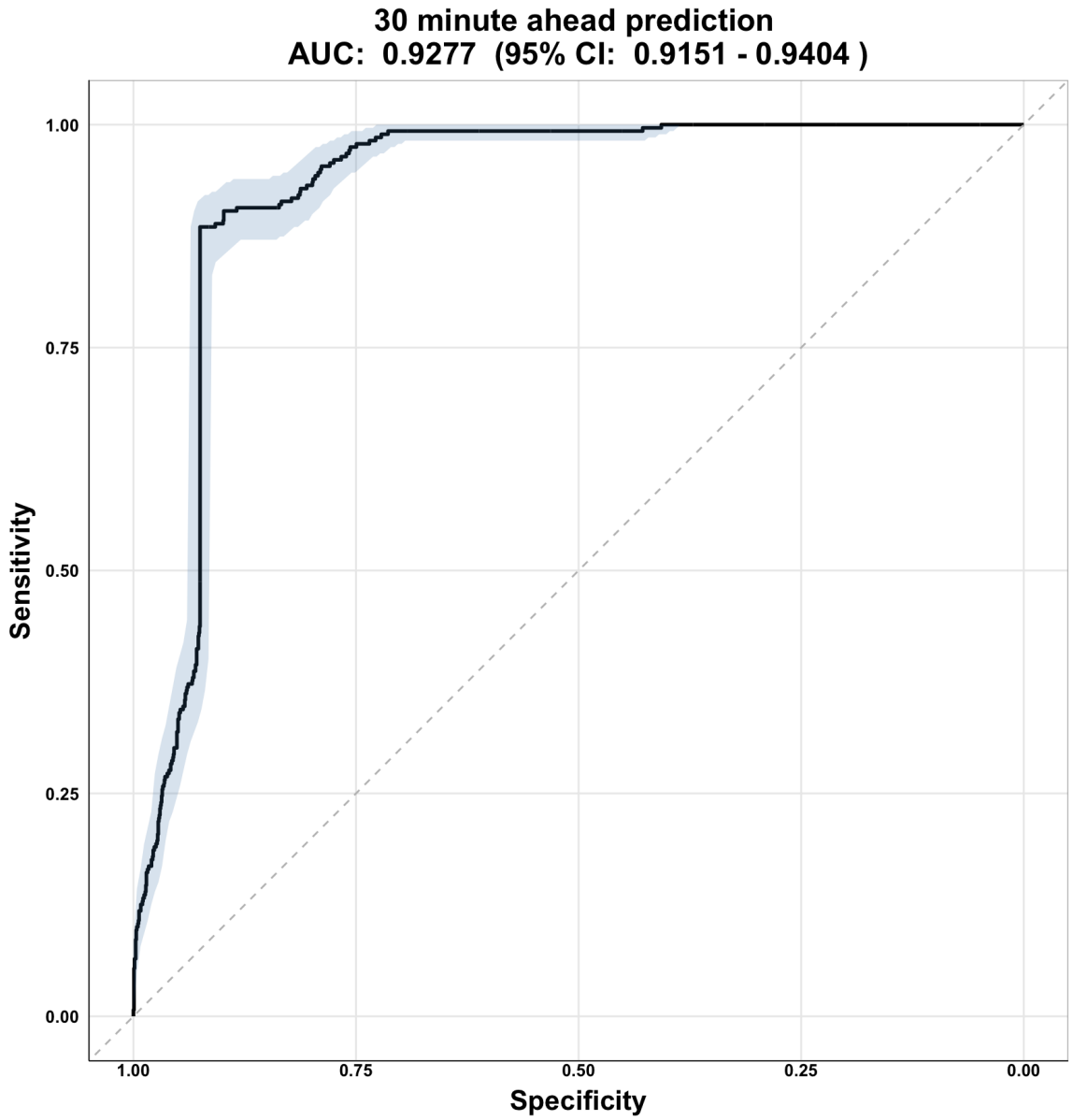

Figure C.4: The ROC curve of the population-level model at time point 6 (30 minutes ahead prediction) for night time predictions (22:00 to 05:59).

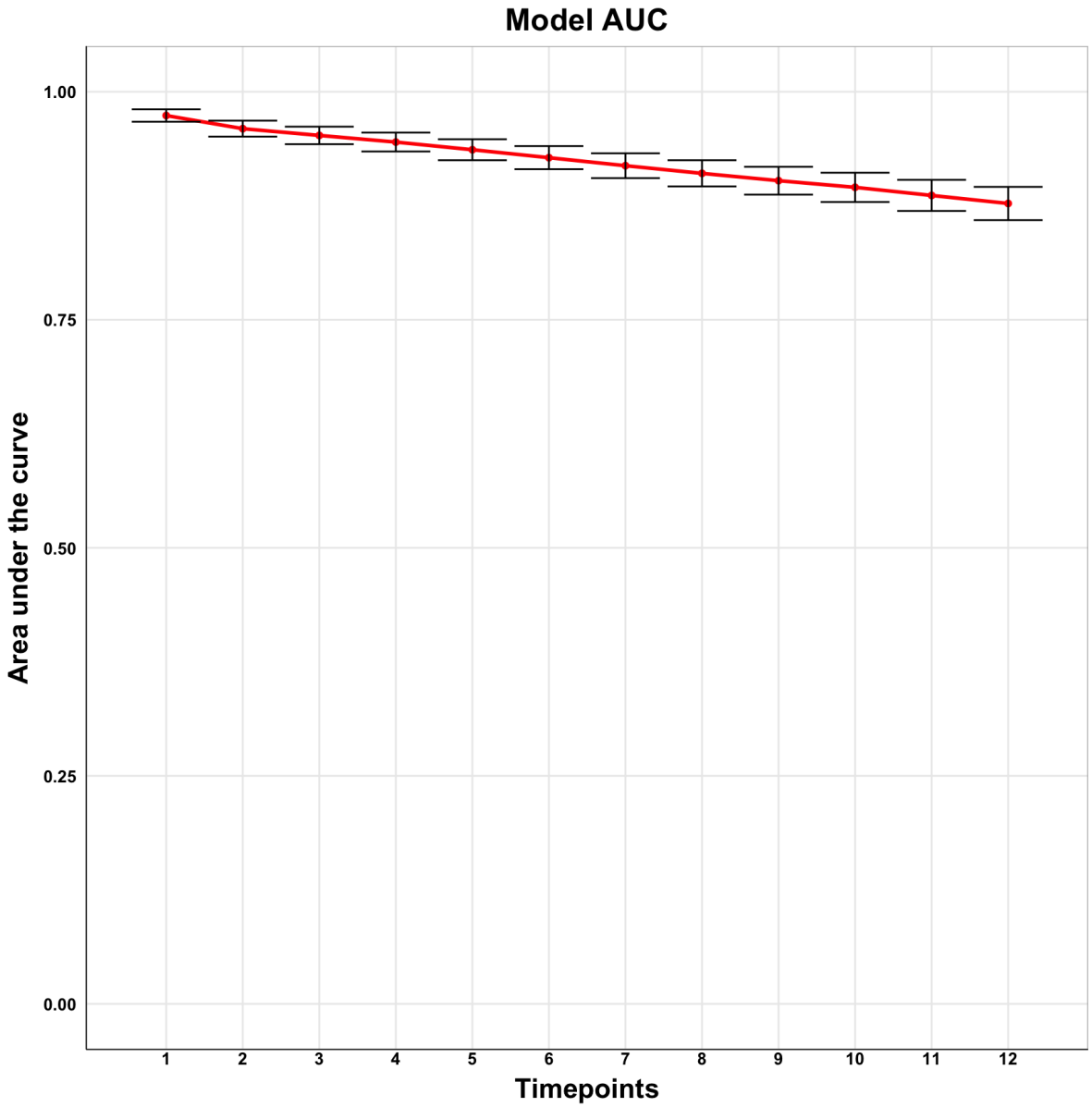

Figure C.5: The AUC of the population-level model at night time (22:00 to 05:59) for a prediction threshold of 0.05 along with its 95% confidence intervals.