Advanced Artificial Intelligence and Machine Learning Driven Data Analyses in Diabetes Mellitus Research

By:

Heydar Khadem

Electronic and Electrical Engineering Department

University of Sheffield

PhD Dissertation

2024

Epigraph

"Data is the oil of the 21st century, and analytics is the combustion engine."

Peter Sondergaard

Acknowledgements

First of all, I would like to express my deepest gratitude to my principal supervisor, Professor Mohammed Benaissa, and my second supervisor, Dr Jackie Elliott, for their unwavering academic and pastoral support during my PhD. Their guidance has been invaluable.

In addition, I would like to extend my sincere appreciation to my esteemed colleagues and coauthors, including Dr. Hoda Nemat, Dr. Mohammad R. Eissa, and Dr. Osamah Alrezj. The intellectual contributions and camaraderie we've shared throughout this journey have been of immeasurable value. Their insights have helped to sharpen my research and enrich my understanding.

I also wish to acknowledge the often unseen yet essential work of the administrative and technical staff. Their tireless efforts have facilitated a seamless progression of my studies, and for this, I am deeply grateful.

Lastly, my gratitude extends to the entire University of Sheffield community. The stimulating academic environment, coupled with a plethora of comprehensive resources, has significantly enhanced my research journey. This institution has provided me with more than just an education, it has been a pivotal part of my growth both academically and personally. My time at the University of Sheffield has been a profoundly enriching experience.

Abstract

Diabetes mellitus is an endocrine disorder of global significance. This PhD research project harnesses the capabilities of machine learning techniques for the meticulous investigation of data associated with diabetes mellitus. The formidable prevalence of the disease across the world, alongside the consequential burdens it imposes on healthcare systems, underscores the paramount importance of this research. The research is characteristically subdivided into three focal domains: predictive analysis, glucose quantification, and risk assessment relating to diabetes. Specifically, the research delves into advanced deep learning architectures for forecasting blood glucose levels, proposes methodologies for improving glucose quantification, and provides thorough risk assessments for COVID-19 patients with pre-existing diabetes mellitus. For each of these three domains, the research deploys state-of-the-art machine learning algorithms as a powerful apparatus to navigate the complexities of diabetes data, culminating in two research publications in reputable, peer-reviewed academic journals. Each publication illuminates the transformative potential of machine learning as a conduit for novel advancements within the respective domain. This in turn contributes to a more nuanced understanding of the disease, enhancement of patient care, and optimisation of healthcare resource allocation. Composed in a publication format, this dissertation is structured as a compilation of the six resultant articles, which are interconnected within the overarching framework of machine learning applications in diabetes research. As a whole, this extensive exploration of diabetes data through machine learning pipelines proffers novel insights and aims to make a substantial contribution to the academic field.

Commentary

In adherence to the University of Sheffield's code of practice, this dissertation adopts a publication format. According to the university's guidelines, a dissertation in this format commences with a commentary, such as this one, which distils the central themes of the PhD research project and its corresponding publication outputs. Following this commentary, a compilation of research publications forms the primary body of the dissertation. The cornerstone of this PhD research project is the analysis of diabetes mellitus data using advanced machine learning techniques.

Diabetes mellitus is a chronic disease affecting millions globally and significantly burdening healthcare systems. Diabetes complications can lead to severe health consequences, including cardiovascular disease, kidney failure, amputations and blindness. Given the global impact of this disease, there is a pressing need for innovative research and solutions to improve patient outcomes and reduce the financial impact on healthcare systems.

Machine learning, with its ability to learn from complex data and adapt, has proven essential in addressing key healthcare challenges. As the volume and complexity of healthcare data continue to grow, machine learning techniques are becoming increasingly important for the analysis and interpretation of this wealth of data. In recent years, the potential of machine learning to address various challenges in diabetes research has gained considerable attention. The surge in the availability of healthcare data, advancements in computational power, and the development of sophisticated machine learning algorithms have provided researchers with new opportunities to tackle diabetes-related challenges, enhance our understanding of the disease, and ultimately improve patient care.

Among the significant contributions of machine learning in diabetes research are its applications in three primary areas: blood glucose level prediction, glucose quantification, and diabetes-related risk assessment. In each of these areas, machine learning techniques have shown promise in uncovering novel insights, guiding clinical decision-making, and ultimately improving patient outcomes. The ongoing

advancements in these fields continue to highlight the potential of machine learning in addressing challenges in diabetes research and contributing to the development of innovative solutions for better disease management and care.

Blood glucose level prediction plays a crucial role in diabetes management. Maintaining optimal glucose levels is essential for preventing acute complications, such as hypoglycemia and hyperglycemia, as well as reducing the risk of long-term complications. Machine learning algorithms have shown great potential in accurately predicting blood glucose levels, enabling patients and healthcare professionals to make better-informed decisions regarding insulin administration, meal planning, and physical activity. This, in turn, can lead to improved glycemic control and an enhanced quality of life for individuals with diabetes.

Glucose quantification is another critical area of diabetes research that has benefited from the application of machine learning techniques. Traditional glucose monitoring methods can be painful, inconvenient, and disruptive to daily life. Spectroscopic alternatives, on the other hand, offer the possibility of continuous, real-time monitoring without the need for blood samples. Machine learning has been employed in the development of advanced signal processing and data analysis techniques for various quantitative analyses of glucose technologies, ultimately aiming to improve the accuracy, reliability, and user-friendliness of these devices.

Assessing diabetes-related risks is essential for early detection, prevention, and management of the disease and its complications. Machine learning algorithms have been applied to identify patterns and relationships in large, complex datasets, which can help uncover risk factors and enable more accurate, personalised risk prediction. This information can be used by healthcare professionals to stratify patients according to their risk levels and allocate resources more effectively. Tailoring interventions to the specific needs of individual patients becomes possible. Ultimately, this leads to better health outcomes and more efficient healthcare systems.

Overall, the application of machine learning in diabetes research holds great promise for advancing our understanding of the disease, improving patient outcomes, and optimising healthcare resource allocation. The current PhD research contributes to this endeavour by developing advanced machine learning techniques focused on the three areas discussed above—blood glucose level prediction, glucose quantification, and diabetes-related risk assessment. In each of these areas, this research has led to two publications in reputable peer-reviewed journals, synopses of which are provided in the subsequent paragraphs.

The first publication is entitled "*Blood Glucose Time Series Forecasting: Advanced GAN Driven Interdependent Deep Learning Topologies*". This work introduces novel deep learning configurations inspired by Generative Adversarial Networks (GANs) to address the complexities of time series forecasting for blood glucose prediction. The paper proposes interdependent deep learning topologies, which significantly enhance both the accuracy and robustness of blood glucose forecasts. Key findings demonstrate that these GAN-based topologies outperform traditional independent methods in predictive performance in blood glucose data. This work provides a new pathway for improving real-time diabetes management by delivering more reliable predictions, thus aiding clinical decision-making. At the time of submitting this dissertation, the article has been submitted to *IEEE Journal of Biomedical Health Informatics*. [1]

The second publication, titled *"Deep Learning Blood Glucose Level Time Series Forecasting: Advanced Nested Stacking Lag Fusion Framework"*, introduces an innovative deep ensemble learning approach to resolve the challenge of look-back or lag optimisation in blood glucose prediction models. The proposed nested stacking framework integrates multiple lag lengths, significantly enhancing predictive accuracy by leveraging complementary temporal patterns. This work shows that the model's ability to incorporate varied lag lengths improves both short- and long-term glucose level predictions, contributing to more precise and adaptive glucose monitoring solutions. The paper is published in *Bioengineering*. [2]

The third publication, *"Classification Before the Regressions for Improvement in Quantification of Glucose using Absorbance Spectroscopy"*, presents a novel pre-classification method that improves glucose quantification from spectral data. By grouping data into three homogeneous classes (hypoglycaemia, euglycaemia, and hyperglycemia) prior to regression analysis, the study enhances predictive accuracy, especially for lower glucose concentrations. This method significantly outperforms conventional approaches that do not employ pre-classification, providing a more robust tool for non-invasive glucose monitoring. This work is published in *Talanta*. [3]

In the fourth publication, *"Signal Fragmentation Based Feature Vector Generation in a Model-Agnostic Framework for Glucose Quantification Using Absorption Spectroscopy"*, a signal fragmentation approach is proposed to improve glucose quantification from spectroscopic data. By dividing the signal into fragments and processing each independently, the study enhances the accuracy of glucose concentration estimates. This fragmentation method, combined with machine learning techniques, proves effective across various modeling strategies, offering a more flexible and reliable approach to glucose monitoring. This paper is published in *Talanta*. [4]

The fifth publication, *"COVID-19 Mortality Risk Assessments for Individuals with and without Diabetes Mellitus: Machine Learning Models Integrated with Interpretation Framework"*, develops machine learning models to predict mortality risks for hospitalized COVID-19 patients with and without diabetes. Using admission data, the study achieves high prediction accuracy and reveals key clinical features contributing to higher mortality in patients with diabetes. The model's interpretability is enhanced through SHapley Additive exPlanations (SHAP), providing actionable insights for clinicians to assess risk and make more informed decisions in real time. Published in *Computers in Biology and Medicine*. [5]

The sixth publication, *"Interpretable Machine Learning for Inpatient COVID-19 Mortality Risk Assessments: Diabetes Mellitus Exclusive Interplay"*, focuses exclusively on COVID-19 patients with pre-existing diabetes. By utilizing machine learning models integrated with SHAP for interpretation, the study uncovers critical clinical features linked to higher mortality in this population. The insights generated by

this research contribute to improving risk management strategies and help clinicians better understand the specific risks faced by diabetic COVID-19 patients. This work is published in *Sensors*. [6]

To encapsulate, the three areas covered in this PhD research, each with its objectives and publication outputs, coalesce into a cohesive framework anchored by advanced machine learning techniques in diabetes mellitus data analysis. The publications demonstrate a significant contribution to the field through rigorous analysis of diabetic data, resulting in novel machine learning techniques that enhance analysis and offer new insights. These outcomes contribute to a deeper understanding of diabetes management and improving patient health.

References

[1] H. Khadem, H. Nemat, J. Elliott, M. Benaissa, Blood Glucose Time Series Forecasting: Advanced GAN Driven Interdependent Deep Learning Topologies, IEE Journal of Biomedical Health Informative (2024).

[2] H. Khadem, H. Nemat, J. Elliott, M. Benaissa, Blood Glucose Level Time Series Forecasting: Nested Deep Ensemble Learning Lag Fusion, Bioengineering. 10 (2023) 1–22. https://doi.org/10.3390/bioengineering10040487.

[3] H. Khadem, M.R. Eissa, H. Nemat, O. Alrezj, M. Benaissa, Classification before regression for improving the accuracy of glucose quantification using absorption spectroscopy, Talanta. 211 (2020) 1–10. https://doi.org/https://doi.org/10.1016/j.talanta.2020.120740.

[4] H. Khadem, H. Nemat, J. Elliott, M. Benaissa, Signal fragmentation based feature vector generation in a model agnostic framework with application to glucose quantification using absorption spectroscopy, Talanta. 243 (2022) 123379. https://doi.org/10.1016/j.talanta.2022.123379.

[5] H. Khadem, H. Nemat, M.R. Eissa, J. Elliott, M. Benaissa, COVID-19 mortality risk assessments for individuals with and without diabetes mellitus: Machine learning models integrated with interpretation framework, Comput. Biol. Med. 144 (2022) 105361. https://doi.org/10.1016/j.compbiomed.2022.105361.

[6] H. Khadem, H. Nemat, J. Elliott, M. Benaissa, Interpretable Machine Learning for Inpatient COVID-19 Mortality Risk Assessments: Diabetes Mellitus Exclusive Interplay, Sensors. 22 (2022) 8757. https://doi.org/10.3390/s22228757.

Declarations

It is of note that I, Heydar Khadem, the author of the dissertation, am the first author of all papers compiled in this dissertation, and the preponderance of the ideas and workload originated from my efforts. Additionally, I affirm that I have secured the necessary permissions to incorporate the published articles within the dissertation, ensuring full compliance with journal policies on reuse and citation.

Furthermore, this PhD research utilised two private datasets, resourced by the research supervisors, alongside one publicly available dataset. The primary contribution of this dissertation lies in the analysis of data and not data acquisition. Hence, it is explicitly stated that the dissertation does not claim ownership of any of the data used. Consequently, appropriate citations and acknowledgements for the usage of each dataset have been included within the respective articles that utilised them.

Table of Contents

List of Figures

Publication 1

Publication 2

Publication 3

Publication 4

Publication 5

Publication 6

List of Tables

Publication 1

Publication 2

Publication 3

Publication 4

Publication 5

Publication 6

List of Abbreviations

AB	AdaBoost
AL	Adversarial Learning
ALPO4	Alkaline Phosphatase
ALT	Alanine Transaminase
APTT	Activated Partial Thromboplastin Time
ASE	Average Surveillance Error
AUC	Area Under the Curve
BGL	Blood Glucose Level
BGLP	Blood Glucose Level Prediction
BGV	Blood Gas Value
BMI	Body Mass Index
CGM	Continuous Glucose Monitoring
CKD	Chronic Kidney Disease
CL	Collaborative Learning
CLD	Chronic Liver Disease
COPD	Chronic Obstructive Pulmonary Disease
COVID-19	Coronavirus Disease-2019
CRP	C-Reactive Protein
CT	Computed Tomography
CTPA	Computed Tomography Pulmonary Angiogram
DBP	Diastolic Blood Pressure
D-dimer	Disseminated Intravascular Coagulation

DL	Deep Learning
DM	Diabetes Mellitus
DNAR	Do Not Attempt Resuscitation
eGFR	Estimated Glomerular Filtration Rate
Eus	Euglycaemic Events
FI	First Inpatient
FiO2	Fraction Of Inspired Oxygen
FN	False Negative
FP	False Positive
FTIR	Fourier Transform Infrared
GAN	Generative Adversarial Network
GB	Gradient Boosting
Hb	Haemoglobin
HbA1c	Glycated Haemoglobin
HF	Heart Failure
HCO3	Bicarbonate
HR	Highest Requirement
HV	Highest Value
Hypers	Hyperglycaemic Events
Hypos	Hypoglycaemic Events
IHD	Ischemic Heart Disease
IPLS	Interval Partial Least Squares
K	Potassium
LDA	Linear Discriminant Analysis
LaV	Last Value

LAYBA	Latest Available Within One Year Before Admission
LR	Logistic Regression
LV	Lowest Value
LYM	Lymphocytes
MAE	Mean Absolute Error
MAPD	Mean Absolute Percentage Deviation
MAPE	Mean Absolute Percentage Error
MCC	Matthews Correlation Coefficient
MIR	Mid-Infrared
ML	Machine Learning
MN	Monocytes
MSC	Multivariate Scatter Correction
MSE	Mean Square Error
Na	Sodium
NEUT	Neutrophils
NHS	National Health Service
NIR	Near-Infrared
NLRL	Neutrophils-Lymphocytes Ratio Labelled
O2	Oxygen
OA	On Admission
PBC	Positive Blood Culture
PC	Presenting Complaint
PCA	Principal Component Analysis
PCR	Principal Component Regression
PE	Pulmonary Embolism

PH	Prediction Horizon
PID	Patient Identity
PLSR	Partial Least Squares Regression
PT	Prothrombin Time
PVD	Peripheral Vascular Disease
R2	Coefficient of Determination
RBGLR	Ratio Of Blood Glucose Level Readings
RF	Random Forest
RPLD	Reported Pre-existing Lung Disease
RR	Respiratory Rate
RMSE	Root Mean Square Error
SBP	Systolic Blood Pressure
SD	Standard Deviation
SE	Surveillance Error
SG	Savitzky-Golay
SHAP	SHapley Additive exPlanations
SFFVG	Signal Fragmentation Based Feature Vector Generation
SOB	Shortness Of Breath
SVC	Support Vector Classifier
T1DM	Type 1 Diabetes Mellitus
TIA	Transient Ischemic Attack
TN	True Negative
TP	True Positive
TS	Time Series
TSF	Time Series Forecasting

VRII	Variable Rate Intravenous Insulin Infusion
WCC	White Cell Count

Publication 1.

Blood Glucose Time Series Forecasting: Advanced GAN Driven Interdependent Deep Learning Topologies [1]

Abstract. This article presents new deep learning architectures, leveraging generative adversarial networks, for time series forecasting (TSF) with application to blood glucose prediction. For composing the new structures, three deep learning modules are contrived as precursors; a stem regressor, an auxiliary discriminator, and an auxiliary regressor. The stem regressor, as a compulsory element of all structures, is appointed to execute the primary task of TSF. The module intakes sequences of lag observations and outputs forecasts of future values over a predefined prediction horizon. The auxiliary discriminator is an adversarial compartment that demarcates actual prediction horizon data sequences from those synthesised by the stem regressor. The auxiliary regressor collaborates with the stem regressor by exploiting estimates about the prediction horizon to forecast data excursion over an expanded timespan. The three structures are created depending on whether the stem regressor is retrofitted with the auxiliary discriminator, auxiliary regressor, or both. To instantiate the structures' effectiveness, they are exercised to generate time series blood glucose level forecasting systems. To this end, the well-established publicly available Ohio type 1 diabetes mellitus (T1DM) datasets were scrutinised. The systems developed undergo in-depth mathematical, clinical, and comparative evaluation analysis. The outcomes substantiate the supremacy of the proposed interdependent learning setups over standard benchmarked independent learning procedures.

Keywords. Deep Learning, Time Series, Generative Adversarial Networks, Blood Glucose, Electricity Transformer Temperature

---

[1] At the time of submitting this dissertation, this article was under review at the IEEE Journal of Biomedical Helath Informatics. Authors: **H. Khadem**, H. Nemat, J. Elliott, M. Benaissa.

1. Introduction

Time series (TS) is a collection of time-indexed data points [1]. TS analysis refers to developing methodologies for deriving meaningful properties of TS data [2,3]. Time series forecasting (TSF) is a scope of TS analysis that adopts specialised techniques to project underlying historical patterns of TS data into the future [4]. TSF has found a variety of applications in science and technology [5]. Consequently, extensive research is underway to develop reliable TSF pipelines [6–8]. For instance, studies such as [9–14] have created effective TSF models in a wide range of applications leveraging classical techniques, such as autoregressive, moving average, exponential smoothing, and autoregressive integrated with moving average.

While traditional TSF approaches have been broadly used they have some limitations e.g., tuning difficulty, significant domain knowledge demand, limited capability to deal with complex patterns, and limited ability to handle outliers and missing values [15,16]. Thus, more advanced data analysis algorithms, namely machine learning (ML), have been invoked to further the TSF domain [17–20]. As representatives, in [21–24], ML techniques such as penalised linear methods or nonlinear regression trees have been investigated for generating functional TSF models.

A privileged ML domain to accomplish complicated computing tasks, including TSF, is deep learning (DL) [25–30]. On the one hand, given DL's competence in mapping intricate nonlinear dynamics, it has proven effective for TSF [31,32]. On the other hand, recent technological breakthroughs have facilitated acquiring large TS data typically needed for DL analysis [33,34]. Some of the recent studies that have recruited DL techniques such as recurrent neural networks, convolutional neural networks, and transformers to manufacture dependable TSF models include [35–40]. Despite the efficacy of DL-based TSF analysis, further improvements are still desired in many areas. This has encouraged the generation of compound techniques with more analytical potential.

One way to further enrich the analysis is to incorporate advanced complementary techniques like generative adversarial networks (GANs) [41]. GANs are a type of deep learning model composed of two neural networks, a generator and a discriminator, that work in opposition to improve the model's ability to generate realistic data. The generator's mission is to create lifelike images, and the discriminator's task is to distinguish between actual and artificial images. The crux of GAN is that the two subnetworks are intertwined and trained in parallel through an adversarial loop. In each training iteration, the generator learns to mock more realistic images. At the same time, the discriminator masters differentiating between real and fake images more precisely [42].

After establishing its efficacy in simulating image data, GAN was deployed to synthesise other data types, including TS data [43–46]. Later, customised variations of GAN also emerged for TSF applications. In these modified versions, the generator subnetwork was a regressor predicting a future value based on historical data. The discriminator was a classifier determining whether the predicted value by the regressor resembled the actual reference data point [47–52].

By equipping TSF systems with techniques such as GAN, the analyses have further matured [41]. Notwithstanding all the improvements so far, continued progress in TSF is required to deal with challenging real-world problems [53]. In pursuit of this objective, one approach is to devise advanced TSF infrastructures atop the underlying foundations such as GAN. Such engineering upgrades the repertoire of TSF tools and provides potent alternative solutions to cope with more complex problems.

Backed up by the state-of-the-art paradigm of GAN, the present work offers three interdependent setups for DL sequence-to-sequence univariate TSF. The proposed setups concomitantly train a stem regressor, responsible for performing predictions over a specified prediction horizon (PH), with one or two auxiliary modules. The first platform interlinks the stem regressor with an auxiliary discriminator. The discriminator's task is to detect whether a sequence within a PH is real or synthesised by the stem regressor. During the training, the stem regressor forms an adversarial interaction with the auxiliary discriminator. The stem regressor is educated to produce sequences that, aside from optimising the prediction

performance, curtail the accuracy of the discriminator. In the second platform, the stem regressor is interconnected with an auxiliary regressor. The auxiliary regressor utilises the stem regressor's output sequences to make forecasts of data entities within a specific timeframe beyond the PH. The stem regressor shapes a collaborative relationship with the auxiliary regressor in the training stage. The stem module learns to output sequences that, in addition to minimising the prediction error, lead to increased performance for the auxiliary module. Finally, the third platform interrelates the stem regressor with both the auxiliary regressor and discriminator described above.

To showcase the efficacy of the propositions, the three designed interdependent platforms described above are then tasked with creating blood glucose level prediction (BGLP) systems for individuals with type 1 diabetes mellitus (T1DM). Due to the erratic nature of the phenomenon, automated BGLP for T1DM patients constitutes a TSF research conundrum [54], desiring further enhancements [55]. Accurate BGLP contributes to optimal and sustainable glycaemic control, a crucial goal in managing T1DM [56–58]. In turn, the effective management of the disease reduces the risk of associated acute and chronic complications [59,60]. In recent times, the growing use of wearable sensing devices by people with T1DM, particularly continuous glucose monitoring (CGM), enables the automated collection of vast data in demand for DL analysis [61]. Hence, as in other research areas, DL methods integrated with cutting-edge tools such as GAN have gained utility in generating TSF models to perform BGLP for T1DM patients [62]. Despite many studies dedicated to this topic, constructing decisive BGLP systems remains a TSF challenge.

Overall, the main contributions and objectives of the work include:

- Advancing incorporation of GAN-driven approaches in TFS analysis
- Introducing three novel interdependent platforms for TSF with demonstrated capability
- Evolving GAN-based TSF analysis by devising collaborative interactions between components

The remainder of the paper is organised as follows. Section 2 conducts a concise literature survey on the use of GANS in TS analysis. Section 3 presents three new interdependent structures for univariate time

series forecasting. In section 4, the recommended structures are examined to create BGLP systems. Section 5, experimentally validates the generated systems and discusses the results. Section Finally, the work is summarised and concluded in section 6.

2. Related work

In this section, we overview some of the significant works that have used GAN for TS analysis in various fields. A more comprehensive literature survey of the topic can be found in review articles such as [63–66].

Article [67] presents a promising new approach for modelling financial TS data using GAN. The article, first, argues that traditional models, such as autoregressive or moving average techniques, struggle to capture the complex dynamics of financial data. Then, in response, the work proposes using GAN and shows that this approach outperforms traditional models in terms of accuracy and ability to capture complex patterns in the data. The work also discusses how GAN could be used in various areas of finance, such as fraud detection.

Another article [68] proposes a novel DL methodology that employs GAN for anomaly detection in TS data. The authors assess the efficacy of the method on benchmark datasets and demonstrate its superior performance over several state-of-the-art anomaly detection techniques. This research offers a promising solution to detecting anomalies in time series data that can have far-reaching applications in various domains.

Article [69] proposes a new approach to addressing the problem of missing values in multivariate TS data. The authors introduce a specialized GAN architecture designed for imputing TS data. The evaluation analysis of real-world datasets demonstrates that the model outperforms several existing imputation methods. This research offers a promising solution to the challenge of imputing missing values in multivariate time series data that has practical applications in diverse fields, such as finance, healthcare, and environmental monitoring.

Another research article [70] presents a new method for predicting the hourly photovoltaic power output using conditional GAN. The proposed approach addresses the issue of limited training data by using conditional GAN to generate synthetic data, which helps to increase the size of the training dataset. A DL model then uses the augmented dataset to predict the future output. The study compares the performance of this approach with other popular TSF methods and finds that it outperforms them in terms of accuracy. The new method's potential for improving the accuracy of photovoltaic power forecasting could play a crucial role in integrating renewable energy sources into the power grid.

Article [71] introduces a new approach for TS prediction and classification using a combination of GAN and recurrent neural networks with an attention mechanism. The recurrent neural network unit is used to analyze the temporal patterns in the data, while the GAN unit is employed to create synthetic data that can be used to enhance the training dataset. The study compares the performance of this new approach with other popular methods used for TS prediction and classification, and the results show that it is more accurate and efficient.

Finally, [43] presents a new method for TS resampling that addresses the issue of unevenly spaced data. The proposed method uses GAN to generate synthetic data that can fill in the gaps between the original data points. GAN is trained to learn the statistical patterns of the original data and then used to generate synthetic data points that complete the missing values. The study compares the performance of this method with other popular techniques used for resampling TS data, and the results show that it is more accurate. The authors conclude that this new approach has the potential to improve the accuracy of time series resampling.

Overall, GAN has shown promising results in TS analysis and has outperformed traditional methods in several applications. However, there are still many open research questions and challenges to be addressed in this field, such as designing more effective GAN-driven architectures for TSF.

3. Methods

In this research, three interdependent frameworks are introduced for DL sequence-to-sequence TSF. For building these interrelated learning setups, three components are determined; a stem regressor, an auxiliary discriminator, and an auxiliary regressor. The stem regressor is assigned to accomplish the principal task of TSF. The module receives a determined length of historical data and predicts future data over a certain PH (Figure 1). The auxiliary discriminator is responsible for further assessing the similarity between actual PH sequences and those synthesised by the stem regressor. To do so, the auxiliary component intakes a sequence that contains a certain length of true historical data, true or synthesised PH data, and true post-PH data, and detects whether the sequence contains authentic or synthesised PH data (Figure 2). The auxiliary regressor's mission is to test how PH sequences synthesised by the stem regressor are informative about the future. To this end, the auxiliary component utilises the sequences produced by the stem regressor to make predictions over a period after the PH. In detail, the element intakes a stem module's pair of input and expand the predictions over a predefined post-PH window. For simplicity, the post-PH period is given a length identical to that of the PH (Figure 3).

Since the stem regressor performs the primary TSF predictions, it is a compulsory part of all systems. The two auxiliary components, on the other hand, can subserviently interconnect with the stem unit to undertake supplementary assignments. When interlinked with the auxiliary discriminator, the stem regressor interacts with it adversarially and returns sequences that, in addition to optimising its own predictive performance, demote the accuracy of the discriminator. In contrast, if interrelated with the auxiliary regressor, the stem regressor treats it collaboratively and outputs sequences that, along with minimising prediction error, promote the functionality of the auxiliary compartment. Depending on whether the stem module is interconnected to either or both of the auxiliary compartments, the three interdependent platforms are formed as described below.

3.1. Adversarial Learning

Adversarial learning is a DL technique in which two models, typically a generator and a discriminator, are trained simultaneously. As illustrated in Figure 1, in the adversarial learning framework, the stem regressor is assimilated with an auxiliary discriminator. The two modules are then trained in parallel. In the training phase, the stem regressor forms an adversarial interaction with the auxiliary discriminator. For this purpose, to update the weights of the system the loss functions given in Eqs. 1 and 2 are optimised for the auxiliary discriminator and the stem regressor, respectively. With these loss functions, the auxiliary discriminator learns to label true data as 0 and synthesised data as 1. On the other hand, the stem regressor is educated to produce outcomes that, besides improving the prediction accuracy, undermine the performance of the discriminator by minimising its performance to correctly label synthesised data as 1.

$$L_{AD} = \mathrm{E}(\mathrm{AD}(PH_x), 1) + \mathrm{E}(\mathrm{AD}(PH_{\hat{x}}), 0) \quad (1)$$

$$L_{SR} = \mathrm{E}(SR(H_x), PH_x) + \mathrm{E}(\mathrm{AD}(PH_{\hat{x}}), 1) \quad (2)$$

where, L_a: loss a, AD: auxiliary discriminator, E(a, b): error between a and b, AD(a): auxiliary discriminator's evaluation of a, PH_x: real prediction horizon sequence, $PH_{\hat{x}}$: synthesised prediction horizon sequence, SR: stem regressor, H_x:real history sequence, SR(a): stem regressor's evaluation of a.

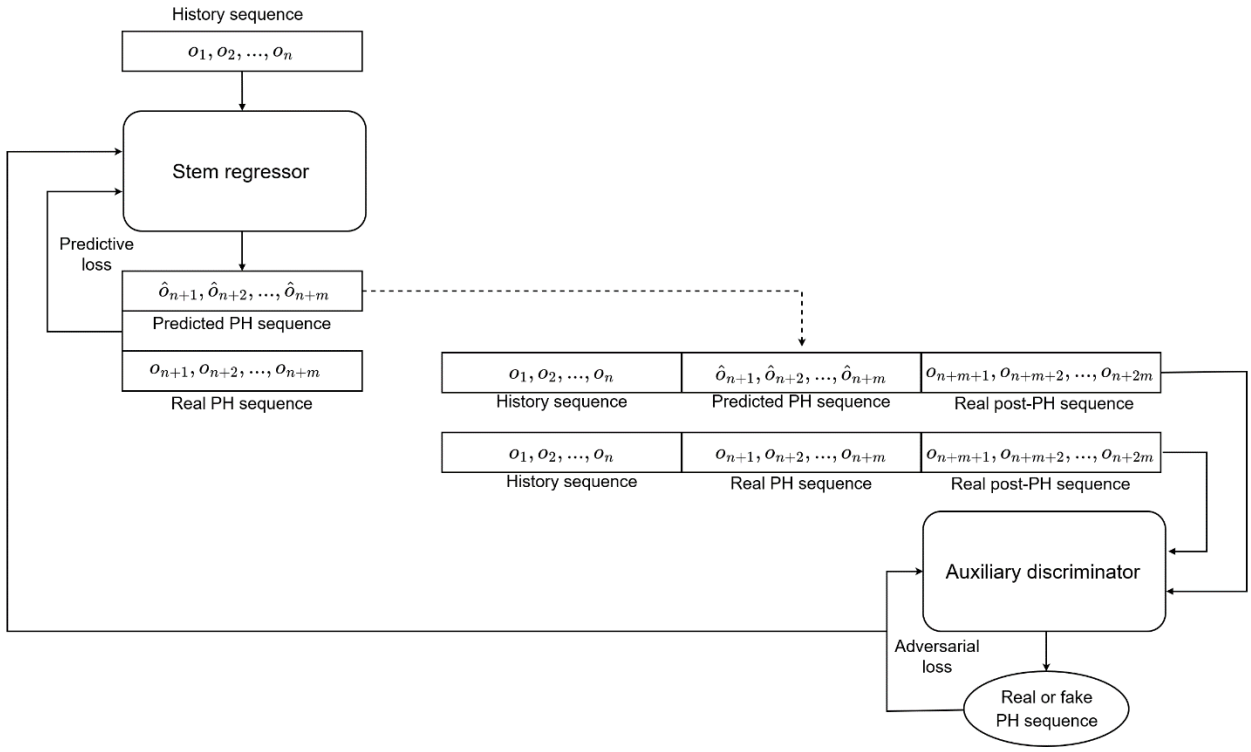

Figure 1. General block diagram of the adversarial learning layout where the stem regressor is internally attached to an auxiliary discriminator. Then, both compartments are trained simultaneously.
Note. N represents the number of values in the history window; M represents the number of data values in the prediction horizon and the post-prediction horizon interval.
Abbreviations. o: observation; PH: prediction horizon.

3.2. Collaborative Learning

The stem regressor is interlinked with the auxiliary regressor in this framework, as shown in Figure 2. The two modules are then trained simultaneously. During the training, the stem regressor works collaboratively with the auxiliary regressor. To do so, the loss functions in Eqs. 3 and 4 are used respectively for the auxiliary regressor and the stem regressor. This way, the stem compartment learns to produce outcomes that, in conjunction with optimising its own performance (minimising PH prediction errors), promote the performance of the auxiliary constituent (minimising post-PH prediction errors) as well.

$$L_{AR} = \mathrm{E}(\mathrm{AR}(H_x + PH_{\hat{x}}), PPH_x) \quad (3)$$

$$L_{SR} = \mathrm{E}(SR(H_x), PH_x) + \mathrm{E}(\mathrm{AR}(H_x + PH_{\hat{x}}), PPH_x) \quad (4)$$

where, L_a: loss a, AR: auxiliary regressor, AR(a): auxiliary regressor's evaluation of a , E(a, b): error between a and b, H_x: real history sequence, $PH_{\hat{x}}$: synthesised prediction horizon sequence, PPH_x: real post prediction horizon sequence, SR: stem regressor, PH_x: real predictin horizon sequence, SR(a): stem regressor's evaluation of a.

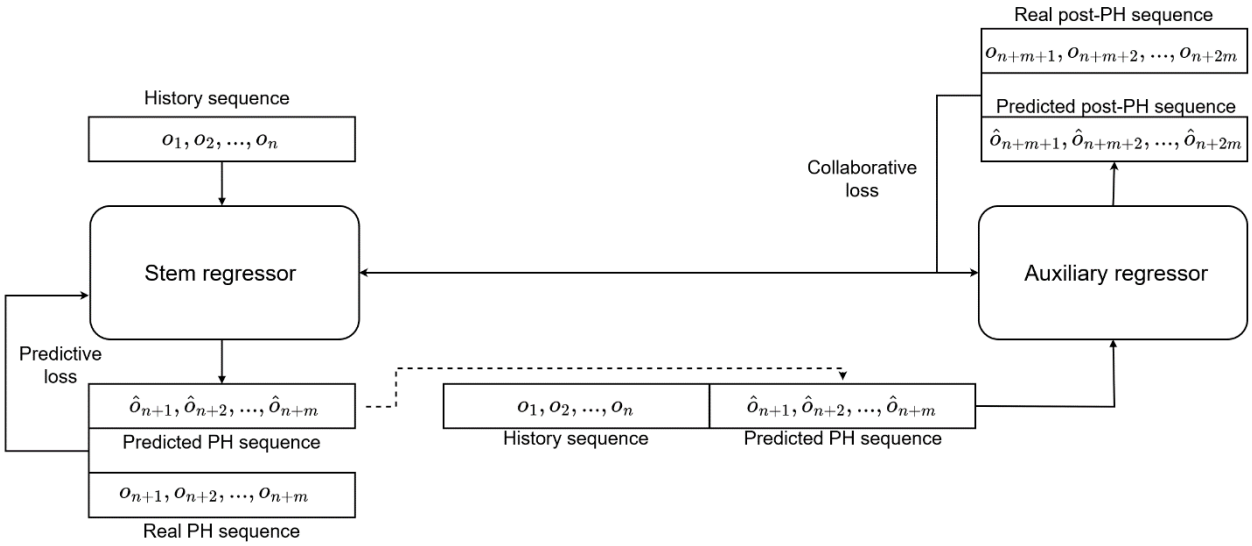

Figure 2. General block diagram of collaborative learning layout where the stem regressor is attached to an auxiliary regressor. Then, both compartments are trained simultaneously.
Note. N represents the number of values in the history window; M represents the number of values in the prediction horizon and the post-prediction horizon interval.
Abbreviations. o: observation; PH: prediction horizon.

3.3. Adversarial Collaborative Learning

As represented in Figure 3, the last learning framework interlinks the stem regressor with both the auxiliary discriminator and regressor. In this scenario, the three modules are trained together, where the stem regressor bounds an adversarial relationship with the auxiliary discriminator and a collaborative relationship with the auxiliary regressor. To this end, the loss functions in Eqs. 5, 6, and 7 are considered for the auxiliary discriminator, auxiliary regressor, and stem regressor. Using these loss functions, with similar explanations given for the two previous learning scenarios, during the training stage of adversarial collaborative learning, the stem regressor learns to generate outcomes that, along with optimising the

prediction error, decrease the discriminator's performance and increase the auxiliary regressor's performance.

$$L_{AD} = \mathrm{E}(\mathrm{AD}(PH_x), 1) + \mathrm{E}(\mathrm{AD}(PH_{\hat{x}}), 0) \quad (5)$$

$$L_{AR} = \mathrm{E}(\mathrm{AR}(H_x + PH_{\hat{x}}), PPH_x) \quad (6)$$

$$L_{SR} = \mathrm{E}(SR(H_x), PH_x) + \mathrm{E}(\mathrm{AD}(PH_{\hat{x}}), 1) + \mathrm{E}(\mathrm{AR}(H_x + PH_{\hat{x}}), PPH_x) \quad (7)$$

where, L_a: loss a, AD: auxiliary discriminator, E(a, b): error between a and b, AD(a): auxiliary discriminator's evaluation of a, PH_x: real prediction horizon sequence, $PH_{\hat{x}}$: synthesised prediction horizon sequence, AD: auxiliary regressor, AR(a): auxiliary regressor's evaluation of a, H_x: real history sequence, PPH_x: real post prediction horizon sequence, SR: stem regressor, SR(a): stem regressor's evaluation of a.

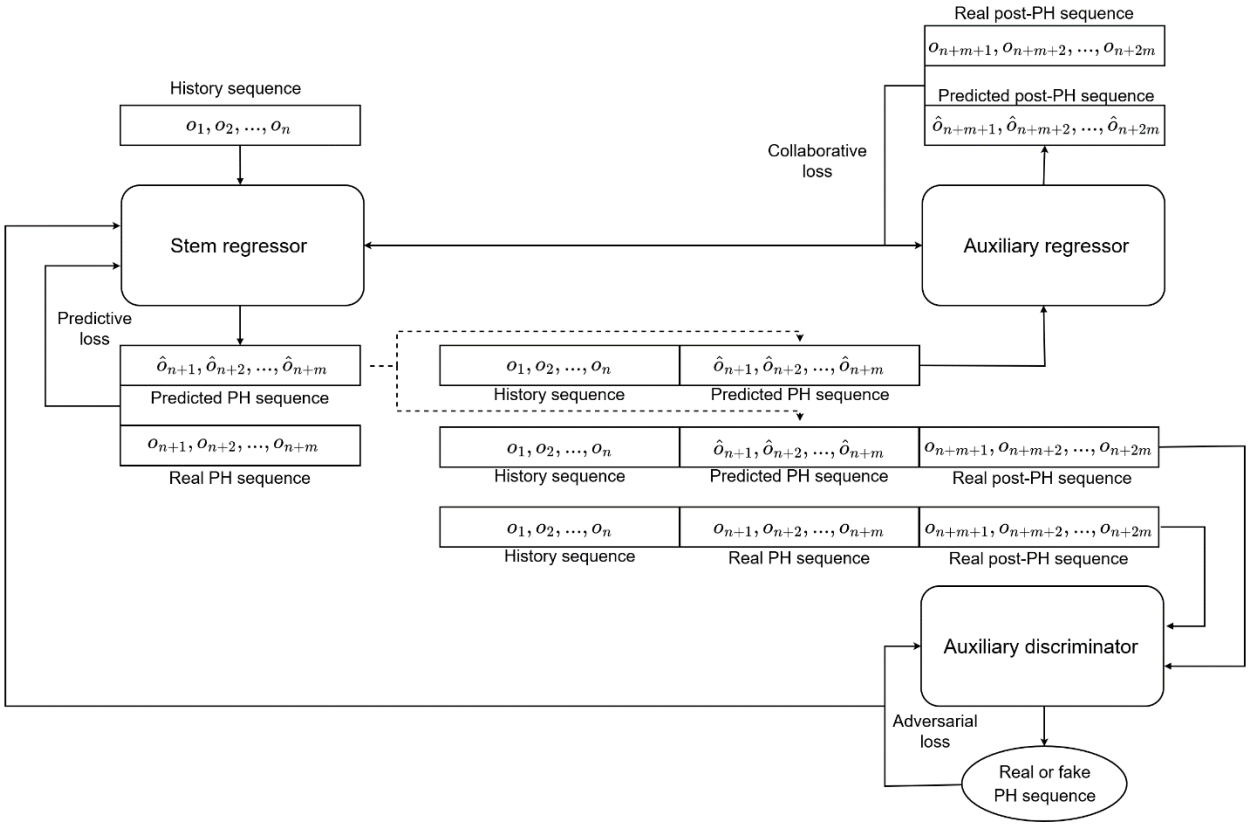

Figure 3. General block diagram of collaborative adversarial learning layout where the stem regressor is attached to an auxiliary discriminator and an auxiliary regressor. Then, all three compartments are trained simultaneously.
Note. N represents the number of values in the history window; M represents the number of values in the prediction horizon and the post-prediction horizon interval.
Abbreviations. o: observation; PH: prediction horizon.

4. Instantiation

The frameworks described in section 3 are instituted to case study BGLP 30 and 60 minutes into the future. Two multilayer perceptron networks with demonstrated usefulness in TSF are allotted as the stem and auxiliary regressors [72]. For simplicity, similar architectures are considered for both networks. They are formed of an input layer, a 50-unit dense layer followed by a 20-unit dense layer and another dense layer as output. The output layer needs six units when PH is 30 minutes and 12 units when PH is 60 minutes. For all layers, ReLU is chosen as the activation function. For the auxiliary discriminator, a convolutional neural network, which generally has excellent classification capacity, was selected [73]. It consists of an input layer, a 20-unit Conv1D layer, a 10-unit Conv1D layer, and a single-unit dense layer as output with a sigmoid activation function and binary cross entropy as the loss function. The number of units in all networks is fined tuned using a grid search approach. The search space for the regressors includes 25, 50 and 100 (5, 10 and 20) nodes for the first (second) layer. The search space for the discriminator is 20, 30 and 40 (5, 10, 15) nodes for the first (second) layer. The values are selected based on the best performance observing training data only. The rest of the hyperparameters are preset and commonly used values.

For developing BGLP systems, two T1DM datasets are studied. Initially, the data are preprocessed according to the requirements of our problem space. Subsequently, to generate the BGLP systems, the stem regressor is trained by exploring the three proposed interdependent setups, along with conventional independent learning. In the independent learning approach, the stem module is trained solo without amalgamating any auxiliary component. In all training scenarios, ADAM is used as the optimiser; epoch size, batch sizes, and learning rate are set at 600, 128, and 0.002, respectively. Rigorous evaluation analyses are then conducted on the generated systems.

4.1. Data

For developing BGLP systems, this work investigates two Ohio T1DM datasets [74] which are well-studied in this field of TSF research [75–80]. Each dataset encompasses eight weeks' worth of diabetes-related attributes for a cohort of six individuals with T1DM [74]. The first dataset compiles data for four females and two males aged between 40 and 60 years [74]. It was released in 2018 for the first BGLP challenge [74]. The second dataset comprises data for one female and five males within the age range of 20–80 years old [74]. This dataset was disseminated for the second BGLP challenge in 2020 [74]. Hereafter, this paper refers to the former dataset as Ohio T1DM 2018 and the latter as Ohio T1DM 2020.

In alignment with previous studies on developing univariate TSF systems for BGLP, the current work studies CGM data included in the datasets [81]. The modality has been collected by Medtronic Enlite sensors in five-minute intervals [74]. For each individual, the collected data points for the last ten days are allocated as the testing set and the preceding data points (46 days) as the training set by the data collection team [74]. All generated BGLP systems are trained only using the training set, and the testing set remains unseen for the evaluation analysis. Table 1 outlines some statistical properties of the CGM data in the training and testing sets of the Ohio T1DM datasets. A comprehensive description of the datasets can be found in the datasets documentation [74].

Table 1. Statistical characteristics of CGM data in Ohio T1DM datasets for each participant's training and testing set.

Dataset	PID	Sex	Age	Set	Statistical property							
					Samples	Range (mg/dl)	Mean (mg/dl)	SD (mg/dl)	Missings (%)	Hypos (%)	Eus (%)	Hypers (%)
Ohio T1DM 2018	559	female	40–60	Training	10655	40–400	167.53	70.44	12.06	3.65	55.98	40.37
				Testing	2444	45–400	168.93	67.78	14.81	3.03	59.86	37.11
	563	male	40–60	Training	11013	40–400	146.94	50.51	8.80	2.82	72.81	24.36
				Testing	2569	62–313	167.38	46.15	4.71	0.70	60.45	38.85
	570	male	40–60	Training	10981	46–377	187.5	62.33	5.73	1.97	42.97	55.07
				Testing	2672	60–388	215.71	66.99	5.05	0.41	29.04	70.55
	575	female	40–60	Training	11865	40–400	141.77	60.27	10.43	8.71	68.62	22.66
				Testing	2589	40–342	150.49	60.53	4.94	5.37	63.50	31.13
	588	female	40–60	Training	12639	40–400	164.99	50.51	3.69	1.04	63.56	35.40
				Testing	2606	66–354	175.98	48.66	3.42	0.15	53.26	46.58
	591	female	40–60	Training	10846	40–397	156.01	58.03	17.59	3.94	63.97	32.09
				Testing	2759	43–291	144.83	51.42	3.15	5.18	67.27	27.55
Ohio T1DM 2020	540	male	20–40	Training	11914	40–369	136.78	54.75	9.76	7.08	72.66	20.25
				Testing	2360	52–400	149.94	66.46	6.74	5.64	68.18	26.19
	544	male	40–60	Training	10533	48–400	165.12	60.08	19.11	1.47	63.78	34.75
				Testing	2715	62–335	156.48	54.14	15.47	1.22	68.29	30.50
	552	male	20–40	Training	8661	45–345	146.88	54.63	22.30	3.89	72.05	24.06
				Testing	1792	47–305	138.11	50.23	85.71	3.57	80.02	16.41
	567	female	20–40	Training	10750	40–400	154.43	60.88	24.91	6.75	63.40	29.84
				Testing	2388	40–351	146.25	55.00	20.18	8.33	67.38	24.29
	584	male	40–60	Training	12027	40–400	192.34	65.29	9.13	0.80	47.69	51.51
				Testing	2661	41–400	170.48	60.76	12.40	1.01	61.86	37.13
	596	male	60–80	Training	10858	40–367	147.17	49.34	25.35	2.08	73.99	23.93
				Testing	2663	49–305	146.98	50.79	9.76	2.78	75.07	22.16

Note. CGM: continuous glucose monitoring; T1DM: type 1 diabetes mellitus; PID: patient identification; SD: standard deviation; Hypos: hypoglycaemic events; Eus: euglycaemic events; Hypers: hyperglycaemic events.

4.2. Preprocessing

This subsection reports the preprocessing analysis operated on the Ohio T1DM datasets before proceeding with the BGLP modelling phase.

4.2.1. Imputation

In the first stage of the preprocessing, missing CGM values are handled. Linear interpolation is implemented to fill in missing values in the training set. However, missing values in the testing set are imputed utilising linear extrapolation. This technique avoids information leakage by ascertaining that systems do not manipulate future information in the evaluation stage. Ergo, the resultant systems would be functional for real-time predictions.

4.2.2. Reframing

The next preprocessing stage is translating the sequence-to-sequence BGLP task to a supervised ML problem. For this purpose, a window with the length of history plus PH is rolled over the CGM series, creating a set of associated vectors. Each vector is then split into pairs of input and output sequences according to the length of history and PH. This operation renders a subset of associated input and output sequences necessary for supervised ML [82–84]. To exemplify, a rolling window with a length of 90 minutes forms a set of vectors for BGLP 30 minutes in advance from 60 minutes of lag observations. Then, each vector is subdivided; the first 60 minutes are set as an input sequence and the last 30 minutes as the associated output sequence. It merits mentioning that considering the sampling frequency of CGM values in the Ohio T1DM datasets, every five-minute interval corresponds to one timestep.

4.3. Evaluation

The performance of the developed systems in making predictions across the entire testing set is rigorously reviewed from mathematical and clinical perspectives. For a thorough assessment, several evaluation metrics are considered for each viewpoint as follows.

4.3.1. Mathematical Evaluation

BGLP errors are measured via three expansively used regression metrics; root mean square error (RMSE), mean absolute error (MAE), and mean absolute percentage error (MAPE), as in Eqs. 8, 9, and 10, respectively. Moreover, the coefficient of determination (r^2), calculated as Eq. 11, is leveraged to rate the correlation between the reference and predicted BGLs.

$$RMSE = \sqrt{\left(\sum_{i=1}^{N}(BGL_i - \hat{BGL}_i)^2\right)/n} \quad (8)$$

$$MAE = \left(\left|BGL_i - \hat{BGL}_i\right|\right)/N \quad (9)$$

$$MAPE = \left(\left(\sum_{i=1}^{N}\left|\left(BGL_i - \hat{BGL}_i\right)/BGL_i\right|\right)/\mathrm{N}\right) \times 100 \quad (10)$$

$$r^2 = 1 - ((\sum_{i=1}^{N} (BGL_i - \hat{BGL}_i)^2) \Big/ (\sum_{i=1}^{N} (BGL_i - \overline{BGL})^2)) \quad (11)$$

where, N represents the size of the testing set, and the hat symbol denotes predicted BGLs.

4.3.2. Clinical Evaluation

The following evaluation criteria were employed to assess the performance of the generated systems from the clinical point of view.

Surveillance error (SE) quantifies the clinical risk of BGLP errors by assigning a value to each prediction [85]. A detailed explanation of the calculation for SE is included in the original article [85]. Succinctly, 0 < SE < 0.5 reflects no clinical risk, 0.5 < SE < 1.5 slight clinical risk, 1.5 < SE < 2.5 moderate clinical risk, 2.5 < SE < 3.5 high clinical risk, and 3.5 < SE extreme clinical risk. This work utilises the percentage of predictions with no clinical risk (SE < 0.5) and the average surveillance error (ASE) for predictions across the entire testing set as evaluation metrics.

Matthews correlation coefficient (MCC) measures the quality of binary classification [86]. The metric is specifically beneficial when the distribution of data in the two classes is unbalanced [86]. In this work, MCC is calculated as Eq. 12 to score the fulfilment of the systems' predictions in correctly prognosticating the occurrence of adverse glycaemic events (BGL < 70 mg/dl or >180 mg/dl) as opposed to euglycaemic events (70 mg/dl < BGL < 180 mg/dl).

$$MCC = (TP \times TN - FP \times FN) / \sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)} \quad (12)$$

where TP (true positive) represents the number of adverse glycaemic events foresaw truly by the BGLP system, TN (true negative) represents the number of euglycaemic events foresaw truly, FP (false positive) represents the number of adverse glycaemic events foresaw falsely, and FN (false negative) represents the number of euglycaemic events foresaw falsely.

5. Results and Discussion

This section presents the outcomes of the evaluation analyses for generated BGLP systems and the associated discussions. In response to the stochastic essence of DL algorithms, which causes randomness in the outcomes, each system runs five times, and the results achieved are presented in the form of mean and standard deviation (SD).

Tables Table and Table display the evaluation results for systems created respectively using the Ohio T1DM 2018 dataset and Ohio T1DM 2020 dataset. Each table is compartmentalised with the results of 12 scenarios, i.e., BGLP modelling for six data contributors in two PHs. For each scenario, four systems are created using the three interdependent learning platforms and independent learning approach, one a piece.

Monochromatic colour coding is applied in the tables to visualise intra-scenario comparisons between learning platforms. To this end, the cells are shaded using four grey colours, where the darkest to lightest colour codes the outcomes for each metric from the first to the fourth rank. According to the tables, no learning structure is universally superior, and there are circumstances where each interdependent layout dominated the others. For instance, for BGLP 60 minutes into the future, the best learning setup was adversarial learning for subject 559, collaborative learning for subject 570, and adversarial collaborative learning for subject 540. These outcomes imply that for a new given BGLP problem, there is a prospect of superiority for each interdependent learning structure. This finding promises the potential utility of the proposed interdependent learning forums to address challenging TSF tasks.

An inter-framework comparative analysis is also performed, considering all evaluation metrics in all scenarios. For this purpose, the results of colour-coded rankings in Tables 2 and 3 are combined and encapsulated in the pie charts shown in Figure 4. Each chart illustrates a summary of the rankings achieved using each learning structure for all evaluation metrics across the aggregate pool of scenarios considered for both datasets. For cohesion, the charts are colour-coded using the same pallet and approach as applied to the tables. As the charts highlight, for the sample TSF problem investigated in this research, the consensus

was that adversarial collaborative learning delivered the highest performance overall. This learning platform produced outcomes for the evaluation metrics placed first, second, third, and fourth rank in, respectively, 39%, 20%, 19%, and 17% of the cases. Therefore the average ranking of the evaluation metrics produced by this platform was 2.14. Collaborative learning was the second-best platform overall, with an average ranking of 2.22. Adversarial learning with an average ranking of 2.49 and independent learning with an average ranking of 3.04 was third and fourth, respectively. Statistical significance between the models' performance was then verified by conducting the Friedman test [87]. Friedman [88] is a non-parametric statistical test as an alternative to the ANOVA [89] test with no normality assumption. The null hypothesis in this analysis was that the models have identical ranking distribution and the significance level was 0.05. The null hypothesis was rejected by a resultant p-value of 0.0265.

Overall, the results obtained underscore the potential efficacy of the proposed learning frameworks in advancing BGLP when compared with the traditional independent learning mechanism. Apropos the monitored capability of the proposed interdependent learning structures, the following explanations are provided. Adversarial interaction of the stem regressor with the discriminator enforces the stem regressor to produce sequences that, along with optimising prediction errors, adhere to the data distribution within the PH. This way, further assessments are dictated on the sequences generated by the stem regressor. In detail, the output sequences are compelled to resemble real CGM sequences and fit the PH's data pattern. Likewise, a collaborative relationship of the stem regressor with the auxiliary regressor also imposes exterior assessments on the outputs of the stem regressor. In this manner, the sequences produced by the stem regressor are directed to have increased knowledge of the future. As a result, the trained stem regressor yields sequences that possess reinforced future predictivity. As a result of dual interactions with both the auxiliary modules, the stem regressor learns to return sequences with optimum prediction errors that are also more consistent with the data dynamics in the PH and hold improved future predictivity.

In addition to the evaluation results above, here, a complementary benchmark analysis is performed to further demonstrate the potential of the proposed interdependent topologies. For this purpose, the outcomes

ACL in creating BGLP models are studied side-by-side with existing approaches. For comparative analysis, the benchmark experimental design in the literature is used [90].

Table 4 displays the results of BGLP over the Ohio T1DM datasets for 9 existing approaches alongside ACL. Detailed descriptions of the benchmarked models are given in [90], here a brief explanation of them is given in the table footnote. According to the results in the table, ACL is amongst the top-performance models in all scenarios, further verifying the capability of the propositions.

Accurate blood BGLP represents a significant advancement in the management of T1D, enabling more informed decisions regarding insulin dosing and other therapeutic interventions. The superior performance of the proposed interdependent learning topologies in minimising prediction errors directly enhances the reliability of BGLP, which is crucial for preventing dangerous hypoglycaemic and hyperglycaemic events. The ability of these models to maintain high accuracy across diverse patient datasets underscores their potential for widespread clinical application, particularly in optimising CGM systems.

When integrated with CGM systems, these predictive models can provide real-time, personalised insights into glucose trends, allowing for more precise and proactive diabetes management. This could lead to improved glycaemic control, reduced frequency of complications, and ultimately a better quality of life for patients with T1D. Among the proposed frameworks, ACL is particularly noteworthy for its adaptability to individual patient data, which supports its role in personalised medicine. By incorporating these advanced models into clinical practice, healthcare providers can enhance the safety and efficacy of diabetes management, leading to better patient outcomes and potentially reducing the overall burden of T1D on healthcare systems.

6. Denouement

This work designed three novel interdependent learning environments upon principles of GAN for sequence-to-sequence univariate TSF. A stem regressor assigned to perform the primary task of TSF is trained within these learning environments to generate TSF systems. Therein, the stem regressor undergoes beneficiary adversarial, collaborative, or adversarial collaborative interactions with auxiliary components. To instantiate the efficacy of the propositions, the new setups were exploited to create BGLP systems, a challenging real-life TSF problem desiring improved analysis. In this sense, experiments were conducted on two well-known and publicly available Ohio T1DM datasets. The generated systems were evaluated vigorously regression-wise and clinical-wise. The evaluation results promised the effectiveness and compatibility of the propositions. All in all, the interpretation of the results supported the capability of the proposed interdependent platforms to supply alternative learning mechanisms that could transcend the conventional independent learning approaches. The proposed frameworks may also pave the way for the construction of more revolutionised TSF forums. In this sense, further work can involve customising the frameworks for application in other fields such as finance, energy, or climate forecasting. This can potentially extend the applicability and usefulness of the proposed methods and identify new insights in various domains.

Table 2. Evaluation results for BGLP systems created using Ohio T1DM 2018 dataset.

Scenario	Learning	Evaluation metric						
		RMSE±SD (mg/dl)	MAE±SD (mg/dl)	MAPE±SD (%)	r^2±SD (%)	MCC±SD (%)	SE<0.5±SD (%)	ASE±SD
PID 559 PH 30	IL	19.36±0.43	13.45±0.22	8.79±0.19	91.71±1.20	81.08±0.86	89.96±0.56	0.189±0.002
	AL	18.87±0.32	13.29±0.14	8.65±0.12	92.12±1.12	80.01±0.98	90.29±0.95	0.185±0.003
	CL	19.05±0.34	13.17±0.18	8.78±0.17	91.97±1.01	80.68±0.89	89.52±0.51	0.184±0.001
	ACL	18.77±0.18	13.09±0.12	8.55±0.10	92.20±0.92	81.37±0.87	89.60±0.62	0.184±0.002
PID 559 PH 60	IL	33.22±0.65	25.08±0.40	17.97±0.38	75.58±0.96	64.43±0.88	75.78±0.92	0.350±0.003
	AL	32.54±0.94	23.88±0.47	16.19±0.41	76.57±0.89	65.57±0.74	78.67±0.91	0.329±0.005
	CL	32.07±0.56	24.26±0.32	16.00±0.27	75.80±0.75	65.27±0.62	78.11±0.62	0.327±0.002
	ACL	32.54±0.49	24.12±0.24	16.55±0.20	76.57±0.67	64.71±0.54	77.43±0.42	0.336±0.007
PID 563 PH 30	IL	19.68±0.21	13.37±0.19	8.23±0.17	81.81±0.67	74.17±0.86	91.68±1.01	0.183±0.004
	AL	19.05±0.18	14.02±0.16	8.19±0.15	80.42±0.71	73.46±0.51	91.64±0.95	0.188±0.002
	CL	19.17±0.22	13.61±0.16	8.13±0.13	81.75±0.72	73.60±0.42	91.56±0.62	0.187±0.003
	ACL	18.75±0.12	13.05±0.10	8.15±0.09	82.37±0.51	77.10±0.49	91.95±0.79	0.182±0.001
PID563 PH 60	IL	29.97±0.84	21.87±0.53	13.81±0.42	57.81±0.74	57.97±0.28	81.23±0.34	0.305±0.006
	AL	30.57±0.54	22.06±0.32	13.65±0.24	56.11±0.62	56.61±0.59	80.71±0.47	0.301±0.004
	CL	30.65±0.62	21.99±0.36	13.63±0.21	55.88±0.28	56.99±0.39	80.52±0.51	0.301±0.001
	ACL	31.04±0.41	22.64±0.26	13.65±0.14	54.73±0.39	53.13±0.62	79.50±0.61	0.306±0.006
PID 570 PH 30	IL	16.25±0.0.35	11.41±0.28	5.68±0.14	94.05±0.81	86.49±0.56	96.44±0.79	0.107±0.002
	AL	16.08±0.0.30	11.25±0.12	5.73±0.12	94.17±0.45	86.80±0.68	96.40±1.25	0.109±0.004
	CL	16.47±0.28	11.21±0.14	5.60±0.11	93.88±0.31	86.79±0.98	96.62±0.68	0.105±0.003
	ACL	15.86±0.38	11.07±0.11	5.58±0.09	94.18±0.64	86.73±0.57	96.58±0.72	0.104±0.002
PID 570 PH 60	IL	29.89±0.61	22.26±0.42	10.91±0.20	79.86±0.96	75.79±0.95	89.89±0.65	0.203±0.003
	AL	28.38±0.84	20.65±0.24	10.31±0.18	81.85±0.92	76.76±0.78	90.11±0.91	0.193±0.004
	CL	27.37±0.65	19.94±0.18	10.39±0.16	83.10±0.54	77.55±0.69	90.26±0.57	0.193±0.003
	ACL	27.5±0.23	20.05±0.12	10.48±0.08	82.95±0.53	77.88±0.62	90.41±0.64	0.195±0.002
PID 575 PH 30	IL	23.01±0.44	14.34±0.32	10.13±0.12	85.59±0.50	76.08±0.87	86.37±0.84	0.222±0.004
	AL	23.00±0.0.31	14.85±0.21	10.54±0.29	83.91±0.74	77.23±0.32	88.00±0.69	0.224±0.004
	CL	22.45±0.61	15.11±0.53	10.16±0.16	86.28±0.64	77.31±0.57	86.75±0.72	0.219±0.0025
	ACL	22.52±0.25	15.16±0.21	10.03±0.17	86.20±0.81	77.81±0.61	86.99±0.56	0.217±0.003
PID 575 PH 60	IL	36.82±0.53	26.30±0.32	18.90±0.20	63.12±0.69	52.24±0.62	70.89±0.85	0.396±0.006
	AL	36.25±0.62	25.97±0.28	18.80±0.24	66.26±0.82	52.37±0.54	71.80±0.92	0.395±0.007
	CL	35.15±0.87	25.57±0.42	19.38±0.14	66.09±0.86	54.11±0.89	70.85±0.59	0.398±0.005
	ACL	36.55±0.82	26.07±0.29	19.07±0.19	64.65±0.58	52.02±0.38	70.03±0.67	0.397±0.004
PID 588 PH 30	IL	19.25±0.51	13.78±0.32	8.29±0.21	83.95±0.87	71.09±0.84	92.02±0.69	0.183±0.001
	AL	18.81±0.39	13.67±0.24	8.14±0.22	84.67±0.59	73.05±0.89	91.91±0.92	0.198±0.002
	CL	18.41±0.42	13.58±0.26	8.16±0.13	85.32±0.74	75.24±0.52	92.02±0.51	0.184±0.001
	ACL	18.72±0.35	13.64±0.14	8.25±0.08	84.82±0.92	76.24±0.43	91.73±0.68	0.182±0.003
PID 588 PH 60	IL	31.04±0.42	23.06±0.26	14.18±0.19	58.25±0.62	59.40±0.63	78.17±0.54	0.309±0.003
	AL	30.92±0.18	22.63±0.12	13.90±0.16	56.56±0.42	57.36±0.51	80.05±0.58	0.277±0.003
	CL	31.19±0.59	22.97±0.23	13.62±0.18	58.56±0.20	58.36±0.24	79.07±0.92	0.303±0.004
	ACL	31.75±0.51	23.14±0.27	13.56±0.23	56.32±0.54	56.01±0.26	80.30±67	0.289±0.004
PID 591 PH 30	IL	21.13±0.17	15.61±0.14	11.95±0.13	81.92±0.54	62.66±0.52	82.79±0.74	0.264±0.003
	AL	21.41±0.25	15.24±0.21	11.83±0.15	82.46±0.68	65.11±0.46	82.61±0.72	0.262±0.002
	CL	21.22±0.20	15.68±0.15	11.63±0.13	82.77±0.41	63.59±0.49	82.90±0.46	0.259±0.001
	ACL	21.25±0.24	15.49±0.16	12.13±0.14	82.73±0.86	62.97±0.52	82.10±0.61	0.268±0.003
PID 591 PH 60	IL	33.48±0.56	25.88±0.28	21.08±0.25	57.11±0.42	44.65±0.51	68.38±0.88	0.418±0.008
	AL	32.75±0.84	25.15±0.40	20.75±0.24	58.19±0.68	44.61±0.62	69.12±0.71	0.407±0.006
	CL	33.20±0.54	25.53±0.27	19.73±0.19	58.96±0.77	45.69±0.66	69.23±0.75	0.418±0.002
	ACL	33.00±0.20	25.45±0.16	20.08±0.12	58.98±0.20	46.52±0.51	69.44±0.35	0.413±0.001

Note. BGLP: blood glucose level prediction; T1DM: type 1 diabetes mellitus; RMSE: root mean square error; MAE: mean absolute error; MAPE: mean absolute percentage error; r^2: coefficient of determination; SE: surveillance error; ASE: average surveillance error; MCC: Matthews correlation coefficient; PID: patient identity; PH: prediction horizon; IL: independent learning; AL: adversarial learning; CL: collaborative learning; ACL: adversarial collaborative learning.

Note. Outcomes of each evaluation metric for a given modelling scenario are colour-coded from dark grey for best to light grey for worst outcomes.

Table 3. Evaluation results for BGLP systems created using Ohio T1DM 2020 dataset.

Scenario	Learning	Evaluation metric						
		RMSE±SD (mg/dl)	MAE±SD (mg/dl)	MAPE±SD (%)	r^2±SD (%)	MCC±SD (%)	SE<0.5±SD (%)	ASE±SD
PID 540 PH 30	IL	21.49±0.21	15.93±0.16	11.81±0.14	89.72±0.68	73.84±0.74	81.05±0.84	0.228±0.003
	AL	20.90±0.12	15.63±0.10	11.17±0.08	90.01±0.86	74.46±0.76	82.54±0.67	0.238±0.004
	CL	21.68±0.32	16.08±0.15	11.02±0.12	89.58±0.95	74.70±0.84	81.98±0.93	0.231±0.002
	ACL	21.18±0.44	15.75±0.24	11.07±0.18	90.26±0.62	73.98±0.62	84.69±0.74	0.234±0.001
PID 540 PH 60	IL	40.10±0.62	30.59±0.81	21.71±0.45	64.88±0.54	53.71±0.47	61.20±0.56	0.424±0.005
	AL	38.76±0.54	29.77±0.65	21.96±0.42	65.54±0.62	56.92±0.58	62.85±0.48	0.419±0.006
	CL	39.01±0.53	30.06±0.62	21.20±0.35	65.50±0.42	54.88±0.52	65.12±0.69	0.416±0.007
	ACL	38.84±0.84	29.48±0.68	20.67±0.32	66.43±0.61	57.62±0.48	64.51±0.38	0.404±0.001
PID 544 PH 30	IL	18.02±0.61	13.06±0.36	8.94±0.19	88.61±0.69	79.21±0.69	91.15±1.08	0.192±0.005
	AL	17.99±0.28	12.70±0.18	8.35±0.12	88.74±0.79	79.45±0.62	91.60±1.12	0.180±0.001
	CL	17.98±0.29	12.61±0.17	8.23±0.11	88.74±0.86	79.67±0.57	92.08±0.85	0.175±0.002
	ACL	18.01±0.32	12.72±0.16	8.21±0.15	88.72±0.92	79.69±0.54	91.89±0.60	0.174±0.002
PID544 PH 60	IL	31.39±0.59	23.65±0.56	16.05±0.46	65.71±0.60	58.59±0.74	74.03±0.64	0.336±0.005
	AL	30.93±0.42	23.21±0.35	15.83±0.23	66.72±0.55	60.77±0.42	75.14±0.56	0.328±0.001
	CL	30.85±0.62	23.25±0.31	16.01±0.22	66.88±0.43	61.05±0.65	75.07±0.48	0.333±0.001
	ACL	31.47±0.24	23.92±0.21	16.65±0.15	65.54±0.39	61.25±0.32	74.70±0.72	0.341±0.002
PID 552 PH 30	IL	16.73±0.34	12.50±0.14	9.70±0.10	89.71±0.98	74.76±0.67	89.40±0.86	0.205±0.001
	AL	16.78±0.13	12.50±0.08	9.51±0.08	89.65±0.86	74.98±0.54	89.87±0.81	0.202±0.003
	CL	16.89±0.21	12.66±0.11	9.70±0.09	89.52±0.84	75.14±0.84	89.40±0.82	0.205±0.002
	ACL	16.69±0.18	12.25±0.11	9.35±0.10	89.76±0.69	74.97±0.46	90.25±0.39	0.197±0.003
PID 552 PH 60	IL	29.42±0.84	21.82±0.54	16.53±0.45	68.22±0.38	59.36±0.42	75.85±0.84	0.331±0.006
	AL	30.45±0.54	23.76±0.42	19.72±0.36	65.93±0.46	59.38±0.65	73.50±0.85	0.369±0.003
	CL	29.40±0.62	22.24±0.42	17.52±0.32	68.18±0.86	60.13±0.42	75.64±0.86	0.342±0.001
	ACL	28.96±0.49	22.18±0.36	17.71±0.22	69.02±0.84	61.41±0.38	75.55±0.94	0.341±0.002
PID 567 PH 30	IL	20.68±0.42	14.74±0.21	11.01±0.15	85.76±0.58	64.45±0.48	82.42±0.59	0.258±0.003
	AL	21.11±0.65	15.25±0.20	11.45±0.14	85.02±0.75	62.95±0.52	81.81±0.68	0.268±0.005
	CL	20.88±0.35	14.98±0.15	11.18±0.11	85.34±0.69	64.03±0.68	82.24±0.84	0.260±0.002
	ACL	20.57±0.21	14.38±0.12	10.28±0.10	85.78±0.76	68.08±0.63	85.15±0.20	0.232±0.002
PID 567 PH 60	IL	38.41±0.48	29.59±0.35	23.81±0.16	50.62±0.68	58.04±0.54	61.18±0.18	0.503±0.004
	AL	35.99±0.81	26.68±0.54	19.89±0.32	56.46±0.41	57.99±0.16	64.43±0.58	0.462±0.003
	CL	37.42±0.74	28.78±0.42	22.71±0.33	52.77±0.68	59.13±0.28	62.91±0.92	0.486±0.002
	ACL	36.93±0.59	27.72±0.37	21.26±0.25	54.09±0.82	58.03±0.38	66.04±0.54	0.440±0.002
PID 584 PH 30	IL	21.78±0.19	15.83±0.13	10.56±0.10	87.18±0.91	77.08±0.68	88.02±0.91	0.211±0.001
	AL	22.35±0.22	16.12±0.11	11.20±0.08	86.40±0.57	77.28±0.57	87.54±0.21	0.229±0.002
	CL	21.32±0.32	15.24±0.13	9.89±0.09	87.52±0.65	78.05±0.86	89.24±0.52	0.203±0.003
	ACL	22.41±0.26	16.44±0.14	11.25±0.08	86.26±0.42	77.24±0.91	87.54±0.84	0.229±0.000
PID 584 PH 60	IL	36.14±0.96	26.53±0.74	17.06±0.65	64.14±0.87	64.12±0.54	74.74±0.62	0.343±0.008
	AL	36.37±0.71	26.79±0.62	17.41±0.42	63.68±0.54	63.21±0.58	74.55±0.61	0.340±0.006
	CL	36.82±0.65	27.65±0.42	18.74±0.35	62.78±0.25	61.75±0.68	72.39±0.58	0.366±0.005
	ACL	36.04±0.58	26.55±0.38	17.34±0.32	64.34±0.36	63.27±0.35	74.85±0.46	0.343±0.004
PID 596 PH 30	IL	17.92±0.29	12.75±0.21	9.54±0.16	87.83±0.88	75.67±1.06	88.81±0.89	0.203±0.002
	AL	17.68±0.41	12.44±0.18	9.28±0.13	87.70±0.75	74.33±0.96	89.07±0.92	0.209±0.002
	CL	17.42±0.18	12.11±0.11	8.84±0.10	88.04±0.71	74.58±0.52	89.40±0.81	0.193±0.001
	ACL	17.83±0.13	12.35±0.06	8.89±0.05	87.48±0.68	73.53±0.38	89.26±0.50	0.196±0.002
PID 596 PH 60	IL	29.41±0.59	21.33±0.42	15.89±0.31	68.77±0.84	54.61±0.68	78.63±0.88	0.321±0.003
	AL	29.04±0.65	20.91±0.43	15.48±0.32	66.80±0.51	58.69±0.54	79.47±0.96	0.313±0.000
	CL	28.93±0.71	20.87±0.49	15.31±0.36	67.05±0.68	57.50±0.49	79.14±0.72	0.312±0.001
	ACL	29.21±0.48	20.99±0.36	15.10±0.22	66.41±0.40	56.29±0.68	79.44±0.13	0.311±0.004

Note. BGLP: blood glucose level prediction; T1DM: type 1 diabetes mellitus; RMSE: root mean square error; MAE: mean absolute error; MAPE: mean absolute percentage error; r^2: coefficient of determination; SE: surveillance error; ASE: average surveillance error; MCC: Matthews correlation coefficient; PID: patient identity; PH: prediction horizon; IL: independent learning; AL: adversarial learning; CL: collaborative learning; ACL: adversarial collaborative learning.

Note. Outcomes of each evaluation metric for a given modelling scenario are colour-coded from dark grey for best to light grey for worst outcomes.

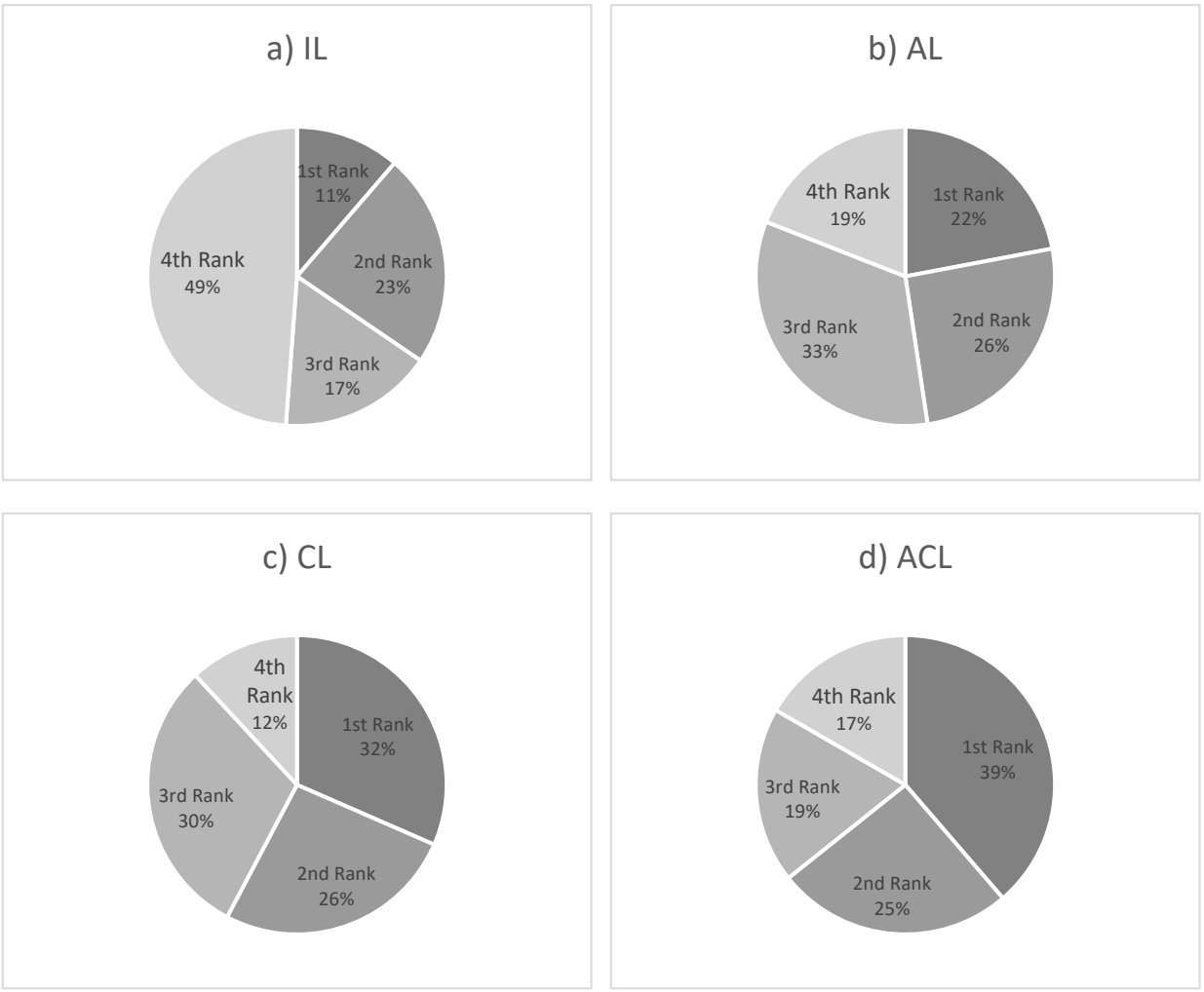

Figure 4. Summary of the evaluation metrics rankings obtained using each learning framework across all the scenarios considered for both studied datasets.
Abbreviations. IL: independent learning; AL: adversarial learning; CL: collaborative learning; ACL: adversarial collaborative learning.

Table 4. Results of comparative analysis of mean accuracy performance for blood glucose level prediction over Ohio T1DM dataset.

Model	PH 30 min		PH 60 min	
	RMSE (mg/dl)	MAPE (%)	RMSE (mg/dl)	MAPE±SD (%)
Baseline	28.32	13.51	41.02	20.37
Poly	57.26	31.09	57.27	31.04
AR	20.70	9.62	33.20	16.73
ARX	20.61	9.59	33.43	16.73
SVR	20.10	9.08	32.27	15.38
GP	20.01	9.16	31.97	15.92
ELM	25.38	11.56	35.14	16.91
FFNN	21.00	9.33	32.93	15.83
LSTM	20.46	9.24	32.88	16.00
ACL	19.38	9.31	32.81	16.84

Note. Baseline: a naive model that copies the last observation in the history window as future predictions; Poly: a polynomial regression model; AR and ARX: models from the ARIMAX family; SVR: a support vector regression model, GP: a Gaussian Process model; ELM: an Extreme Learning Machines model; FFNN: a Feed-forward Neural Network model; LSTM: a Long Short-Term Memory Recurrent Neural Network model; ACL: adversarial collaborative learning; RMSE: root mean square error; MAPE: mean absolute prediction error.

Data Availability

Instructions on attaining the Ohio T1DM datasets can be found at this address.

Code Availability

We have made our source codes accessible on this Gitlab repository. For these implementations, we scripted in Python (3.6.7) [91]. The third-party libraries used include TensorFlow [92], Keras [93], Pandas [94], NumPy [95], Sklearn [96], and statsmodels [97].

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Maçaira PM, Tavares Thomé AM, Cyrino Oliveira FL, Carvalho Ferrer AL. Time series analysis with explanatory variables: A systematic literature review. Environ Model Softw 2018;107:199–209. https://doi.org/10.1016/J.ENVSOFT.2018.06.004.

[2] Zou Y, Donner R V., Marwan N, Donges JF, Kurths J. Complex network approaches to nonlinear time series analysis. Phys Rep 2019;787:1–97. https://doi.org/10.1016/J.PHYSREP.2018.10.005.

[3] Broomhead D. S. and Jones Roger. Time-series analysis. Proc. R. Soc. London. A. Math. Phys. Sci., vol. 423, The Royal Society London; 1989, p. 103–21. https://doi.org/10.1098/RSPA.1989.0044.

[4] Sibeijn M, Pequito S. A time-reversed model selection approach to time series forecasting. Sci Rep 2022;12:1–14. https://doi.org/10.1038/s41598-022-15120-x.

[5] Meisenbacher S, Turowski M, Phipps K, Rätz M, Müller D, Hagenmeyer V, et al. Review of automated time series forecasting pipelines. Wiley Interdiscip Rev Data Min Knowl Discov 2022:e1475.

[6] Athiyarath S, Paul M, Krishnaswamy S. A Comparative Study and Analysis of Time Series Forecasting Techniques. SN Comput Sci 2020 13 2020;1:1–7. https://doi.org/10.1007/S42979-020-00180-5.

[7] Aijaz I, Agarwal P. A Study on Time Series Forecasting using Hybridization of Time Series Models

and Neural Networks. Recent Adv Comput Sci Commun 2019;13:827–32. https://doi.org/10.2174/1573401315666190619112842.

[8] Semenoglou AA, Spiliotis E, Makridakis S, Assimakopoulos V. Investigating the accuracy of cross-learning time series forecasting methods. Int J Forecast 2021;37:1072–84. https://doi.org/10.1016/J.IJFORECAST.2020.11.009.

[9] Aamir M, Shabri A. Modelling and forecasting monthly crude oil price of Pakistan: A comparative study of ARIMA, GARCH and ARIMA Kalman model. AIP Conf Proc 2016;1750:060015. https://doi.org/10.1063/1.4954620.

[10] Khan S, Alghulaiakh H. ARIMA Model for Accurate Time Series Stocks Forecasting. IJACSA) Int J Adv Comput Sci Appl n.d.;11:2020.

[11] Ab Razak NH, Aris AZ, Ramli MF, Looi LJ, Juahir H. Temporal flood incidence forecasting for Segamat River (Malaysia) using autoregressive integrated moving average modelling. J Flood Risk Manag 2018;11:S794–804. https://doi.org/10.1111/JFR3.12258.

[12] Das R, Middya AI, Roy S. High granular and short term time series forecasting of PM 2.5 air pollutant - a comparative review. Artif Intell Rev 2022;55:1253–87. https://doi.org/10.1007/S10462-021-09991-1/TABLES/7.

[13] Zhao X, Chen Y, Xu C, Dharmawan PAS, Indradewi AAD. Double exponential smoothing brown method towards sales forecasting system with a linear and non-stationary data trend. J Phys Conf Ser 2021;1810:012026. https://doi.org/10.1088/1742-6596/1810/1/012026.

[14] Sulandari W, Suhartono, Subanar, Rodrigues PC. Exponential Smoothing on Modeling and Forecasting Multiple Seasonal Time Series: An Overview. Https://DoiOrg/101142/S0219477521300032 2021;20. https://doi.org/10.1142/S0219477521300032.

[15] Xie J, Wang Q. Benchmarking Machine Learning Algorithms on Blood Glucose Prediction for Type i Diabetes in Comparison with Classical Time-Series Models. IEEE Trans Biomed Eng 2020;67:3101–24. https://doi.org/10.1109/TBME.2020.2975959.

[16] Karthikeswaren R, Kayathwal K, Dhama G, Arora A. A Survey on Classical and Deep Learning based Intermittent Time Series Forecasting Methods. Proc Int Jt Conf Neural Networks 2021;2021-July. https://doi.org/10.1109/IJCNN52387.2021.9533963.

[17] Cheng W, Wang Y, Peng Z, Ren X, Shuai Y, Zang S, et al. High-efficiency chaotic time series prediction based on time convolution neural network. Chaos, Solitons & Fractals 2021;152:111304. https://doi.org/10.1016/J.CHAOS.2021.111304.

[18] Tealab A. Time series forecasting using artificial neural networks methodologies: A systematic

review. Futur Comput Informatics J 2018;3:334–40. https://doi.org/10.1016/J.FCIJ.2018.10.003.

[19] Ramadevi B, Bingi K. Chaotic Time Series Forecasting Approaches Using Machine Learning Techniques: A Review. Symmetry (Basel) 2022;14:955. https://doi.org/10.3390/SYM14050955.

[20] Ensafi Y, Amin SH, Zhang G, Shah B. Time-series forecasting of seasonal items sales using machine learning – A comparative analysis. Int J Inf Manag Data Insights 2022;2:100058. https://doi.org/10.1016/J.JJIMEI.2022.100058.

[21] Syrgkanis V, Zampetakis M, Abernethy J, Agarwal S. Estimation and Inference with Trees and Forests in High Dimensions. Proc Mach Learn Res 2020;125:3453–4. https://doi.org/10.1214/15-AOS1321.

[22] Konzen E, Ziegelmann FA. LASSO-Type Penalties for Covariate Selection and Forecasting in Time Series. J Forecast 2016;35:592–612. https://doi.org/10.1002/FOR.2403.

[23] Corsi F. A Simple Approximate Long-Memory Model of Realized Volatility. J Financ Econom 2009;7:174–96. https://doi.org/10.1093/jjfinec/nbp001.

[24] Masini RP, Medeiros MC, Mendes EF. Machine learning advances for time series forecasting. J Econ Surv 2023;37:76–111. https://doi.org/10.1111/JOES.12429.

[25] Semenoglou A-A, Spiliotis E, Assimakopoulos V. Image-based time series forecasting: A deep convolutional neural network approach. Neural Networks 2023;157:39–53. https://doi.org/10.1016/J.NEUNET.2022.10.006.

[26] Garg A, Zhang W, Samaran J, Savitha R, Foo CS. An Evaluation of Anomaly Detection and Diagnosis in Multivariate Time Series. IEEE Trans Neural Networks Learn Syst 2022;33:2508–17. https://doi.org/10.1109/TNNLS.2021.3105827.

[27] De Oliveira JFL, Silva EG, De Mattos Neto PSG. A Hybrid System Based on Dynamic Selection for Time Series Forecasting. IEEE Trans Neural Networks Learn Syst 2022;33:3251–63. https://doi.org/10.1109/TNNLS.2021.3051384.

[28] Cichos F, Gustavsson K, Mehlig B, Volpe G. Machine learning for active matter. Nat Mach Intell 2020;2:94–103. https://doi.org/10.1038/s42256-020-0146-9.

[29] Lim B, Zohren S. Time-series forecasting with deep learning: a survey. Philos Trans R Soc A 2021;379. https://doi.org/10.1098/RSTA.2020.0209.

[30] Ismail Fawaz H, Forestier G, Weber J, Idoumghar L, Muller PA. Deep learning for time series classification: a review. Data Min Knowl Discov 2019 334 2019;33:917–63. https://doi.org/10.1007/S10618-019-00619-1.

[31] Torres JF, Hadjout D, Sebaa A, Martínez-Álvarez F, Troncoso A. Deep Learning for Time Series Forecasting: A Survey. Big Data 2021;9:3–21.

https://doi.org/10.1089/BIG.2020.0159/ASSET/IMAGES/LARGE/BIG.2020.0159_FIGURE10.JPEG.

[32] Lara-Benítez P, Carranza-García M, Riquelme JC. An Experimental Review on Deep Learning Architectures for Time Series Forecasting. Int J Neural Syst 2021;31.

[33] Liang W, Tadesse GA, Ho D, Li FF, Zaharia M, Zhang C, et al. Advances, challenges and opportunities in creating data for trustworthy AI. Nat Mach Intell 2022 48 2022;4:669–77. https://doi.org/10.1038/s42256-022-00516-1.

[34] Chan KM, Chong ZL, C Khoo MB, - al, Sun AY, Scanlon BR. How can Big Data and machine learning benefit environment and water management: a survey of methods, applications, and future directions. Environ Res Lett 2019;14:073001. https://doi.org/10.1088/1748-9326/AB1B7D.

[35] Zhou H, Zhang S, Peng J, Zhang S, Li J, Xiong H, et al. Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting. 35th AAAI Conf Artif Intell AAAI 2021 2021;12B:11106–15. https://doi.org/10.1609/aaai.v35i12.17325.

[36] Lazcano A, Herrera PJ, Monge M. A Combined Model Based on Recurrent Neural Networks and Graph Convolutional Networks for Financial Time Series Forecasting. Math 2023, Vol 11, Page 224 2023;11:224. https://doi.org/10.3390/MATH11010224.

[37] Gasparin A, Lukovic S, Alippi C. Deep learning for time series forecasting: The electric load case. CAAI Trans Intell Technol 2022;7:1–25. https://doi.org/10.1049/CIT2.12060.

[38] Du L, Gao R, Suganthan PN, Wang DZW. Bayesian optimization based dynamic ensemble for time series forecasting. Inf Sci (Ny) 2022;591:155–75. https://doi.org/10.1016/J.INS.2022.01.010.

[39] Alassafi MO, Jarrah M, Alotaibi R. Time series predicting of COVID-19 based on deep learning. Neurocomputing 2022;468:335–44. https://doi.org/10.1016/J.NEUCOM.2021.10.035.

[40] Rahimilarki R, Gao Z, Jin N, Zhang A. Convolutional neural network fault classification based on time-series analysis for benchmark wind turbine machine. Renew Energy 2022;185:916–31. https://doi.org/10.1016/J.RENENE.2021.12.056.

[41] BrophyEoin, WangZhengwei, SheQi, WardTomás. Generative Adversarial Networks in Time Series: A Systematic Literature Review. ACM Comput Surv 2021.

[42] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville and YB. Generative Adversarial Nets. Adv. neural Inf. pro- cessing Syst., 2014, p. 2672–2680.

[43] Dahl CM, Sørensen EN. Time series (re)sampling using Generative Adversarial Networks. Neural Networks 2022;156:95–107. https://doi.org/10.1016/J.NEUNET.2022.09.010.

[44] Yoon J, Jarrett D, van der Schaar M. Time-series Generative Adversarial Networks. Adv Neural Inf

Process Syst 2019;32.

[45] Cichosz SL, Xylander AAP. A Conditional Generative Adversarial Network for Synthesis of Continuous Glucose Monitoring Signals. J Diabetes Sci Technol 2021:1_4.

[46] Li H, Xu Y, Ke D, Su K. μ-law SGAN for generating spectra with more details in speech enhancement. Neural Networks 2021;136:17–27. https://doi.org/10.1016/J.NEUNET.2020.12.017.

[47] Zhou X, Pan Z, Hu G, Tang S, Zhao C. Stock Market Prediction on High-Frequency Data Using Generative Adversarial Nets. Math Probl Eng 2018;2018. https://doi.org/10.1155/2018/4907423.

[48] Rüttgers M, Lee S, Jeon S, You D. Prediction of a typhoon track using a generative adversarial network and satellite images. Sci Reports 2019 91 2019;9:1–15. https://doi.org/10.1038/s41598-019-42339-y.

[49] Koochali A, Schichtel P, Dengel A, Ahmed S. Probabilistic Forecasting of Sensory Data with Generative Adversarial Networks - ForGAN. IEEE Access 2019;7:63868–80. https://doi.org/10.1109/ACCESS.2019.2915544.

[50] Han J, Wang C. SSR-TVD: Spatial Super-Resolution for Time-Varying Data Analysis and Visualization. IEEE Trans Vis Comput Graph 2022;28:2445–56. https://doi.org/10.1109/TVCG.2020.3032123.

[51] Elazab A, Wang C, Gardezi SJS, Bai H, Hu Q, Wang T, et al. GP-GAN: Brain tumor growth prediction using stacked 3D generative adversarial networks from longitudinal MR Images. Neural Networks 2020;132:321–32. https://doi.org/10.1016/J.NEUNET.2020.09.004.

[52] Chu J, Dong W, Huang Z. Endpoint prediction of heart failure using electronic health records. J Biomed Inform 2020;109:103518. https://doi.org/10.1016/J.JBI.2020.103518.

[53] Cheng C, Sa-Ngasoongsong A, Beyca O, Le T, Yang H, Kong Z, et al. Time series forecasting for nonlinear and non-stationary processes: a review and comparative study. IIE Trans 2015;47:1053–71. https://doi.org/10.1080/0740817X.2014.999180.

[54] Zhu T, Wang W, Yu M. A novel blood glucose time series prediction framework based on a novel signal decomposition method. Chaos, Solitons & Fractals 2022;164:112673. https://doi.org/10.1016/J.CHAOS.2022.112673.

[55] Khadem H, Nemat H, Elliott J, Benaissa M. Signal fragmentation based feature vector generation in a model agnostic framework with application to glucose quantification using absorption spectroscopy. Talanta 2022;243:123379. https://doi.org/10.1016/j.talanta.2022.123379.

[56] Woldaregay AZ, Årsand E, Botsis T, Albers D, Mamykina L, Hartvigsen G. Data-Driven Blood Glucose Pattern Classification and Anomalies Detection: Machine-Learning Applications in Type 1 Diabetes. J Med Internet Res 2019;21:e11030. https://doi.org/10.2196/11030.

[57] Khadem H, Nemat H, Elliott J, Benaissa M. Multi-Lag Stacking for Blood Glucose Level Prediction. Knowl. Discov. Healthc. Data, vol. 2675, 2020, p. 146–50.

[58] Nemat H, Khadem H, Elliott J, Benaissa M. Data fusion of activity and CGM for predicting blood glucose levels. Knowl. Discov. Healthc. Data, vol. 2675, 2020, p. 120–4.

[59] Marcus Y, Eldor R, Yaron M, Shaklai S, Ish-Shalom M, Shefer G, et al. Improving blood glucose level predictability using machine learning. Diabetes Metab Res Rev 2020;36. https://doi.org/10.1002/dmrr.3348.

[60] Nemat H, Khadem H, Eissa MR, Elliott J, Benaissa M. Blood Glucose Level Prediction: Advanced Deep-Ensemble Learning Approach. IEEE J Biomed Heal Informatics 2022;26:2758–69. https://doi.org/10.1109/JBHI.2022.3144870.

[61] Woldaregay AZ, Årsand E, Walderhaug S, Albers D, Mamykina L, Botsis T, et al. Data-driven modeling and prediction of blood glucose dynamics: Machine learning applications in type 1 diabetes. Artif Intell Med 2019;98:109–34. https://doi.org/10.1016/j.artmed.2019.07.007.

[62] Zhu T, Yao X, Li K, Herrero P, Georgiou P. Blood glucose prediction for type 1 diabetes using generative adversarial networks. CEUR Workshop Proc 2020;2675:90–4.

[63] Zhang D, Ma M, Xia L. A comprehensive review on GANs for time-series signals. Neural Comput Appl 2022;34:3551–71. https://doi.org/10.1007/S00521-022-06888-0/FIGURES/9.

[64] Brophy E, Wang Z, She Q, Ward T. Generative Adversarial Networks in Time Series: A Systematic Literature Review. ACM Comput Surv 2022;55:31. https://doi.org/10.1145/3559540.

[65] Brophy E, Wang Z, Lab BA, Qi C, Bytedance S, Lab AI, et al. Generative adversarial networks in time series: A survey and taxonomy 2021.

[66] Festag S, Denzler J, Spreckelsen C. Generative adversarial networks for biomedical time series forecasting and imputation: A systematic review. J Biomed Inform 2022;129:104058. https://doi.org/10.1016/j.jbi.2022.104058.

[67] Takahashi S, Chen Y, Tanaka-Ishii K. Modeling financial time-series with generative adversarial networks. Phys A Stat Mech Its Appl 2019;527:121261. https://doi.org/10.1016/J.PHYSA.2019.121261.

[68] Geiger A, Liu D, Alnegheimish S, Cuesta-Infante A, Veeramachaneni K. TadGAN: Time Series Anomaly Detection Using Generative Adversarial Networks. Proc - 2020 IEEE Int Conf Big Data, Big Data 2020 2020:33–43. https://doi.org/10.1109/BIGDATA50022.2020.9378139.

[69] Luo Y, Cai X, Zhang Y, Xu J, Yuan X. Multivariate Time Series Imputation with Generative Adversarial Networks. 32nd Conf. Neural Inf. Process. Syst., n.d.

[70] Huang X, Li Q, Tai Y, Chen Z, Liu J, Shi J, et al. Time series forecasting for hourly photovoltaic

power using conditional generative adversarial network and Bi-LSTM. Energy 2022;246:123403. https://doi.org/10.1016/J.ENERGY.2022.123403.

[71] Zhou K, Wang W, Hu T, Deng K. Time Series Forecasting and Classification Models Based on Recurrent with Attention Mechanism and Generative Adversarial Networks. Sensors 2020, Vol 20, Page 7211 2020;20:7211. https://doi.org/10.3390/S20247211.

[72] Hua Y, Zhao Z, Li R, Chen X, Liu Z, Zhang H. Deep Learning with Long Short-Term Memory for Time Series Prediction. IEEE Commun Mag 2019;57:114–9. https://doi.org/10.1109/MCOM.2019.1800155.

[73] Liu CL, Hsaio WH, Tu YC. Time Series Classification with Multivariate Convolutional Neural Network. IEEE Trans Ind Electron 2019;66:4788–97. https://doi.org/10.1109/TIE.2018.2864702.

[74] Marling C, Bunescu R. The OhioT1DM Dataset for Blood Glucose Level Prediction: Update 2020. Int. Work. Knowl. Discov. Healthc. Data, vol. 2675, NIH Public Access; 2020, p. 71–4.

[75] Daniels J, Herrero P, Georgiou P. A Multitask Learning Approach to Personalized Blood Glucose Prediction. IEEE J Biomed Heal Informatics 2022;26:436–45. https://doi.org/10.1109/JBHI.2021.3100558.

[76] Yang T, Yu X, Ma N, Wu R, Li H. An autonomous channel deep learning framework for blood glucose prediction. Appl Soft Comput 2022;120:108636. https://doi.org/10.1016/j.asoc.2022.108636.

[77] Zhu T, Li K, Chen J, Herrero P, Georgiou P. Dilated Recurrent Neural Networks for Glucose Forecasting in Type 1 Diabetes. J Healthc Informatics Res 2020;4:308–24. https://doi.org/10.1007/s41666-020-00068-2.

[78] Martinsson J, Schliep A, Eliasson B, Mogren O. Blood Glucose Prediction with Variance Estimation Using Recurrent Neural Networks. J Healthc Informatics Res 2020;4:1–18. https://doi.org/10.1007/S41666-019-00059-Y/FIGURES/8.

[79] Shuvo MMH, Islam SK. Deep Multitask Learning by Stacked Long Short-Term Memory for Predicting Personalized Blood Glucose Concentration. IEEE J Biomed Heal Informatics 2023:1–12. https://doi.org/10.1109/JBHI.2022.3233486.

[80] Nemat H, Khadem H, Elliott J, Benaissa M. Causality analysis in type 1 diabetes mellitus with application to blood glucose level prediction. Comput Biol Med 2023;153:106535. https://doi.org/10.1016/j.compbiomed.2022.106535.

[81] Jeon J, Leimbigler PJ, Baruah G, Li MH, Fossat Y, Whitehead AJ. Predicting Glycaemia in Type 1 Diabetes Patients: Experiments in Feature Engineering and Data Imputation. J Healthc Informatics Res 2019 41 2019;4:71–90. https://doi.org/10.1007/S41666-019-00063-2.

[82] HYNDMAN RJ, ATHANASOPOULOS. Forecasting: principles and practice - Rob J Hyndman, George Athanasopoulos. OTexts 2018. https://books.google.co.uk/books?hl=en&lr=&id=_bBhDwAAQBAJ&oi=fnd&pg=PA7&dq=Forecasting+Principles+and+Practice&ots=Til1ykZIEH&sig=3JpyoZzvfMbDNMNpJ9wP7B__1cc#v=onepage&q=Forecasting Principles and Practice&f=false (accessed March 12, 2023).

[83] Souza RC. Practical Time Series Analysis Prediction with Statistics and Machine Learning. vol. 21. 2001.

[84] Brownlee J. Time series forecasting as supervised learning. 2016.

[85] Klonoff DC, Lias C, Vigersky R, Clarke W, Parkes JL, Sacks DB, et al. The surveillance error grid. J Diabetes Sci Technol 2014;8:658–72. https://doi.org/10.1177/1932296814539589.

[86] Zhu Q. On the performance of Matthews correlation coefficient (MCC) for imbalanced dataset. Pattern Recognit Lett 2020;136:71–80. https://doi.org/10.1016/J.PATREC.2020.03.030.

[87] Someetheram V, Marsani MF, Mohd Kasihmuddin MS, Zamri NE, Muhammad Sidik SS, Mohd Jamaludin SZ, et al. Random Maximum 2 Satisfiability Logic in Discrete Hopfield Neural Network Incorporating Improved Election Algorithm. Math 2022, Vol 10, Page 4734 2022;10:4734. https://doi.org/10.3390/MATH10244734.

[88] Friedman M. A Comparison of Alternative Tests of Significance for the Problem of m Rankings on JSTOR. Ann Math Stat 1940;11:86–92.

[89] Fisher R. Statistical Methods and Scientific Induction. J R Stat Soc Ser B 1955;17:69–78. https://doi.org/10.1111/J.2517-6161.1955.TB00180.X.

[90] De Bois M, Yacoubi MAE, Ammi M. GLYFE: review and benchmark of personalized glucose predictive models in type 1 diabetes. Med Biol Eng Comput 2022;60:1–17. https://doi.org/10.1007/s11517-021-02437-4.

[91] Van Rossum G, Drake FL. Python 3 Reference Manual. Scotts Valley, CA: CreateSpace; 2009.

[92] Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, et al. Tensorflow: A system for large-scale machine learning. 12th Symp. Oper. Syst. Des. Implement., 2016, p. 265–83.

[93] Chollet F, others. Keras 2015.

[94] McKinney W. Data structures for statistical computing in python. 9th Python Sci. Conf., vol. 445, 2010, p. 51–6.

[95] Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D, et al. Array programming with {NumPy}. Nature 2020;585:357–362. https://doi.org/10.1038/s41586-020-2649-2.

[96] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine

learning in Python. J Mach Learn Res 2011;12:2825–30.

[97] Seabold S, Perktold J. statsmodels: Econometric and statistical modeling with python. 9th Python Sci. Conf., 2010.

Publication 2.

Blood Glucose Level Time Series Forecasting: Nested Deep Ensemble Learning Lag Fusion[2]

Abstract. Blood glucose level prediction is a critical aspect of diabetes management. It enables individuals to make informed decisions about their insulin dosing, diet, and physical activity. This, in turn, improves their quality of life and reduces the risk of chronic and acute complications. One conundrum in developing time-series forecasting models for blood glucose level prediction is to determine an appropriate length for look-back windows. On the one hand, studying short histories foists the risk of information incompletion. On the other hand, analysing long histories might induce information redundancy due to the data shift phenomenon. Additionally, optimal lag lengths are inconsistent across individuals because of the domain shift occurrence. Therefore, in bespoke analysis, either optimal lag values should be found for each individual separately or a global suboptimal lag value should be used for all. The former approach degenerates the analysis's congruency and imposes extra perplexity. With the latter, the fine-tunned lag is not necessarily the optimum option for all individuals. To cope with this challenge, this work suggests an interconnected lag fusion framework based on nested meta-learning analysis that improves the accuracy and precision of predictions for personalised blood glucose level forecasting. The proposed framework is leveraged to generate blood glucose prediction models for patients with type 1 diabetes by scrutinising two well-established publicly available Ohio type 1 diabetes datasets. The models developed undergo vigorous evaluation and statistical analysis from mathematical and clinical perspectives. The results achieved underpin the efficacy of the proposed method in blood glucose level time-series prediction analysis.

---

[2] This article was published and featured in Bioengineering. 10 (2023) 1–22. Authors: **H. Khadem**, H. Nemat, J. Elliott, M. Benaissa.

Keywords. Deep Learning; Time-Series Forecasting; Blood Glucose; Diabetes; Ensemble Learning; Artificial Neural Network

1. Introduction

Type 1 diabetes is a chronic metabolic disorder [1]. The disease is currently incurable [2,3]. Nevertheless, its effective management can dramatically mitigate the symptoms and the risk of associated short-term and long-term complications [4,5]. Accordingly, people with type 1 diabetes and their potential carers are normally educated on the standard practices to control the illness [6–8].

Self-management of type 1 diabetes is, however, burdensome and prone to human errors [9–11]. Hence, automating the management tasks would be highly beneficial [12,13]. Some developments have already been made related to this concern [14–16]. For example, technological breakthroughs, such as continuous glucose monitoring biosensors [17,18] and insulin pumps [19,20], nowadays, serve myriads of type 1 diabetes patients. The former, in a minimally invasive fashion, takes regular snapshots of blood glucose levels in alignment with the general advice on a frequent review of glycaemic state [21,22]. The latter semiautomates insulin administration, requiring minimum user interference [23–25]. Moreover, there are ongoing efforts to develop fully noninvasive continuous blood glucose level monitoring sensors to help more effective diabetes management [26–29].

Despite the advancements achieved so far, continued progress in the automation process is still demanded to further facilitate and effectuate the management of type 1 diabetes [30,31]. In this respect, engineering accurate blood glucose predictor devices would be game-changing [32,33]. Such instruments can provide early warning about possible adverse glycaemic events so that automated or nonautomated pre-emptive measures can be taken [34,35]. Additionally, these devices are a prerequisite for the advent of a closed-loop artificial pancreas as the current vision for the ultimate automated management of type 1 diabetes [36,37].

For predicting blood glucose levels, physiological, data-driven, and hybrid modelling approaches can be pursued [38,39]. In the data-driven approach, also used in this research, current and past values of diabetes-management-related variables are studied to project future blood glucose excursion [38,40].

For constructing data-driven blood glucose level predictors, one of the three main categories of time-series forecasting approaches is typically used: classical time-series forecasting, traditional machine learning, or deep learning analysis. Among these, deep learning, as a member of the modern artificial intelligence family, has proven potency in solving complicated computational tasks, including complex time-series forecasting [41–46].

Predicting the blood glucose levels of individuals with type 1 diabetes is a convoluted forecasting mission due to the highly erratic behaviour of the phenomenon [47]. Thus, in line with many other time-series forecasting areas, deep learning has gained enormous popularity in the blood glucose level prediction realm [48,49]. Subsequently, extensive research has been underway to advance the analysis. Notwithstanding all the enhancements in this field so far, there still exist challenges to be addressed adequately [50]. This work contributes to addressing one such challenge.

When applying deep learning algorithms for data-driven time-series blood glucose level forecasting, lag observations of data are studied to predict specific future values. Here, a quandary is to select the appropriate length of history to be investigated. This issue is even more pronounced when considering the fact that due to the significant discrepancy in the blood glucose profile across type 1 diabetes patients, the common practice is to generate personalised models. In this circumstance, finding an optimal length of history separately for each individual entails further disparity and complexity in the analysis. To address this difficulty, the present work suggests a compound lag fusion approach by exploiting the potential of nested ensemble learning over typical ensemble learning analysis. This is the first paper, to the best of our knowledge, that incorporates nested meta-learning analysis in the field of blood glucose level prediction.

The rest of the article is outlined as follows. Section 2 reviews some recent studies on type 1 diabetes blood glucose level prediction. Section 3 concisely describes the datasets used in this research. Section 4

explains model development and assessment analysis. Section 5 presents the results of the model assessment analysis along with the relevant discussions. Finally, Section 6 summarises and concludes the work.

2. Literature Survey

In the following, a number of recent articles on data-driven blood glucose level prediction are succinctly overviewed. For further alignment with the contents of this study, the focus of this overview is on the application of state-of-the-art machine learning techniques and the use of Ohio type 1 diabetes datasets for model development and evaluation. A more comprehensive review of the latest revolutions in the blood glucose level prediction area can be studied at these references [51–54].

A recent article offered a multitask approach for blood glucose level prediction by experimenting on the Ohio datasets [55]. The methods are based on the concept of transfer learning. The study explicitly targets addressing the challenge of the need for extensively large amounts of data for personalised blood glucose level prediction. For this purpose, it suggests pre-training a model on a source domain and a multitask model on the whole dataset and then using these learning experiences in constructing personalised models. The authors showcase the efficacy of their propositions by comparing the performance of their approach with sequential transfer learning and subject-isolated learning.

An autonomous channel setup was recently presented for deep learning blood glucose level prediction using the Ohio datasets [56]. The proposed method chose the history lengths for different variables adaptively by affecting the time-dependency scale. The crux is to avoid dismissing useful information from variables with enduring influence and engaging uninformative data from variables with transient impact at the same time. The models generated in the study undergo comparison analysis with standard non-autonomous channel structures deploying mathematical and clinical assessments.

A deep learning approach based on dilated recurrent neural networks accompanied by transfer learning concepts is introduced for blood glucose level prediction [57]. In the study, personalised models are created

for individuals with type 1 diabetes using an Ohio dataset. The method is examined for short-term forecasting tasks. Its supremacy over standard methods, including autoregressive models, support vector regression, and conventional neural networks, is shown.

Another study suggests an efficient method for univariate blood glucose level prediction [58]. In the analysis, recurrent neural networks were used as learners. The learners are trained in an end-to-end approach to predict future blood glucose levels 30 and 60 minutes in advance using only histories of blood glucose data. The models are developed and assessed using an Ohio dataset. The results achieved are comparable with the state-of-the-art research on the dataset. In addition to accuracy analysis, the study investigates the certainty of predictions. To do so, a parameterised univariate Gaussian is tasked with calculating the standard deviation of the predictions as a representative of uncertainty.

Employing the concepts of the Internet of things, a study compares four broadly used models of glycaemia, including support vector machine, Bayesian regularised neural network, multilayer perceptron, and Gaussian approach [59]. These models are used to investigate the possibility of completing the data collected from 25 individuals with type 1 diabetes by mapping intricate patterns of data. The findings highlight the potential of such analysis in contributing to improved diabetes management. Further, among the approaches examined, Bayesian regularised neural networks outperform others by delivering the best root mean square error and coefficient of determination.

3. Material

For generating blood glucose level prediction models, this study uses two well-established, publicly accessible Ohio type 1 diabetes datasets [60]. The first dataset includes data for six individuals with type 1 diabetes. The participants' age at the time of data collection was in the range of 40 to 60 years. The sample comprised four females and two males. This dataset was initially released for the first blood glucose challenge in Knowledge Discovery at the Healthcare Data conference in 2018. This dataset is referred to as the Ohio 2018 dataset hereafter. The second dataset also contains six people with type 1 diabetes, different

from those in the first dataset. The data contributors in this dataset were in an age range of 20 to 80 years at the point of data acquisition. Five of them were male and one female. This dataset was originally distributed for the second blood glucose level prediction challenge in Knowledge Discovery at the Healthcare Data conference in 2020. Hereafter, we refer to this dataset as the Ohio 2020 dataset.

Both datasets contain diabetes-related modalities, including blood glucose, physical activity, carbohydrate intake, and bolus insulin injection. Blood glucose and bolus insulin data were collected automatically using physiological sensors. For the former, a Medtronic Enlite continuous glucose monitoring device was used. For the latter, patients in the Ohio 2018 dataset wore a Basis Peak fitness band that collected heart rate data as a representative of physical activity. Alternatively, subjects in the Ohio 2020 dataset wore an Empatica Embrace fitness band that tracked the magnitude of acceleration as a representative of physical activity data. On the other hand, carbohydrate and bolus insulin data were self-reported by individuals in both datasets.

In both datasets, data were collected for eight weeks. The data come with the training and testing set already separated by the data collection and distribution team. The last ten days of data are allocated as a testing set and the remaining former data points as the training set. In the present study, using training sets only, bespoke predictive models are created for future values of blood glucose levels from historical values of blood glucose itself as the indigenous variable, along with exogenous variables of physical activity, carbohydrate intake, and bolus insulin injection. The testing sets are then used to evaluate the generated models. Table 1 displays individuals' identification number, sex, and age information together with a short representation of the statistical properties of blood glucose as the intrinsic variable in the dataset. A more comprehensive description of the Ohio datasets and the data collection process can be found in the original documentation [60].

Table 1. Demographic information of contributors and summary of statistical properties of blood glucose data (the focal modality) in the Ohio datasets.

Dataset	PID	Sex	Age	Set	Blood Glucose Data							
					Count	Range (mg/dL)	Mean (mg/dL)	SD (mg/dL)	MR (%)	HOR (%)	ER (%)	HRR (%)
2018	559	female	40–60	Train	10,655	40–400	167.53	70.44	12.06	3.65	55.98	40.37
				Test	2444	45–400	168.93	67.78	14.81	3.03	59.86	37.11
	563	male	40–60	Train	11,013	40–400	146.94	50.51	8.80	2.82	72.81	24.36
				Test	2569	62–313	167.38	46.15	4.71	0.70	60.45	38.85
	570	male	40–60	Train	10,981	46–377	187.5	62.33	5.73	1.97	42.97	55.07
				Test	2672	60–388	215.71	66.99	5.05	0.41	29.04	70.55
	575	female	40–60	Train	11,865	40–400	141.77	60.27	10.43	8.71	68.62	22.66
				Test	2589	40–342	150.49	60.53	4.94	5.37	63.50	31.13
	588	female	40–60	Train	12,639	40–400	164.99	50.51	3.69	1.04	63.56	35.40
				Test	2606	66–354	175.98	48.66	3.42	0.15	53.26	46.58
	591	female	40–60	Train	10,846	40–397	156.01	58.03	17.59	3.94	63.97	32.09
				Test	2759	43–291	144.83	51.42	3.15	5.18	67.27	27.55
2020	540	male	20–40	Train	11,914	40–369	136.78	54.75	9.76	7.08	72.66	20.25
				Test	2360	52–400	149.94	66.46	6.74	5.64	68.18	26.19
	544	male	40–60	Train	10,533	48–400	165.12	60.08	19.11	1.47	63.78	34.75
				Test	2715	62–335	156.48	54.14	15.47	1.22	68.29	30.50
	552	male	20–40	Train	8661	45–345	146.88	54.63	22.30	3.89	72.05	24.06
				Test	1792	47–305	138.11	50.23	85.71	3.57	80.02	16.41
	567	female	20–40	Train	10,750	40–400	154.43	60.88	24.91	6.75	63.40	29.84
				Test	2388	40–351	146.25	55.00	20.18	8.33	67.38	24.29
	584	male	40–60	Train	12,027	40–400	192.34	65.29	9.13	0.80	47.69	51.51
				Test	2661	41–400	170.48	60.76	12.40	1.01	61.86	37.13
	596	male	60–80	Train	10,858	40–367	147.17	49.34	25.35	2.08	73.99	23.93
				Test	2663	49–305	146.98	50.79	9.76	2.78	75.07	22.16

Note. PID: patient identification; SD: standard deviation; MR: missingness rate; HOR: hypoglycaemic rate; ER: euglycaemic rate; HRR: hyperglycaemic rate. Hypoglycaemia, euglycaemia, and hyperglycaemia refer to when the blood glucose level is, respectively, less than 70 mg/dL, between 70 and 180 mg/dL, and more than 180 mg/dL. Both hypoglycaemia and hyperglycaemia are adverse glycaemic events.

4. Methods

This section explicates the methodological implementations for blood glucose level prediction model generation and evaluation. First, some curation steps performed to prepare the data for formal prediction modelling analysis are explained. Next, time-series forecasting models constructed for blood glucose level prediction are described. After that, the criteria considered for evaluating the generated predictive models are presented. Finally, statistical analysis operated on the model outputs is outlined.

4.1. Data Curation

The following pre-modelling curation steps are operated on the raw data to render the ensuing formal deep learning prediction modelling analysis more effective.

4.1.1. Missingness Treatment

The first data curation stage deals with the missing values presented in the automatically collected blood glucose and physical activity data. At the beginning and end of the blood glucose and physical activity series, there are some timespans where data are absent. This unavailability occurred because the subject did not start and finish wearing the sensing devices exactly at the same time. As an initial missing value treatment step, the head and tail of all series are trimmed by removing the void timestamps so that variables start and end from the same point. Afterwards, the linear interpolation technique is used to fill in missing values in the training sets of blood glucose and physical activity. Alternatively, for the testing sets of these modalities, the linear extrapolation technique is used to fill in missing values. This technique precludes future value observation in the evaluation stage, so the models created possess applicability for real-time monitoring.

4.1.2. Sparsity Handling

The sparsity of the self-reported carbohydrate and bolus insulin data is the next pre-modelling issue to be addressed. A reasonable assumption as to the unavailable values of these modalities in the majority of timestamps is that there has been no occurrence to be reported in those points. Therefore, for these two modalities, as a simple yet acceptable practice, zero values are assigned to non-reported timestamps.

4.1.3. Data Alignment

Another data curation step is to unify the frequency of exogenous modalities and align their timestamps with the blood glucose level as the indigenous variable. Initially, acceleration data are downsampled from a one-minute frequency to a five-minute frequency. For this purpose, the entries in the nearest neighbourhood to blood glucose timestamps are kept, and the remaining data points are removed. Following that, timestamps of all extrinsic variables are aligned with those of blood glucose levels with the minimum possible shifts.

4.1.4. Data Transformation

As the next data curation step, as a common practice, feature values are converted into a standardised form that machine learning models can analyse more effectively. For each variable, first, the average of training set values is subtracted from all values in both the training and testing sets. Then, all obtained values are divided by the standard deviation of the training set to make unit variance variables.

4.1.5. Stationarity Inspection

Stationary time-series data have statistical characteristics, including variance and mean, that do not change over time. In this data treatment step, the stationarity condition in the time-series data is satisfied. By conducting the feature transformation step explained in Section 4.1.4, the variances in the series are stabilised. To stabilise the mean of the series, the first-order differencing method is applied. Subsequently, the outcomes are examined using two prevalent statistical tests of Kwiatkowski–Phillips–Schmidt–Shin [61] and Augmented Dickey–Fuller [62], where both confirm the stationary of the series.

4.1.6. Problem Reframing

The final data curation phase translates the time-series blood glucose level prediction question to the supervised machine learning language. Hence, pairs of independent and dependent variables need to be constructed from the time-series data. To this end, a rolling window approach is used to appoint sequences of lag observations for blood glucose, physical activity, carbohydrate, and bolus insulin as the independent variables and sequences of blood glucose in the prediction horizon as the dependent variable.

5. Modelling

This subsection describes time-series forecasting models created for blood glucose level prediction 30 and 60 min into the future. This work undertakes a sequence-to-sequence fashion for multi-step-ahead time-series prediction. Prior to explaining the formal modelling process, it is useful to provide a brief explanation of stacking as an ensemble learning variation used in this work.

5.1. Preliminary

Ensemble learning is an advanced machine learning method that attempts to improve analysis performance by combining the decisions of multiple models [63]. Stacking is a type of ensemble learning in which a meta-learner intakes predictions of a number of base learners as an input feature to make final decisions [64].

5.2. Model Development

The diagram in Figure 1 displays the procedure contrived in this work for model creation. According to the diagram, the models are constructed by training three categories of learners: non-stacking, stacking, and nested stacking. The models generated based on the block diagram in Figure 1 are described below.

A non-stacking model takes a specific length of historical blood glucose, physical activity, carbohydrate, and bolus insulin data as multivariate input and returns a sequence of forecasted future blood glucose levels over a predefined prediction horizon of 30 or 60 min. According to the diagram in Figure 1, for each prediction horizon of 30 and 60 min, eight non-stacking models are created in aggregate. For this purpose, a multilayer perceptron network and a long short-term memory network are trained separately on four different lag lengths of 30, 60, 90, and 120 min.

A stacking model is a meta-model that takes sequence predictions from four non-stacking models with a homogenous learner (multilayer perceptron network or long short-term memory network) as multivariate input and fuses them to generate new prediction outputs. According to v, for each prediction horizon of 30 and 60 min, two stacking models are created, one with multilayer perceptron networks and the other with long short-term memory networks as the underlying embedded learners.

A nested stacking model is a nested meta-model. It receives the outcomes of the two stacking models described above as multivariate inputs and returns new predictions. As can be seen in Figure 1, two nested stacking models are generated for each prediction horizon of 30 and 60 min; one employs a multilayer perceptron network and the other a long short-term memory network as the nested stacking learner.

According to Figure 1, in all model creation scenarios, the learners recruited are either multilayer perceptron or long short-term memory networks. For simplicity and coherency, all multilayer perceptron networks have similar architectures consisting of an input layer, a hidden dense layer with 100 nodes, followed by another dense layer as output. Additionally, all long short-term memory networks are the vanilla type with an input layer, a hidden 100-node LSTM layer, and a dense output layer. Given the five-minute resolution of time-series data investigated, the number of nodes in the output layer is 6 and 12 for 30 min and 60 min prediction horizons, respectively. In all networks, He uniform is set as the initialiser, Adam as the optimiser, ReLU as the activation function, and mean square error as the loss function. Moreover, in all training scenarios, epoch size and batch size are set to 100 and 32, respectively. In addition, the learning rate is initiated from 0.01, and then using the ReduceLROnPlateau callback, it is reduced by a factor of 0.1 once the validation loss reduction stagnates with the patience of ten iterations.

5.3. Model Assessment

This section describes the analyses performed to validate the functionality of the developed blood glucose level prediction models. The generated models are assessed from regression, clinical, and statistical perspectives, as discussed below.

5.3.1. Regression Evaluation

Four broadly applied regression metrics are determined to verify the performance of the constructed models from a mathematical viewpoint. Mean absolute error (Equation (1)), root mean square error (Equation (2)), and mean absolute percentage error (Equation (3)) rate the accuracy of predictions. Further, the coefficient of determination (Equation (4)) measures the correlation between the reference and predicted blood glucose levels.

$$\mathrm{MAE} = \left(\sum_{i=1}^{N} \left| \mathrm{BGL_i} - \widehat{\mathrm{BGL}}_i \right| \right) / \mathrm{N} \tag{1}$$

$$\mathrm{RMSE} = \sqrt{(\sum_{i=1}^{N}(\mathrm{BGL_i} - \widehat{\mathrm{BGL}}_i)^2)/N} \tag{2}$$

$$\mathrm{MAPE} = ((\sum_{i=1}^{N}\left|(\mathrm{BGL_i} - \widehat{\mathrm{BGL}}_i)/\mathrm{BGL_i}\right|)/N) \times 100 \tag{3}$$

$$r^2 = 1 - ((\sum_{i=1}^{N}(\mathrm{BGL_i} - \widehat{\mathrm{BGL}}_i)^2)(\sum_{i=1}^{N}(\mathrm{BGL_i} - \overline{\mathrm{BGL}})^2)) \tag{4}$$

where MAE: mean absolute error; BGL: blood glucose level; N: the size of the testing set; RMSE: root mean square error; MAPE: mean absolute prediction error; r^2: coefficient of determination.

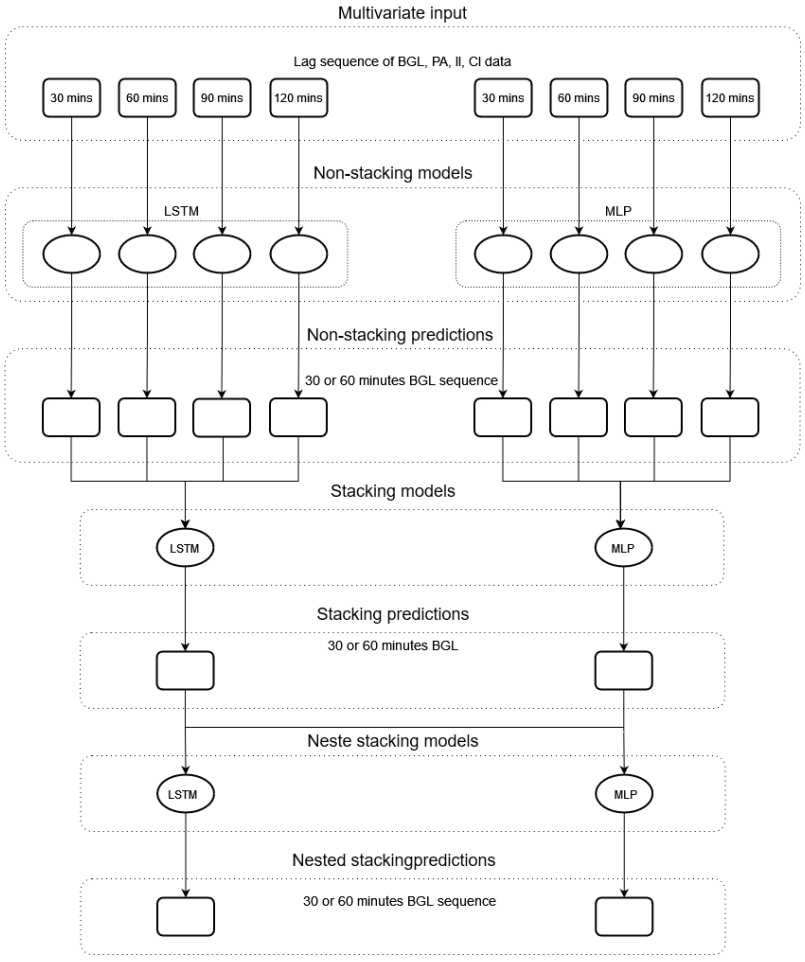

Figure 1. Blueprint for generating non-stacking, stacking, and nested stacking blood glucose level prediction models. Rectangular and oval blocks represent sequences of lag or future data and regression learners, respectively. Note. BGL: blood glucose level; PA: physical activity; II: insulin injection; CI: carbohydrate intake; LSTM: long short-term memory; MLP: multilayer perceptron.

5.3.2. Clinical Evaluation

Two criteria are employed to evaluate the developed models from a clinical standpoint. One criterion is Matthew's correlation coefficient [65]. It is a factor fundamentally used for assessing the effectuality of binary classifications. In this work, this metric, calculated as Equation (5), is exploited to investigate the potency of the blood glucose prediction models in discriminating adverse glycaemic events from euglycaemic events. Hereby, an adverse glycaemic event is defined as a blood glucose level lower than 70 mg/dL (hypoglycaemia) or more than 180 mg/dL (hyperglycaemia), and a euglycaemia event as a blood glucose level between 70 mg/dL and 180 mg/dL.

$$\mathrm{MCC} = (\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}) / \sqrt{(\mathrm{TP} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})} \quad (5)$$

where TP: true positive (the count of correctly predicted adverse glycaemic events); TN: true negative (the count of correctly predicted euglycaemic events); FP: false positive (the count of falsely predicted adverse glycaemic events); FN: false negative (the count of falsely predicted euglycaemic events).

The other considered clinical evaluation criterion is surveillance error [66]. It is based on error grid analysis to identify the clinical risk of inaccuracies in blood glucose level predictions. Detailed calculations of surveillance error can be found in the original article [66]. However, a concise elucidation of the outcome of the calculations is as follows. A unitless error value is measured for each predicted blood glucose level. Errors smaller than 0.5 indicate clinically risk-free predictions. Errors between 0.5 and 1.5 indicate clinically slight-risk predictions. Errors between 1.5 and 2.5 indicate clinically moderate-risk predictions. Errors between 2.5 and 3.5 indicate clinically high-risk predictions. Finally, errors bigger than 3.5 indicate clinically critical-risk predictions. We adopt two evaluation metrics based on surveillance error calculation outcomes. One is the average of surveillance errors across the entire testing set, and the other is the proportion of obtained surveillance errors less than 0.5 (clinically riskless predictions) across the entire testing set.

5.3.3. Statistical Analysis

Statistical analysis is conducted for further side-by-side performance assessment for different models. In this sense, the non-parametric Friedman test is exercised to compare the outcomes of different models [67]. This test is privileged for inter-model comparative analysis across multiple datasets with no normality assumption requirement as opposed to the counterpart ANOVA test [68]. In this study, the test is assigned to compare the performance of different types of models considering individuals as independent data sources. To do so, a significant level of five percent is considered to examine the consistency of results achieved for evaluation metrics. The null hypothesis for the test is that the results of the non-stacking, stacking, and nested stacking models have identical distributions. In the next step, for cases where the global Friedman test detects the existence of a statistically significant difference amongst the models' performance, the local Nemenyi test [69], as a post hoc procedure, compares the models in a pairwise manner. In this multi-comparison analysis, the Holm–Bonferroni method is used to adjust the significance level [70]. Finally, the heuristic critical difference approach is employed to visualise the outcomes of the post hoc analysis [71]. The statistical tests are operated on all evaluation metrics in both prediction horizons of 30 and 60 min. Both multilayer perceptron and long short-term memory networks are examined as learners separately.

6. Results and Discussion

This section presents the outcomes of model assessment analyses and the relevant discussion. Initially, the results of regression-wise and clinical-wise evaluation investigations are given for the non-stacking, stacking, and nested stacking models. Therein, for each metric, mean and standard deviation values achieved over five model runs are reported, a common practice in deep learning to counteract the stochastic nature of the analysis. After presenting the evaluation results, the results of the statistical analysis performed for more detailed comparison inspections between different types of models are exhibited.

The full evaluation results of the non-stacking models are compartmentalised in four tables given in Appendix A. Table A1 is dedicated to models with multilayer perceptron learners created on the Ohio 2018 dataset, Table A2 to models with multilayer perceptron learners created on the Ohio 2020 dataset, Table A3 to models with long short-term memory learners created on the Ohio 2018 dataset, and Table A4 to models with long short-term memory learners created on the Ohio 2020.

In the non-stacking analysis, there are four modelling scenarios for each patient: blood glucose level prediction 30 and 60 min in advance, once assigning multilayer perceptron and once long short-term memory as the learner. As can be seen in the Appendix A tables, for each scenario, four models are created by training the learner on 30, 60, 90, or 120 min of historical data separately. Additionally, there are four parallel modelling scenarios for stacking and nested stacking analysis: blood glucose level prediction 30 and 60 min in advance, once employing multilayer perceptron and once long short-term memory as the last-level learner. On the other hand, one model is created for each scenario in stacking and nested stacking analysis because different lags are not separately studied.

To compare the stacking and nested stacking analyses with the non-stacking analyses, initially, for each patient, one of the four non-stacking models created for each modelling scenario is selected as the representative. Then, the representative non-stacking models are studied in parallel with the counterpart stacking and nested stacking models. To select the representative non-stacking models, first, the best evaluation metrics achieved in each modelling scenario are marked in bold font in the Appendix A tables. Subsequently, the model delivering the highest number of best-obtained evaluation metrics, highlighted in grey in the tables, is deemed as the representative. For eligibility, the results for these models are given in Table 2. Moreover, the complete evaluation results for the stacking and nested stacking models are recorded in Table 3 and Table 4 respectively.

After picking the representative non-stacking models, the overall performance of these models is compared with the stacking and nested stacking counterparts. To this end, first, the Friedman test is conducted on these models' outcomes. p-values less than a significance level of 5% reveal scenarios in

which there is a statistically meaningful distinction in the outputs of the three types of models for a specific evaluation metric. To elicit the performance difference for these cases, critical difference analysis integrated with the post hoc Nemenyi test is used. The results of the critical difference analysis are shown in Figure 2. These diagrams show the average ranking of the modelling approaches in generating superior outcomes for a given evaluation metric. In each figure, models with statistically different average rankings are linked via a thick horizontal line. From Figure 2, the nested stacking models yielded superior evaluation outcomes overall. These findings substantiate the effectiveness of the propositions in addressing the challenge of lag optimisation while conducting enhanced outcomes.

It is noteworthy that, according to the highlighted models in the Appendix A tables, an inconsistency in the efficient lag to be investigated for different patients, prediction horizons, and learners can be observed. In detail, the optimal lag is 30 min in 19 cases, 60 min in 19 cases, 90 min in 5 cases, and 120 min in 5 cases. Such disparity further accentuates the utility of the nested stacking analyses that efficaciously circumvent the lag optimisation process.

Table 2. The evaluation results for the best non-stacking models created using Ohio datasets.

Dataset	PID	Learner	PH	Evaluation Metric						
				RMSE ± SD (mg/dL)	MAE ± SD (mg/dL)	MAPE ± SD (%)	r2 ± SD (%)	MCC ± SD (%)	SE < 0.5 ± SD (%)	ASE ± SD
2018	559	MLP	30	19.65 ± 0.06	13.56 ± 0.03	8.78 ± 0.03	90.75 ± 0.05	0.77 ± 0.00	0.90 ± 0.00	0.19 ± 0.00
			60	31.36 ± 0.06	22.78 ± 0.06	15.18 ± 0.07	76.30 ± 0.08	0.63 ± 0.00	0.79 ± 0.00	0.31 ± 0.00
		LSTM	30	23.12 ± 0.43	16.60 ± 0.66	11.10 ± 0.63	87.19 ± 0.47	0.74 ± 0.01	0.86 ± 0.01	0.24 ± 0.01
			60	36.08 ± 1.47	25.38 ± 0.84	16.62 ± 0.25	68.60 ± 2.56	0.59 ± 0.02	0.75 ± 0.01	0.34 ± 0.01
	563	MLP	30	18.71 ± 0.05	13.46 ± 0.06	8.47 ± 0.04	82.97 ± 0.09	0.74 ± 0.00	0.91 ± 0.00	0.19 ± 0.00
			60	30.65 ± 0.01	21.69 ± 0.04	13.46 ± 0.04	54.36 ± 0.04	0.57 ± 0.01	0.81 ± 0.00	0.30 ± 0.00
		LSTM	30	21.59 ± 0.64	15.33 ± 0.45	9.69 ± 0.19	77.31 ± 1.34	0.72 ± 0.01	0.89 ± 0.00	0.22 ± 0.00
			60	33.02 ± 0.62	24.13 ± 0.61	15.07 ± 0.18	47.03 ± 2.01	0.51 ± 0.01	0.75 ± 0.02	0.33 ± 0.01
	570	MLP	30	17.44 ± 0.03	12.47 ± 0.03	6.38 ± 0.03	93.34 ± 0.03	0.86 ± 0.00	0.96 ± 0.00	0.12 ± 0.00
			60	29.00 ± 0.14	20.97 ± 0.13	10.73 ± 0.04	81.62 ± 0.18	0.79 ± 0.00	0.91 ± 0.00	0.20 ± 0.00
		LSTM	30	22.92 ± 1.49	16.16 ± 1.15	8.04 ± 0.65	88.47 ± 1.52	0.81 ± 0.02	0.94 ± 0.01	0.15 ± 0.01
			60	35.80 ± 1.50	26.75 ± 1.85	12.68 ± 0.43	71.95 ± 2.31	0.75 ± 0.00	0.88 ± 0.01	0.23 ± 0.01
	575	MLP	30	24.12 ± 0.06	16.05 ± 0.10	11.43 ± 0.09	84.48 ± 0.07	0.73 ± 0.00	0.86 ± 0.00	0.24 ± 0.00
			60	35.63 ± 0.17	25.66 ± 0.20	18.91 ± 0.17	66.19 ± 0.32	0.57 ± 0.01	0.71 ± 0.00	0.38 ± 0.00
		LSTM	30	27.20 ± 0.57	18.25 ± 0.45	13.14 ± 0.71	80.24 ± 0.82	0.69 ± 0.00	0.82 ± 0.02	0.28 ± 0.01
			60	38.09 ± 0.03	27.47 ± 0.52	20.48 ± 1.20	61.36 ± 0.07	0.54 ± 0.02	0.70 ± 0.00	0.41 ± 0.01
	588	MLP	30	18.07 ± 0.35	13.50 ± 0.15	8.29 ± 0.01	85.66 ± 0.56	0.76 ± 0.01	0.93 ± 0.00	0.18 ± 0.00
			60	30.36 ± 0.11	22.68 ± 0.13	14.16 ± 0.12	59.60 ± 0.28	0.58 ± 0.00	0.77 ± 0.00	0.31 ± 0.00
		LSTM	30	19.23 ± 0.11	14.16 ± 0.11	8.53 ± 0.12	83.77 ± 0.19	0.74 ± 0.00	0.92 ± 0.00	0.19 ± 0.00
			60	30.46 ± 0.60	22.48 ± 0.39	14.04 ± 0.23	59.33 ± 1.61	0.60 ± 0.01	0.79 ± 0.01	0.30 ± 0.01
	591	MLP	30	22.98 ± 0.11	16.61 ± 0.05	12.99 ± 0.03	80.32 ± 0.18	0.65 ± 0.01	0.80 ± 0.00	0.29 ± 0.00
			60	34.98 ± 0.05	26.93 ± 0.08	21.91 ± 0.13	54.41 ± 0.12	0.39 ± 0.00	0.65 ± 0.00	0.45 ± 0.00
		LSTM	30	26.33 ± 0.42	19.55 ± 0.24	15.65 ± 0.40	74.16 ± 0.83	0.60 ± 0.00	0.75 ± 0.01	0.34 ± 0.01
			60	36.51 ± 0.20	28.36 ± 0.26	23.32 ± 0.27	50.32 ± 0.54	0.37 ± 0.02	0.63 ± 0.00	0.47 ± 0.00
2020	540	MLP	30	22.88 ± 0.13	17.45 ± 0.10	12.71 ± 0.04	87.60 ± 0.14	0.68 ± 0.00	0.81 ± 0.00	0.27 ± 0.00
			60	39.84 ± 0.14	30.49 ± 0.12	22.96 ± 0.13	62.48 ± 0.27	0.52 ± 0.00	0.66 ± 0.00	0.44 ± 0.00
		LSTM	30	24.84 ± 0.42	18.48 ± 0.70	13.81 ± 1.24	85.37 ± 0.49	0.67 ± 0.02	0.80 ± 0.01	0.29 ± 0.02
			60	41.36 ± 0.58	30.69 ± 0.37	22.40 ± 0.20	59.56 ± 1.12	0.50 ± 0.02	0.66 ± 0.00	0.44 ± 0.00
	544	MLP	30	17.37 ± 0.03	12.14 ± 0.03	8.21 ± 0.03	88.26 ± 0.04	0.78 ± 0.00	0.92 ± 0.00	0.18 ± 0.00
			60	28.49 ± 0.03	20.74 ± 0.04	14.16 ± 0.05	68.32 ± 0.07	0.63 ± 0.00	0.78 ± 0.00	0.30 ± 0.00
		LSTM	30	21.23 ± 0.53	15.00 ± 0.49	9.93 ± 0.35	82.45 ± 0.87	0.76 ± 0.01	0.89 ± 0.00	0.21 ± 0.01
			60	30.45 ± 0.12	22.09 ± 0.45	14.81 ± 0.52	63.83 ± 0.29	0.59 ± 0.02	0.78 ± 0.01	0.31 ± 0.01
	552	MLP	30	14.06 ± 0.03	8.25 ± 0.11	6.48 ± 0.09	86.18 ± 0.05	0.75 ± 0.00	0.92 ± 0.00	0.14 ± 0.00
			60	23.83 ± 0.03	14.57 ± 0.10	11.75 ± 0.12	60.36 ± 0.09	0.64 ± 0.00	0.84 ± 0.00	0.22 ± 0.00
		LSTM	30	16.72 ± 0.44	10.31 ± 0.24	8.04 ± 0.22	80.45 ± 1.01	0.71 ± 0.02	0.90 ± 0.01	0.16 ± 0.01
			60	25.47 ± 0.30	16.27 ± 0.24	13.02 ± 0.27	54.73 ± 1.05	0.61 ± 0.01	0.83 ± 0.01	0.24 ± 0.01
	567	MLP	30	22.72 ± 0.04	16.47 ± 0.04	12.48 ± 0.03	84.80 ± 0.05	0.64 ± 0.00	0.80 ± 0.00	0.28 ± 0.00
			60	38.38 ± 0.02	29.51 ± 0.04	23.24 ± 0.06	56.68 ± 0.04	0.46 ± 0.00	0.64 ± 0.00	0.47 ± 0.00
		LSTM	30	24.64 ± 0.97	17.85 ± 0.81	13.48 ± 0.66	82.10 ± 1.41	0.60 ± 0.01	0.78 ± 0.01	0.31 ± 0.01
			60	40.13 ± 1.22	30.57 ± 1.14	25.05 ± 1.96	52.61 ± 2.86	0.45 ± 0.01	0.62 ± 0.02	0.50 ± 0.03
	584	MLP	30	22.78 ± 0.04	16.92 ± 0.04	11.34 ± 0.03	85.49 ± 0.05	0.77 ± 0.00	0.87 ± 0.00	0.23 ± 0.00
			60	35.99 ± 0.05	27.29 ± 0.02	18.40 ± 0.03	63.67 ± 0.11	0.60 ± 0.00	0.72 ± 0.00	0.37 ± 0.00
		LSTM	30	25.31 ± 1.32	18.27 ± 0.95	11.49 ± 0.52	82.05 ± 1.89	0.75 ± 0.01	0.86 ± 0.01	0.23 ± 0.01
			60	41.45 ± 1.58	31.50 ± 1.91	21.43 ± 2.17	51.75 ± 3.64	0.55 ± 0.03	0.67 ± 0.04	0.42 ± 0.04
	596	MLP	30	17.87 ± 0.08	12.89 ± 0.06	9.67 ± 0.03	86.99 ± 0.12	0.74 ± 0.00	0.89 ± 0.00	0.20 ± 0.00
			60	35.99 ± 0.05	27.29 ± 0.02	18.40 ± 0.03	63.67 ± 0.11	0.60 ± 0.00	0.72 ± 0.00	0.37 ± 0.00
		LSTM	30	19.96 ± 0.28	14.31 ± 0.03	10.83 ± 0.18	83.78 ± 0.45	0.70 ± 0.01	0.87 ± 0.00	0.23 ± 0.00
			60	30.28 ± 0.72	22.17 ± 0.71	16.97 ± 0.45	62.72 ± 1.77	0.56 ± 0.02	0.79 ± 0.00	0.32 ± 0.01

Note. PID: patient identification; PH: prediction horizon; LL: lag length; RMSE: root mean square error; SD: standard deviation; MAE: mean absolute error; MAPE: mean absolute percentage error; r^2: coefficient of determination; MCC: Matthew's correlation coefficient; SE: surveillance error; ASE: average surveillance error.

Table 3. The evaluation results for the stacking models created using Ohio datasets.

Dataset	PID	Learner	PH	Evaluation Metric						
				RMSE ± SD (mg/dL)	MAE ± SD (mg/dL)	MAPE ± SD (%)	r^2 ± SD (%)	MCC ± SD (%)	SE < 0.5 ± SD (%)	ASE ± SD
2018	559	MLP	30	19.00 ± 0.11	13.19 ± 0.08	8.79 ± 0.05	91.35 ± 0.10	0.78 ± 0.00	0.90 ± 0.00	0.19 ± 0.00
			60	31.25 ± 0.41	22.67 ± 0.22	15.22 ± 0.24	76.46 ± 0.61	0.64 ± 0.00	0.79 ± 0.00	0.31 ± 0.00
		LSTM	30	22.90 ± 0.49	15.77 ± 0.17	9.97 ± 0.09	87.43 ± 0.54	0.76 ± 0.01	0.89 ± 0.00	0.21 ± 0.00
			60	34.95 ± 0.17	24.99 ± 0.11	16.61 ± 0.05	70.56 ± 0.29	0.61 ± 0.01	0.76 ± 0.00	0.33 ± 0.00
	563	MLP	30	18.54 ± 0.05	13.03 ± 0.03	8.10 ± 0.00	83.28 ± 0.08	0.74 ± 0.01	0.92 ± 0.00	0.18 ± 0.00
			60	29.87 ± 0.18	21.22 ± 0.14	13.36 ± 0.04	56.67 ± 0.51	0.58 ± 0.01	0.81 ± 0.00	0.30 ± 0.00
		LSTM	30	21.25 ± 0.05	14.97 ± 0.06	9.38 ± 0.02	78.05 ± 0.11	0.73 ± 0.00	0.89 ± 0.00	0.21 ± 0.00
			60	33.20 ± 0.16	23.55 ± 0.07	14.44 ± 0.02	46.46 ± 0.53	0.52 ± 0.00	0.78 ± 0.00	0.32 ± 0.00
	570	MLP	30	17.49 ± 0.11	12.43 ± 0.10	6.36 ± 0.03	93.30 ± 0.09	0.86 ± 0.01	0.96 ± 0.00	0.12 ± 0.00
			60	28.65 ± 0.08	20.90 ± 0.07	10.91 ± 0.04	82.06 ± 0.10	0.78 ± 0.00	0.91 ± 0.00	0.20 ± 0.00
		LSTM	30	21.58 ± 1.50	15.59 ± 1.55	7.70 ± 0.49	89.77 ± 1.44	0.84 ± 0.01	0.94 ± 0.00	0.14 ± 0.01
			60	32.48 ± 0.69	23.55 ± 0.62	11.82 ± 0.06	76.93 ± 0.98	0.76 ± 0.00	0.89 ± 0.00	0.22 ± 0.00
	575	MLP	30	24.21 ± 0.04	15.70 ± 0.09	11.25 ± 0.19	84.36 ± 0.05	0.74 ± 0.00	0.86 ± 0.00	0.24 ± 0.00
			60	36.42 ± 0.41	26.35 ± 0.77	19.85 ± 1.57	64.68 ± 0.79	0.57 ± 0.02	0.71 ± 0.00	0.40 ± 0.02
		LSTM	30	27.73 ± 0.12	18.09 ± 0.09	12.67 ± 0.09	79.48 ± 0.18	0.66 ± 0.00	0.82 ± 0.00	0.27 ± 0.00
			60	38.34 ± 0.09	27.48 ± 0.06	19.59 ± 0.12	60.86 ± 0.18	0.54 ± 0.00	0.68 ± 0.00	0.41 ± 0.00
	588	MLP	30	18.24 ± 0.19	13.51 ± 0.12	8.17 ± 0.02	85.39 ± 0.30	0.75 ± 0.01	0.93 ± 0.00	0.18 ± 0.00
			60	29.65 ± 0.21	21.84 ± 0.18	13.14 ± 0.08	61.46 ± 0.55	0.57 ± 0.01	0.80 ± 0.00	0.29 ± 0.00
		LSTM	30	18.91 ± 0.08	14.03 ± 0.14	8.43 ± 0.25	84.30 ± 0.13	0.75 ± 0.00	0.92 ± 0.00	0.18 ± 0.01
			60	30.67 ± 0.20	22.29 ± 0.25	13.54 ± 0.49	58.76 ± 0.54	0.60 ± 0.01	0.81 ± 0.01	0.29 ± 0.01
	591	MLP	30	22.88 ± 0.07	16.60 ± 0.04	13.03 ± 0.06	80.49 ± 0.12	0.65 ± 0.00	0.80 ± 0.00	0.29 ± 0.00
			60	34.43 ± 0.06	26.80 ± 0.05	22.09 ± 0.09	55.84 ± 0.14	0.41 ± 0.00	0.65 ± 0.00	0.45 ± 0.00
		LSTM	30	25.51 ± 0.01	18.80 ± 0.05	14.79 ± 0.08	75.73 ± 0.03	0.59 ± 0.00	0.76 ± 0.00	0.33 ± 0.00
			60	36.68 ± 0.16	28.44 ± 0.05	23.78 ± 0.03	49.87 ± 0.44	0.42 ± 0.00	0.64 ± 0.00	0.47 ± 0.00
2020	540	MLP	30	22.34 ± 0.02	17.13 ± 0.03	12.58 ± 0.03	88.18 ± 0.02	0.68 ± 0.00	0.82 ± 0.00	0.27 ± 0.00
			60	39.40 ± 0.09	30.32 ± 0.13	22.95 ± 0.10	63.29 ± 0.17	0.52 ± 0.00	0.66 ± 0.00	0.44 ± 0.00
		LSTM	30	24.13 ± 0.14	18.24 ± 0.06	13.57 ± 0.03	86.20 ± 0.17	0.66 ± 0.00	0.80 ± 0.00	0.29 ± 0.00
			60	40.86 ± 0.05	30.62 ± 0.11	23.06 ± 0.18	60.53 ± 0.09	0.51 ± 0.00	0.66 ± 0.00	0.44 ± 0.00
	544	MLP	30	16.96 ± 0.02	12.01 ± 0.05	8.14 ± 0.08	88.81 ± 0.03	0.79 ± 0.00	0.92 ± 0.00	0.18 ± 0.00
			60	28.36 ± 0.17	20.72 ± 0.04	14.21 ± 0.08	68.62 ± 0.37	0.64 ± 0.00	0.78 ± 0.00	0.30 ± 0.00
		LSTM	30	20.85 ± 0.25	14.84 ± 0.20	10.01 ± 0.14	83.08 ± 0.40	0.73 ± 0.00	0.88 ± 0.00	0.22 ± 0.00
			60	31.30 ± 0.23	22.55 ± 0.10	15.44 ± 0.07	61.77 ± 0.57	0.59 ± 0.00	0.76 ± 0.00	0.33 ± 0.00
	552	MLP	30	14.19 ± 0.03	9.00 ± 0.06	7.10 ± 0.03	85.92 ± 0.05	0.72 ± 0.00	0.91 ± 0.00	0.15 ± 0.00
			60	23.78 ± 0.04	15.52 ± 0.20	12.62 ± 0.18	60.53 ± 0.14	0.61 ± 0.01	0.84 ± 0.00	0.23 ± 0.00
		LSTM	30	17.65 ± 0.22	11.92 ± 0.20	9.79 ± 0.21	78.23 ± 0.53	0.69 ± 0.00	0.88 ± 0.01	0.19 ± 0.01
			60	26.93 ± 0.23	17.97 ± 0.17	15.04 ± 0.14	49.39 ± 0.85	0.58 ± 0.01	0.78 ± 0.00	0.28 ± 0.00
	567	MLP	30	22.67 ± 0.22	16.17 ± 0.22	12.39 ± 0.21	84.86 ± 0.29	0.64 ± 0.01	0.81 ± 0.00	0.28 ± 0.00
			60	37.82 ± 0.24	28.14 ± 0.18	22.42 ± 0.23	57.94 ± 0.52	0.48 ± 0.00	0.66 ± 0.00	0.46 ± 0.00
		LSTM	30	23.74 ± 0.09	16.86 ± 0.14	12.96 ± 0.14	83.41 ± 0.13	0.62 ± 0.00	0.79 ± 0.00	0.30 ± 0.00
			60	38.75 ± 0.41	29.24 ± 0.31	23.40 ± 0.46	55.84 ± 0.92	0.47 ± 0.01	0.64 ± 0.01	0.48 ± 0.01
	584	MLP	30	21.89 ± 0.09	15.96 ± 0.14	10.64 ± 0.13	86.60 ± 0.11	0.77 ± 0.00	0.89 ± 0.00	0.22 ± 0.00
			60	35.42 ± 0.42	26.73 ± 0.52	17.97 ± 0.53	64.79 ± 0.83	0.60 ± 0.01	0.73 ± 0.01	0.36 ± 0.01
		LSTM	30	24.79 ± 0.06	18.21 ± 0.08	12.51 ± 0.13	82.82 ± 0.08	0.76 ± 0.00	0.86 ± 0.00	0.25 ± 0.00
			60	38.65 ± 0.29	29.33 ± 0.12	20.14 ± 0.01	58.09 ± 0.63	0.60 ± 0.00	0.70 ± 0.00	0.39 ± 0.00
	596	MLP	30	17.76 ± 0.09	12.85 ± 0.09	9.71 ± 0.11	87.16 ± 0.13	0.75 ± 0.00	0.90 ± 0.00	0.20 ± 0.00
			60	28.80 ± 0.19	21.37 ± 0.13	16.53 ± 0.11	66.29 ± 0.44	0.59 ± 0.01	0.80 ± 0.00	0.31 ± 0.00
		LSTM	30	19.06 ± 0.16	13.55 ± 0.08	10.27 ± 0.06	85.21 ± 0.24	0.72 ± 0.00	0.88 ± 0.00	0.22 ± 0.00
			60	30.01 ± 0.10	22.25 ± 0.10	17.31 ± 0.16	63.39 ± 0.25	0.56 ± 0.00	0.80 ± 0.00	0.32 ± 0.00

Note. PID: patient identification; PH: prediction horizon; LL: lag length; RMSE: root mean square error; SD: standard deviation; MAE: mean absolute error; MAPE: mean absolute percentage error; r^2: coefficient of determination; MCC: Matthew's correlation coefficient; SE: surveillance error; ASE: average surveillance error.

Table 4. The evaluation results for the nested stacking models created using Ohio datasets.

Dataset	PID	Learner	PH	Evaluation Metric						
				RMSE ± SD (mg/dL)	MAE ± SD (mg/dL)	MAPE ± SD (%)	r^2 ± SD (%)	MCC ± SD (%)	SE < 0.5 ± SD (%)	ASE ± SD
2018	559	MLP	30	19.67 ± 0.05	13.54 ± 0.05	8.89 ± 0.03	90.72 ± 0.05	0.79 ± 0.00	0.90 ± 0.00	0.19 ± 0.00
			60	33.44 ± 0.28	23.54 ± 0.16	15.27 ± 0.04	73.05 ± 0.46	0.63 ± 0.00	0.78 ± 0.00	0.31 ± 0.00
		LSTM	30	19.69 ± 0.19	13.51 ± 0.18	8.83 ± 0.17	90.71 ± 0.18	0.79 ± 0.00	0.90 ± 0.00	0.19 ± 0.00
			60	33.93 ± 0.48	23.82 ± 0.28	15.31 ± 0.05	72.25 ± 0.79	0.63 ± 0.01	0.78 ± 0.00	0.31 ± 0.00
	563	MLP	30	18.85 ± 0.10	13.15 ± 0.08	8.27 ± 0.02	82.72 ± 0.19	0.76 ± 0.01	0.91 ± 0.00	0.18 ± 0.00
			60	31.82 ± 0.54	22.38 ± 0.38	13.84 ± 0.11	50.81 ± 1.66	0.55 ± 0.01	0.80 ± 0.01	0.30 ± 0.00
		LSTM	30	19.00 ± 0.07	13.24 ± 0.06	8.31 ± 0.03	82.44 ± 0.13	0.76 ± 0.01	0.91 ± 0.00	0.19 ± 0.00
			60	31.65 ± 0.51	22.37 ± 0.61	13.79 ± 0.10	51.35 ± 1.59	0.55 ± 0.03	0.80 ± 0.01	0.31 ± 0.01
	570	MLP	30	18.34 ± 0.11	12.85 ± 0.08	6.58 ± 0.05	92.64 ± 0.09	0.86 ± 0.00	0.96 ± 0.00	0.12 ± 0.00
			60	31.09 ± 0.28	22.21 ± 0.14	11.54 ± 0.03	78.88 ± 0.38	0.77 ± 0.00	0.89 ± 0.00	0.21 ± 0.00
		LSTM	30	18.57 ± 0.22	13.11 ± 0.12	6.65 ± 0.08	92.45 ± 0.18	0.86 ± 0.00	0.96 ± 0.00	0.12 ± 0.00
			60	31.61 ± 0.60	22.60 ± 0.54	11.53 ± 0.02	78.16 ± 0.84	0.77 ± 0.00	0.90 ± 0.00	0.21 ± 0.00
	575	MLP	30	26.18 ± 0.09	16.60 ± 0.19	12.40 ± 0.27	81.71 ± 0.12	0.73 ± 0.00	0.84 ± 0.00	0.26 ± 0.01
			60	36.98 ± 0.33	26.43 ± 0.50	19.46 ± 1.39	63.57 ± 0.65	0.54 ± 0.01	0.70 ± 0.01	0.40 ± 0.02
		LSTM	30	26.01 ± 0.91	16.47 ± 0.32	12.02 ± 0.66	81.93 ± 1.25	0.73 ± 0.00	0.84 ± 0.01	0.25 ± 0.01
			60	37.05 ± 0.62	26.29 ± 0.28	18.96 ± 0.13	63.44 ± 1.22	0.54 ± 0.00	0.70 ± 0.00	0.39 ± 0.00
	588	MLP	30	18.50 ± 0.11	13.63 ± 0.08	8.11 ± 0.05	84.98 ± 0.17	0.74 ± 0.00	0.93 ± 0.00	0.18 ± 0.00
			60	29.43 ± 0.07	21.42 ± 0.17	13.01 ± 0.42	62.05 ± 0.17	0.62 ± 0.00	0.82 ± 0.01	0.28 ± 0.01
		LSTM	30	18.26 ± 0.14	13.56 ± 0.27	8.23 ± 0.32	85.37 ± 0.22	0.76 ± 0.01	0.93 ± 0.00	0.18 ± 0.01
			60	29.54 ± 0.28	21.33 ± 0.21	12.84 ± 0.09	61.77 ± 0.74	0.62 ± 0.01	0.82 ± 0.00	0.27 ± 0.00
	591	MLP	30	23.07 ± 0.09	16.48 ± 0.04	12.89 ± 0.06	80.16 ± 0.15	0.64 ± 0.01	0.80 ± 0.00	0.29 ± 0.00
			60	35.68 ± 0.11	27.65 ± 0.08	23.12 ± 0.07	52.56 ± 0.29	0.42 ± 0.00	0.65 ± 0.00	0.46 ± 0.00
		LSTM	30	23.08 ± 0.10	16.52 ± 0.07	12.98 ± 0.08	80.14 ± 0.17	0.63 ± 0.00	0.80 ± 0.00	0.29 ± 0.00
			60	35.68 ± 0.21	27.69 ± 0.12	23.16 ± 0.08	52.57 ± 0.55	0.42 ± 0.00	0.65 ± 0.01	0.46 ± 0.00
2020	540	MLP	30	22.36 ± 0.03	16.96 ± 0.05	12.59 ± 0.03	88.15 ± 0.03	0.67 ± 0.00	0.82 ± 0.00	0.27 ± 0.00
			60	38.81 ± 0.26	29.34 ± 0.14	22.04 ± 0.10	64.38 ± 0.47	0.53 ± 0.01	0.68 ± 0.00	0.43 ± 0.00
		LSTM	30	22.39 ± 0.11	16.99 ± 0.09	12.61 ± 0.08	88.12 ± 0.12	0.67 ± 0.01	0.81 ± 0.00	0.27 ± 0.00
			60	38.74 ± 0.18	29.32 ± 0.18	22.05 ± 0.15	64.52 ± 0.33	0.53 ± 0.01	0.68 ± 0.00	0.43 ± 0.00
	544	MLP	30	16.86 ± 0.11	11.89 ± 0.06	8.02 ± 0.06	88.94 ± 0.14	0.78 ± 0.00	0.92 ± 0.00	0.17 ± 0.00
			60	28.92 ± 0.14	20.88 ± 0.05	14.33 ± 0.02	67.36 ± 0.31	0.63 ± 0.00	0.77 ± 0.00	0.30 ± 0.00
		LSTM	30	16.96 ± 0.15	11.95 ± 0.11	8.07 ± 0.09	88.80 ± 0.19	0.78 ± 0.01	0.92 ± 0.00	0.18 ± 0.00
			60	28.84 ± 0.19	20.81 ± 0.10	14.34 ± 0.13	67.54 ± 0.42	0.63 ± 0.00	0.77 ± 0.00	0.30 ± 0.00
	552	MLP	30	13.87 ± 0.16	8.88 ± 0.32	7.07 ± 0.24	86.56 ± 0.32	0.72 ± 0.01	0.92 ± 0.00	0.15 ± 0.01
			60	24.61 ± 0.11	16.04 ± 0.36	13.43 ± 0.30	57.73 ± 0.38	0.60 ± 0.00	0.82 ± 0.00	0.25 ± 0.00
		LSTM	30	13.86 ± 0.02	9.00 ± 0.06	7.13 ± 0.06	86.58 ± 0.03	0.72 ± 0.00	0.92 ± 0.00	0.15 ± 0.00
			60	23.97 ± 0.44	15.47 ± 0.32	12.76 ± 0.38	59.91 ± 1.47	0.61 ± 0.00	0.83 ± 0.01	0.24 ± 0.01
	567	MLP	30	21.81 ± 0.28	15.58 ± 0.14	11.71 ± 0.30	86.00 ± 0.35	0.65 ± 0.01	0.82 ± 0.01	0.27 ± 0.01
			60	37.50 ± 0.18	27.95 ± 0.13	21.97 ± 0.18	58.65 ± 0.39	0.49 ± 0.00	0.66 ± 0.00	0.46 ± 0.00
		LSTM	30	22.02 ± 0.07	15.70 ± 0.05	11.96 ± 0.07	85.72 ± 0.08	0.64 ± 0.00	0.82 ± 0.00	0.27 ± 0.00
			60	37.77 ± 0.25	28.19 ± 0.22	22.38 ± 0.36	58.05 ± 0.55	0.48 ± 0.00	0.66 ± 0.00	0.46 ± 0.00
	584	MLP	30	22.35 ± 0.58	16.74 ± 0.67	11.54 ± 0.54	86.03 ± 0.73	0.77 ± 0.01	0.88 ± 0.01	0.24 ± 0.01
			60	35.77 ± 0.49	27.25 ± 0.49	18.79 ± 0.44	64.11 ± 0.99	0.61 ± 0.01	0.73 ± 0.01	0.37 ± 0.01
		LSTM	30	22.19 ± 0.11	16.54 ± 0.17	11.38 ± 0.17	86.24 ± 0.13	0.77 ± 0.00	0.88 ± 0.00	0.23 ± 0.00
			60	36.02 ± 0.06	27.37 ± 0.12	18.91 ± 0.14	63.60 ± 0.12	0.61 ± 0.00	0.72 ± 0.00	0.37 ± 0.00
	596	MLP	30	17.78 ± 0.24	12.67 ± 0.13	9.52 ± 0.10	87.13 ± 0.35	0.74 ± 0.00	0.89 ± 0.00	0.20 ± 0.00
			60	28.54 ± 0.24	20.79 ± 0.09	15.74 ± 0.27	66.89 ± 0.55	0.58 ± 0.02	0.81 ± 0.00	0.30 ± 0.00
		LSTM	30	17.57 ± 0.25	12.49 ± 0.14	9.35 ± 0.09	87.43 ± 0.36	0.75 ± 0.01	0.89 ± 0.00	0.20 ± 0.00
			60	28.68 ± 0.37	20.97 ± 0.07	15.96 ± 0.31	66.55 ± 0.87	0.58 ± 0.02	0.81 ± 0.00	0.31 ± 0.00

Note. PID: patient identification; PH: prediction horizon; LL: lag length; RMSE: root mean square error; SD: standard deviation; MAE: mean absolute error; MAPE: mean absolute percentage error; r^2: coefficient of determination; MCC: Matthew's correlation coefficient; SE: surveillance error; ASE: average surveillance error.

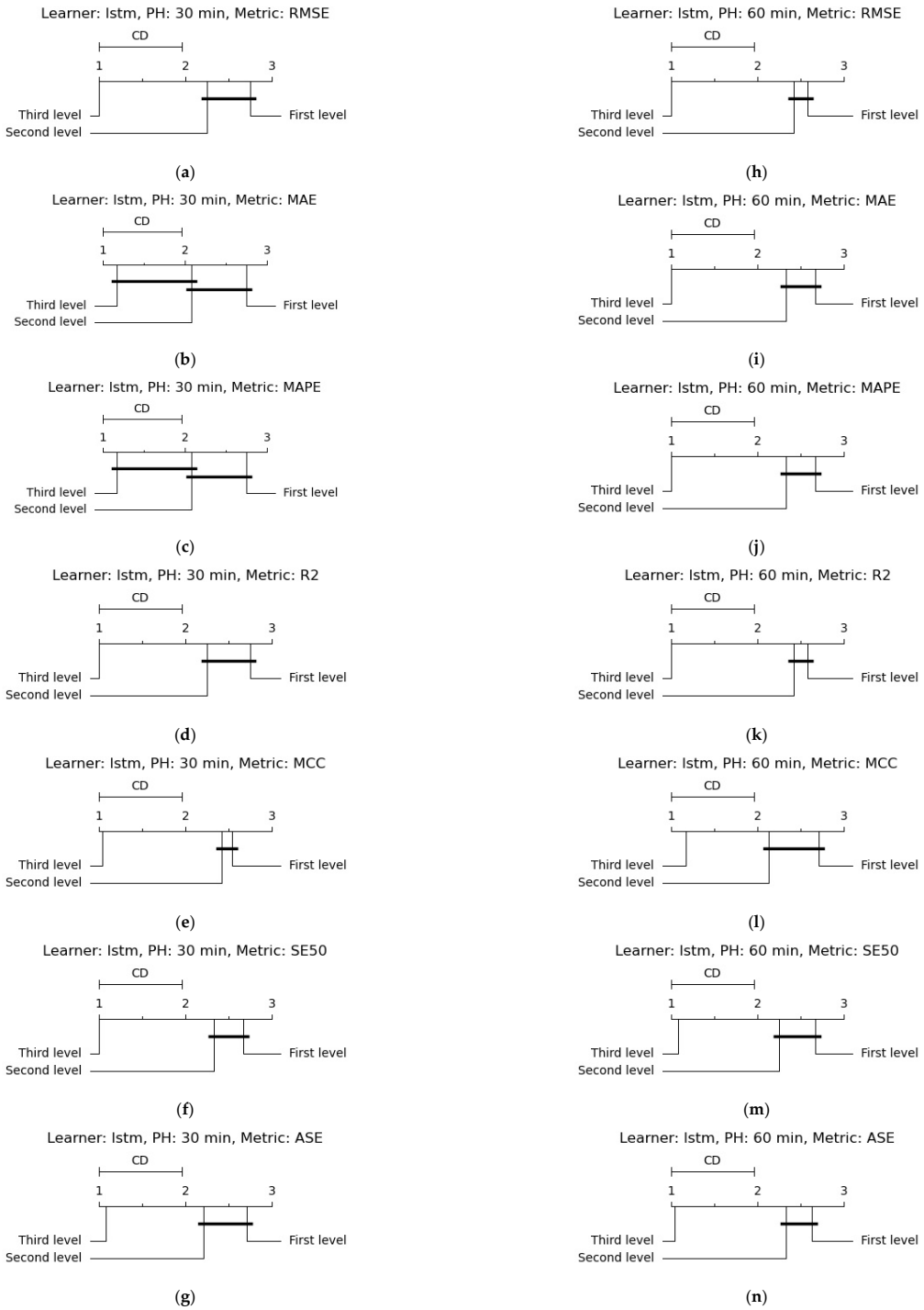

Figure 2. Critical difference diagrams based on Nemenyi test for pairwise comparison of the non-stacking, stacking, and nested stacking modelling approaches: (a) LSTM learner, 30 min PH, and RMSE metric, (b) LSTM learner, 30 min PH, and MAE metric, (c) LSTM learner, 30 min PH, and MAPE metric, (d) LSTM learner, 30 min PH, and r2 metric, (e) LSTM learner, 30 min PH, and MCC metric, (f) LSTM learner, 30 min PH, and SE50 metric, (g) LSTM learner, 30 min PH, and ASE metric, (h) LSTM learner, 60 min PH, and RMSE metric, (i) LSTM learner, 60 min PH, and MAE metric, (j) LSTM learner, 60 min PH, and MAPE metric, (k) LSTM learner, 60 min PH, and r2 metric, (l) LSTM learner, 60 min PH, and MCC metric, (m) LSTM learner, 60 min PH, and SE50 metric, (n) LSTM learner, 60 min PH, and ASE metric. Note. LSTM: long short-term memory; PH: prediction horizon; RMSE: root mean square error; MAE: mean absolute error; MAPE: mean absolute percentage error; r2: coefficient of determination; MCC: Matthew's correlation coefficient; SE: surveillance error; ASE: average surveillance error.

7. Summary and Conclusions

This work offers a nested meta-learning lag fusion approach to address the challenge of history length optimisation in personalised blood glucose level prediction. For this purpose, in lieu of examining different lengths of history from a search space and picking a local optimum for each subject or a global suboptimum for all subjects, all the lags in the search space are studied autonomously, and the results are amalgamated. A multilayer perceptron and long short-term memory network are initially trained on four different lags separately, resulting in four non-stacking models from each network. The outcomes of the four non-stacking multilayer perceptron models are then combined into new outcomes using a stacking multilayer perceptron model. Similarly, a stacking long short-term memory model fuses the results of the four non-stacking long short-term memory models. Finally, the decisions of the two stacking prediction models are ensembled once using a multilayer perceptron and once using a long short-term memory network as a nested stacking model. These investigations are performed for two commonly studied prediction horizons of 30 and 60 min in blood glucose level prediction research. The generated models undergo in-depth regression-wise, clinical-wise, and statistic-wise assessments. The results obtained substantiate the effectiveness of the proposed stacking and nested stacking methods in addressing the challenge of lag optimisation in blood glucose level prediction analysis.

While this study demonstrates the effectiveness of the proposed nested meta-learning lag fusion approach for blood glucose level prediction, several areas for future improvement should be considered. First, enhancing the interpretability of the model will be crucial for gaining deeper insights into the underlying decision-making process, especially for clinical applications where transparency is essential. Second, further optimisation of the model's computational efficiency could facilitate its real-time deployment in wearable glucose monitoring devices. Additionally, expanding the dataset to include larger and more diverse populations would help validate the generalisability of the approach across different patient demographics and conditions. Finally, testing the model in real-world clinical environments,

including its integration with continuous glucose monitoring systems, would provide a more comprehensive understanding of its practical utility and impact on diabetes management.

Software and Code

For developing and evaluating blood glucose level prediction models, this research used Python 3.6 [72] programming. The libraries and packages employed include TensorFlow [73], Keras [73], Pandas [74], NumPy [75], Sklearn [76], SciPy [77], statsmodels [78], scikit-post hocs [79], and cd-diagram [80]. The source code for implementations is available on this Gitlab repository.

Funding

University of Sheffield Institutional Open Access Fund.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A

In this section, the complete outcomes of evaluation analysis on the non-stacking models are provided in four tables, as below.

Table A1. *The evaluation results for non-stacking models created by multilayer perceptron learners using Ohio 2018 dataset.*

PID	PH	LL	Evaluation metric						
			RMSE ± SD (mg/dL)	MAE ± SD (mg/dL)	MAPE ± SD (%)	r^2 ± SD (%)	MCC ± SD (%)	SE < 0.5 ± SD (%)	ASE ± SD
559	30	30	19.96 ± 0.09	13.78 ± 0.11	8.83 ± 0.11	90.45 ± 0.08	0.77 ± 0.00	0.90 ± 0.00	0.19 ± 0.00
		60	19.65 ± 0.06	13.56 ± 0.03	8.78 ± 0.03	90.75 ± 0.05	0.77 ± 0.00	0.90 ± 0.00	0.19 ± 0.00
		90	19.85 ± 0.01	13.73 ± 0.02	8.81 ± 0.04	90.56 ± 0.01	0.77 ± 0.00	0.90 ± 0.00	0.19 ± 0.00
		120	19.88 ± 0.07	13.83 ± 0.05	8.81 ± 0.04	90.53 ± 0.07	0.77 ± 0.00	0.90 ± 0.00	0.19 ± 0.00
	60	30	33.73 ± 0.04	24.46 ± 0.04	16.49 ± 0.05	72.59 ± 0.06	0.58 ± 0.00	0.77 ± 0.00	0.33 ± 0.00
		60	32.04 ± 0.05	23.12 ± 0.09	15.43 ± 0.11	75.26 ± 0.08	0.62 ± 0.01	0.79 ± 0.00	0.31 ± 0.00
		90	31.67 ± 0.05	22.84 ± 0.06	15.23 ± 0.04	75.82 ± 0.08	0.64 ± 0.00	0.79 ± 0.00	0.31 ± 0.00
		120	31.36 ± 0.06	22.78 ± 0.06	15.18 ± 0.07	76.30 ± 0.08	0.63 ± 0.00	0.79 ± 0.00	0.31 ± 0.00
563	30	30	18.71 ± 0.05	13.46 ± 0.06	8.47 ± 0.04	82.97 ± 0.09	0.74 ± 0.00	0.91 ± 0.00	0.19 ± 0.00
		60	18.89 ± 0.03	13.33 ± 0.03	8.30 ± 0.02	82.65 ± 0.05	0.74 ± 0.00	0.91 ± 0.00	0.19 ± 0.00
		90	19.09 ± 0.03	13.42 ± 0.03	8.34 ± 0.02	82.27 ± 0.06	0.74 ± 0.01	0.91 ± 0.00	0.19 ± 0.00
		120	19.29 ± 0.01	13.61 ± 0.00	8.45 ± 0.00	81.91 ± 0.02	0.73 ± 0.01	0.91 ± 0.00	0.19 ± 0.00
	60	30	30.44 ± 0.08	22.46 ± 0.08	14.40 ± 0.06	55.00 ± 0.23	0.49 ± 0.00	0.78 ± 0.00	0.33 ± 0.00
		60	30.43 ± 0.05	21.75 ± 0.02	13.57 ± 0.02	55.02 ± 0.14	0.56 ± 0.01	0.80 ± 0.00	0.30 ± 0.00
		90	30.65 ± 0.01	21.69 ± 0.04	13.46 ± 0.04	54.36 ± 0.04	0.57 ± 0.01	0.81 ± 0.00	0.30 ± 0.00
		120	30.68 ± 0.15	21.72 ± 0.09	13.47 ± 0.05	54.28 ± 0.44	0.57 ± 0.00	0.81 ± 0.00	0.30 ± 0.00
570	30	30	18.24 ± 0.19	13.27 ± 0.15	6.74 ± 0.08	92.71 ± 0.15	0.84 ± 0.00	0.95 ± 0.00	0.13 ± 0.00
		60	17.44 ± 0.03	12.47 ± 0.03	6.38 ± 0.03	93.34 ± 0.03	0.86 ± 0.00	0.96 ± 0.00	0.12 ± 0.00
		90	17.58 ± 0.03	12.54 ± 0.03	6.45 ± 0.01	93.24 ± 0.03	0.86 ± 0.00	0.96 ± 0.00	0.12 ± 0.00
		120	17.71 ± 0.13	12.53 ± 0.11	6.41 ± 0.06	93.13 ± 0.10	0.86 ± 0.00	0.96 ± 0.00	0.12 ± 0.00
	60	30	30.36 ± 0.08	23.08 ± 0.07	11.89 ± 0.03	79.85 ± 0.10	0.74 ± 0.00	0.89 ± 0.00	0.22 ± 0.00
		60	28.89 ± 0.03	21.33 ± 0.04	10.92 ± 0.01	81.76 ± 0.04	0.78 ± 0.00	0.91 ± 0.00	0.20 ± 0.00
		90	28.95 ± 0.10	21.07 ± 0.09	10.82 ± 0.02	81.68 ± 0.13	0.79 ± 0.00	0.91 ± 0.00	0.20 ± 0.00
		120	29.00 ± 0.14	20.97 ± 0.13	10.73 ± 0.04	81.62 ± 0.18	0.79 ± 0.00	0.91 ± 0.00	0.20 ± 0.00
575	30	30	24.12 ± 0.06	16.05 ± 0.10	11.43 ± 0.09	84.48 ± 0.07	0.73 ± 0.00	0.86 ± 0.00	0.24 ± 0.00
		60	24.49 ± 0.04	15.93 ± 0.02	11.39 ± 0.02	84.00 ± 0.06	0.73 ± 0.00	0.85 ± 0.00	0.25 ± 0.00
		90	24.38 ± 0.09	15.97 ± 0.13	11.56 ± 0.11	84.13 ± 0.12	0.74 ± 0.00	0.85 ± 0.00	0.25 ± 0.00
		120	24.35 ± 0.09	16.07 ± 0.12	11.72 ± 0.16	84.17 ± 0.12	0.75 ± 0.00	0.85 ± 0.01	0.25 ± 0.00
	60	30	36.22 ± 0.10	26.77 ± 0.12	19.49 ± 0.10	65.08 ± 0.19	0.51 ± 0.00	0.69 ± 0.00	0.40 ± 0.00
		60	36.27 ± 0.20	26.24 ± 0.25	18.96 ± 0.17	64.96 ± 0.39	0.54 ± 0.01	0.70 ± 0.00	0.39 ± 0.00
		90	35.90 ± 0.23	25.73 ± 0.11	18.79 ± 0.09	65.68 ± 0.44	0.55 ± 0.00	0.70 ± 0.00	0.39 ± 0.00
		120	35.63 ± 0.17	25.66 ± 0.20	18.91 ± 0.17	66.19 ± 0.32	0.57 ± 0.01	0.71 ± 0.00	0.38 ± 0.00
588	30	30	18.80 ± 0.09	13.99 ± 0.09	8.63 ± 0.07	84.49 ± 0.15	0.75 ± 0.00	0.92 ± 0.00	0.19 ± 0.00
		60	18.27 ± 0.42	13.61 ± 0.20	8.36 ± 0.06	85.35 ± 0.68	0.75 ± 0.02	0.93 ± 0.00	0.18 ± 0.00
		90	18.07 ± 0.35	13.50 ± 0.15	8.29 ± 0.01	85.66 ± 0.56	0.76 ± 0.01	0.93 ± 0.00	0.18 ± 0.00
		120	18.44 ± 0.67	13.64 ± 0.37	8.26 ± 0.13	85.06 ± 1.09	0.75 ± 0.02	0.93 ± 0.01	0.18 ± 0.00
	60	30	30.36 ± 0.11	22.68 ± 0.13	14.16 ± 0.12	59.60 ± 0.28	0.58 ± 0.00	0.77 ± 0.00	0.31 ± 0.00
		60	30.72 ± 0.26	22.76 ± 0.25	13.62 ± 0.16	58.65 ± 0.69	0.56 ± 0.01	0.79 ± 0.00	0.30 ± 0.00
		90	30.58 ± 0.05	22.47 ± 0.10	13.41 ± 0.08	59.01 ± 0.13	0.56 ± 0.00	0.80 ± 0.00	0.29 ± 0.00
		120	30.48 ± 0.25	22.39 ± 0.26	13.33 ± 0.19	59.29 ± 0.67	0.57 ± 0.01	0.80 ± 0.00	0.29 ± 0.00
591	30	30	22.89 ± 0.02	16.68 ± 0.02	12.98 ± 0.02	80.47 ± 0.04	0.62 ± 0.00	0.79 ± 0.00	0.29 ± 0.00
		60	22.98 ± 0.11	16.61 ± 0.05	12.99 ± 0.03	80.32 ± 0.18	0.65 ± 0.01	0.80 ± 0.00	0.29 ± 0.00
		90	23.01 ± 0.06	16.71 ± 0.01	13.12 ± 0.02	80.26 ± 0.09	0.64 ± 0.01	0.80 ± 0.00	0.29 ± 0.00
		120	22.97 ± 0.07	16.78 ± 0.05	13.21 ± 0.11	80.32 ± 0.12	0.64 ± 0.01	0.80 ± 0.00	0.29 ± 0.00
	60	30	35.00 ± 0.05	27.27 ± 0.06	22.01 ± 0.07	54.35 ± 0.14	0.36 ± 0.00	0.64 ± 0.00	0.45 ± 0.00
		60	35.93 ± 0.07	27.77 ± 0.02	22.37 ± 0.07	51.89 ± 0.19	0.35 ± 0.00	0.63 ± 0.00	0.46 ± 0.00
		90	34.98 ± 0.05	26.93 ± 0.08	21.91 ± 0.13	54.41 ± 0.12	0.39 ± 0.00	0.65 ± 0.00	0.45 ± 0.00
		120	34.91 ± 0.07	27.12 ± 0.16	22.19 ± 0.25	54.60 ± 0.19	0.39 ± 0.00	0.65 ± 0.00	0.45 ± 0.00

Note. Values in bold indicate the best evaluation outcome for each metric in each learning scenario, and grey highlights denote the best model in each scenario based on the best-achieved evaluation metrics. Note. PID: patient identification; PH: prediction horizon; LL: lag length; RMSE: root mean square error; SD: standard deviation; MAE: mean absolute error; MAPE: mean absolute percentage error; r^2: coefficient of determination; MCC: Matthew's correlation coefficient; SE: surveillance error; ASE: average surveillance error.

Table A2. *The evaluation results for non-stacking models created by multilayer perceptron learners using Ohio 2020 dataset.*

PID	PH	LL	Evaluation Metric						
			RMSE ± SD (mg/dL)	MAE ± SD (mg/dL)	MAPE ± SD (%)	r^2 ± SD (%)	MCC ± SD (%)	SE < 0.5 ± SD (%)	ASE ± SD
540	30	30	23.48 ± 0.04	17.73 ± 0.03	12.88 ± 0.00	86.93 ± 0.04	0.67 ± 0.00	0.81 ± 0.00	0.28 ± 0.00
		60	22.88 ± 0.13	17.45 ± 0.10	12.71 ± 0.04	87.60 ± 0.14	0.68 ± 0.00	0.81 ± 0.00	0.27 ± 0.00
		90	23.41 ± 0.08	17.79 ± 0.04	12.84 ± 0.04	87.02 ± 0.09	0.68 ± 0.00	0.81 ± 0.00	0.28 ± 0.00
		120	23.61 ± 0.13	17.92 ± 0.07	12.86 ± 0.02	86.79 ± 0.15	0.67 ± 0.00	0.81 ± 0.00	0.28 ± 0.00
	60	30	40.74 ± 0.16	31.20 ± 0.15	23.55 ± 0.12	60.76 ± 0.32	0.49 ± 0.00	0.65 ± 0.00	0.45 ± 0.00
		60	39.84 ± 0.14	30.49 ± 0.12	22.96 ± 0.13	62.48 ± 0.27	0.52 ± 0.00	0.66 ± 0.00	0.44 ± 0.00
		90	40.15 ± 0.16	30.68 ± 0.15	23.09 ± 0.14	61.90 ± 0.30	0.52 ± 0.01	0.66 ± 0.00	0.44 ± 0.00
		120	40.38 ± 0.16	30.88 ± 0.14	23.16 ± 0.07	61.45 ± 0.31	0.52 ± 0.00	0.66 ± 0.00	0.44 ± 0.00
544	30	30	17.76 ± 0.06	12.45 ± 0.07	8.47 ± 0.07	87.73 ± 0.09	0.78 ± 0.00	0.91 ± 0.00	0.18 ± 0.00
		60	17.37 ± 0.03	12.14 ± 0.03	8.21 ± 0.03	88.26 ± 0.04	0.78 ± 0.00	0.92 ± 0.00	0.18 ± 0.00
		90	17.61 ± 0.03	12.42 ± 0.04	8.35 ± 0.03	87.94 ± 0.05	0.77 ± 0.00	0.91 ± 0.00	0.18 ± 0.00
		120	17.78 ± 0.10	12.49 ± 0.04	8.39 ± 0.03	87.71 ± 0.13	0.77 ± 0.00	0.91 ± 0.00	0.19 ± 0.00
	60	30	29.25 ± 0.08	21.79 ± 0.08	15.29 ± 0.08	66.61 ± 0.19	0.59 ± 0.00	0.75 ± 0.00	0.32 ± 0.00
		60	28.49 ± 0.03	20.74 ± 0.04	14.16 ± 0.05	68.32 ± 0.07	0.63 ± 0.00	0.78 ± 0.00	0.30 ± 0.00
		90	28.92 ± 0.09	21.03 ± 0.02	14.29 ± 0.04	67.35 ± 0.20	0.63 ± 0.00	0.77 ± 0.00	0.30 ± 0.00
		120	29.14 ± 0.12	21.12 ± 0.09	14.32 ± 0.04	66.86 ± 0.27	0.62 ± 0.00	0.77 ± 0.00	0.31 ± 0.00
552	30	30	14.06 ± 0.03	8.25 ± 0.11	6.48 ± 0.09	86.18 ± 0.05	0.75 ± 0.00	0.92 ± 0.00	0.14 ± 0.00
		60	14.32 ± 0.08	8.91 ± 0.08	7.03 ± 0.06	85.67 ± 0.16	0.73 ± 0.00	0.91 ± 0.00	0.15 ± 0.00
		90	14.47 ± 0.10	9.25 ± 0.09	7.30 ± 0.09	85.36 ± 0.20	0.72 ± 0.00	0.91 ± 0.00	0.15 ± 0.00
		120	14.60 ± 0.08	9.42 ± 0.03	7.44 ± 0.03	85.09 ± 0.16	0.72 ± 0.00	0.91 ± 0.00	0.15 ± 0.00
	60	30	23.83 ± 0.03	14.57 ± 0.10	11.75 ± 0.12	60.36 ± 0.09	0.64 ± 0.00	0.84 ± 0.00	0.22 ± 0.00
		60	23.71 ± 0.06	14.94 ± 0.06	12.07 ± 0.06	60.78 ± 0.18	0.63 ± 0.00	0.84 ± 0.00	0.22 ± 0.00
		90	23.75 ± 0.08	15.44 ± 0.09	12.42 ± 0.06	60.66 ± 0.26	0.64 ± 0.00	0.84 ± 0.00	0.23 ± 0.00
		120	23.87 ± 0.07	15.50 ± 0.09	12.47 ± 0.08	60.25 ± 0.22	0.64 ± 0.00	0.84 ± 0.00	0.23 ± 0.00
567	30	30	22.72 ± 0.04	16.47 ± 0.04	12.48 ± 0.03	84.80 ± 0.05	0.64 ± 0.00	0.80 ± 0.00	0.28 ± 0.00
		60	22.98 ± 0.07	16.63 ± 0.07	12.93 ± 0.07	84.44 ± 0.10	0.64 ± 0.00	0.80 ± 0.00	0.29 ± 0.00
		90	23.48 ± 0.18	17.24 ± 0.15	13.48 ± 0.12	83.77 ± 0.25	0.62 ± 0.00	0.79 ± 0.00	0.31 ± 0.00
		120	24.18 ± 0.20	17.98 ± 0.15	14.18 ± 0.12	82.78 ± 0.29	0.61 ± 0.00	0.78 ± 0.00	0.32 ± 0.00
	60	30	38.38 ± 0.02	29.51 ± 0.04	23.24 ± 0.06	56.68 ± 0.04	0.46 ± 0.00	0.64 ± 0.00	0.47 ± 0.00
		60	39.00 ± 0.07	29.36 ± 0.01	23.95 ± 0.01	55.27 ± 0.15	0.48 ± 0.00	0.64 ± 0.00	0.48 ± 0.00
		90	39.46 ± 0.07	29.96 ± 0.01	24.71 ± 0.03	54.22 ± 0.17	0.46 ± 0.00	0.63 ± 0.00	0.49 ± 0.00
		120	40.39 ± 0.15	30.91 ± 0.08	25.66 ± 0.09	52.01 ± 0.35	0.44 ± 0.00	0.62 ± 0.00	0.51 ± 0.00
584	30	30	23.25 ± 0.08	16.72 ± 0.06	11.00 ± 0.07	84.88 ± 0.10	0.76 ± 0.00	0.87 ± 0.00	0.23 ± 0.00
		60	22.78 ± 0.04	16.92 ± 0.04	11.34 ± 0.03	85.49 ± 0.05	0.77 ± 0.00	0.87 ± 0.00	0.23 ± 0.00
		90	22.80 ± 0.02	17.17 ± 0.03	11.51 ± 0.02	85.47 ± 0.03	0.76 ± 0.00	0.88 ± 0.00	0.24 ± 0.00
		120	23.30 ± 0.10	17.59 ± 0.10	11.79 ± 0.08	84.82 ± 0.13	0.75 ± 0.00	0.87 ± 0.00	0.25 ± 0.00
	60	30	37.53 ± 0.03	27.65 ± 0.22	18.33 ± 0.27	60.48 ± 0.07	0.59 ± 0.00	0.71 ± 0.01	0.37 ± 0.00
		60	35.99 ± 0.05	27.29 ± 0.02	18.40 ± 0.03	63.67 ± 0.11	0.60 ± 0.00	0.72 ± 0.00	0.37 ± 0.00
		90	36.04 ± 0.06	27.64 ± 0.06	18.72 ± 0.07	63.56 ± 0.12	0.59 ± 0.00	0.72 ± 0.00	0.38 ± 0.00
		120	36.39 ± 0.04	27.83 ± 0.09	18.84 ± 0.12	62.85 ± 0.08	0.58 ± 0.00	0.71 ± 0.00	0.38 ± 0.00
596	30	30	18.66 ± 0.09	13.47 ± 0.11	10.09 ± 0.10	85.82 ± 0.14	0.71 ± 0.00	0.89 ± 0.00	0.21 ± 0.00
		60	17.87 ± 0.08	12.89 ± 0.06	9.67 ± 0.03	86.99 ± 0.12	0.74 ± 0.00	0.89 ± 0.00	0.20 ± 0.00
		90	17.87 ± 0.09	12.93 ± 0.06	9.71 ± 0.03	86.99 ± 0.13	0.75 ± 0.00	0.89 ± 0.00	0.20 ± 0.00
		120	17.95 ± 0.05	12.98 ± 0.03	9.76 ± 0.02	86.89 ± 0.07	0.74 ± 0.00	0.90 ± 0.00	0.20 ± 0.00
	60	30	30.46 ± 0.10	22.78 ± 0.08	17.57 ± 0.08	62.29 ± 0.25	0.52 ± 0.00	0.78 ± 0.00	0.33 ± 0.00
		60	29.00 ± 0.13	21.43 ± 0.14	16.36 ± 0.13	65.83 ± 0.30	0.56 ± 0.00	0.80 ± 0.00	0.31 ± 0.00
		90	28.79 ± 0.05	21.35 ± 0.07	16.28 ± 0.07	66.32 ± 0.13	0.57 ± 0.01	0.80 ± 0.00	0.31 ± 0.00
		120	28.83 ± 0.16	21.37 ± 0.16	16.34 ± 0.16	66.22 ± 0.37	0.57 ± 0.01	0.81 ± 0.00	0.31 ± 0.00

Note. Values in bold indicate the best evaluation outcome for each metric in each learning scenario, and grey highlights denote the best model in each scenario based on the best-achieved evaluation metrics. Note. PID: patient identification; PH: prediction horizon; LL: lag length; RMSE: root mean square error; SD: standard deviation; MAE: mean absolute error; MAPE: mean absolute percentage error; r^2: coefficient of determination; MCC: Matthew's correlation coefficient; SE: surveillance error; ASE: average surveillance error.

Table A3. *The evaluation results for non-stacking models created by long short-term memory learners using Ohio 2018 dataset.*

PID	PH	LL	Evaluation Metric						
			RMSE ± SD (mg/dL)	MAE ± SD (mg/dL)	MAPE ± SD (%)	r^2 ± SD (%)	MCC ± SD (%)	SE < 0.5 ± SD (%)	ASE ± SD
559	30	30	23.12 ± 0.43	16.60 ± 0.66	11.10 ± 0.63	87.19 ± 0.47	0.74 ± 0.01	0.86 ± 0.01	0.24 ± 0.01
		60	23.51 ± 0.36	16.79 ± 0.54	11.02 ± 0.64	86.76 ± 0.40	0.74 ± 0.01	0.87 ± 0.01	0.23 ± 0.01
		90	25.50 ± 1.19	17.44 ± 0.64	10.71 ± 0.13	84.39 ± 1.44	0.72 ± 0.03	0.87 ± 0.01	0.23 ± 0.00
		120	32.86 ± 13.20	23.72 ± 10.60	15.55 ± 8.01	71.35 ± 23.13	0.63 ± 0.19	0.78 ± 0.16	0.31 ± 0.15
	60	30	38.39 ± 0.82	27.05 ± 0.53	16.65 ± 0.21	64.46 ± 1.52	0.57 ± 0.01	0.75 ± 0.00	0.35 ± 0.00
		60	38.73 ± 4.41	27.75 ± 3.58	17.37 ± 1.50	63.53 ± 8.42	0.54 ± 0.07	0.73 ± 0.05	0.37 ± 0.05
		90	37.77 ± 3.27	26.72 ± 2.04	16.92 ± 0.47	65.46 ± 6.01	0.58 ± 0.02	0.75 ± 0.01	0.35 ± 0.02
		120	36.08 ± 1.47	25.38 ± 0.84	16.62 ± 0.25	68.60 ± 2.56	0.59 ± 0.02	0.75 ± 0.01	0.34 ± 0.01
563	30	30	21.59 ± 0.64	15.33 ± 0.45	9.69 ± 0.19	77.31 ± 1.34	0.72 ± 0.01	0.89 ± 0.00	0.22 ± 0.00
		60	21.73 ± 0.46	15.52 ± 0.33	9.82 ± 0.32	77.03 ± 0.96	0.73 ± 0.00	0.89 ± 0.00	0.22 ± 0.01
		90	24.91 ± 1.84	17.49 ± 1.38	10.96 ± 1.02	69.71 ± 4.55	0.69 ± 0.03	0.87 ± 0.02	0.24 ± 0.02
		120	24.04 ± 1.89	16.94 ± 1.15	10.65 ± 0.72	71.79 ± 4.43	0.69 ± 0.01	0.87 ± 0.01	0.24 ± 0.01
	60	30	33.02 ± 0.62	24.13 ± 0.61	15.07 ± 0.18	47.03 ± 2.01	0.51 ± 0.01	0.75 ± 0.02	0.33 ± 0.01
		60	34.44 ± 2.48	25.05 ± 2.24	15.80 ± 1.37	42.17 ± 8.46	0.48 ± 0.09	0.74 ± 0.06	0.35 ± 0.03
		90	34.32 ± 1.23	24.45 ± 1.04	15.16 ± 0.63	42.73 ± 4.13	0.52 ± 0.01	0.77 ± 0.02	0.34 ± 0.01
		120	34.13 ± 1.59	24.66 ± 1.10	15.27 ± 0.62	43.33 ± 5.27	0.50 ± 0.02	0.76 ± 0.02	0.34 ± 0.01
570	30	30	24.78 ± 3.96	18.97 ± 3.76	8.84 ± 1.30	86.33 ± 4.12	0.82 ± 0.01	0.94 ± 0.01	0.16 ± 0.02
		60	25.83 ± 5.11	19.99 ± 4.76	9.28 ± 1.87	85.02 ± 5.59	0.81 ± 0.03	0.93 ± 0.02	0.17 ± 0.03
		90	23.09 ± 2.28	17.15 ± 2.09	8.26 ± 0.74	88.25 ± 2.30	0.82 ± 0.01	0.94 ± 0.00	0.15 ± 0.01
		120	22.92 ± 1.49	16.16 ± 1.15	8.04 ± 0.65	88.47 ± 1.52	0.81 ± 0.02	0.94 ± 0.01	0.15 ± 0.01
	60	30	38.34 ± 2.65	29.98 ± 2.52	13.56 ± 0.95	67.77 ± 4.48	0.75 ± 0.01	0.88 ± 0.01	0.25 ± 0.02
		60	35.80 ± 1.50	26.75 ± 1.85	12.68 ± 0.43	71.95 ± 2.31	0.75 ± 0.00	0.88 ± 0.01	0.23 ± 0.01
		90	37.00 ± 2.48	27.94 ± 1.86	13.17 ± 0.99	69.98 ± 4.09	0.75 ± 0.03	0.87 ± 0.02	0.24 ± 0.02
		120	35.80 ± 2.62	25.82 ± 2.70	12.58 ± 0.95	71.89 ± 4.09	0.75 ± 0.02	0.88 ± 0.01	0.23 ± 0.02
575	30	30	27.20 ± 0.57	18.25 ± 0.45	13.14 ± 0.71	80.24 ± 0.82	0.69 ± 0.00	0.82 ± 0.02	0.28 ± 0.01
		60	27.52 ± 0.76	18.26 ± 0.37	13.07 ± 0.32	79.77 ± 1.13	0.69 ± 0.01	0.82 ± 0.00	0.28 ± 0.01
		90	28.37 ± 0.99	18.89 ± 0.88	13.78 ± 0.69	78.51 ± 1.51	0.68 ± 0.01	0.80 ± 0.01	0.30 ± 0.01
		120	29.33 ± 1.12	19.83 ± 1.63	13.69 ± 0.60	77.03 ± 1.74	0.65 ± 0.05	0.80 ± 0.02	0.29 ± 0.01
	60	30	38.09 ± 0.03	27.47 ± 0.52	20.48 ± 1.20	61.36 ± 0.07	0.54 ± 0.02	0.70 ± 0.00	0.41 ± 0.01
		60	39.96 ± 0.84	28.84 ± 0.27	21.39 ± 1.07	57.46 ± 1.78	0.55 ± 0.03	0.68 ± 0.01	0.44 ± 0.01
		90	38.15 ± 0.52	27.58 ± 0.22	20.56 ± 0.49	61.24 ± 1.06	0.52 ± 0.01	0.68 ± 0.01	0.42 ± 0.01
		120	39.47 ± 1.28	28.64 ± 0.43	21.35 ± 0.44	58.48 ± 2.69	0.54 ± 0.01	0.67 ± 0.01	0.43 ± 0.01
588	30	30	19.23 ± 0.11	14.16 ± 0.11	8.53 ± 0.12	83.77 ± 0.19	0.74 ± 0.00	0.92 ± 0.00	0.19 ± 0.00
		60	19.60 ± 0.23	14.57 ± 0.15	8.83 ± 0.07	83.13 ± 0.39	0.74 ± 0.01	0.92 ± 0.00	0.19 ± 0.00
		90	20.33 ± 0.86	15.00 ± 0.73	8.87 ± 0.36	81.84 ± 1.54	0.73 ± 0.01	0.92 ± 0.00	0.19 ± 0.01
		120	21.99 ± 1.74	16.39 ± 1.07	9.64 ± 0.77	78.69 ± 3.39	0.69 ± 0.02	0.91 ± 0.02	0.20 ± 0.02
	60	30	31.32 ± 0.53	23.12 ± 0.56	14.05 ± 0.68	57.00 ± 1.48	0.57 ± 0.01	0.79 ± 0.02	0.30 ± 0.02
		60	30.46 ± 0.60	22.48 ± 0.39	14.04 ± 0.23	59.33 ± 1.61	0.60 ± 0.01	0.79 ± 0.01	0.30 ± 0.01
		90	32.01 ± 0.53	23.06 ± 0.33	14.11 ± 0.47	55.07 ± 1.48	0.58 ± 0.02	0.80 ± 0.01	0.30 ± 0.01
		120	35.57 ± 4.21	25.60 ± 2.74	15.65 ± 1.69	44.02 ± 13.55	0.50 ± 0.08	0.76 ± 0.03	0.33 ± 0.03
591	30	30	26.00 ± 0.54	19.63 ± 0.54	15.81 ± 0.75	74.78 ± 1.04	0.58 ± 0.01	0.74 ± 0.00	0.35 ± 0.01
		60	26.33 ± 0.42	19.55 ± 0.24	15.65 ± 0.40	74.16 ± 0.83	0.60 ± 0.00	0.75 ± 0.01	0.34 ± 0.01
		90	27.44 ± 1.02	20.46 ± 0.58	15.63 ± 0.98	71.90 ± 2.10	0.55 ± 0.05	0.74 ± 0.01	0.34 ± 0.01
		120	27.16 ± 0.88	20.13 ± 0.63	15.75 ± 0.85	72.48 ± 1.78	0.57 ± 0.03	0.74 ± 0.02	0.34 ± 0.01
	60	30	36.51 ± 0.20	28.36 ± 0.26	23.32 ± 0.27	50.32 ± 0.54	0.37 ± 0.02	0.63 ± 0.00	0.47 ± 0.00
		60	37.52 ± 0.93	28.36 ± 0.32	22.47 ± 0.57	47.52 ± 2.58	0.36 ± 0.04	0.63 ± 0.01	0.47 ± 0.00
		90	37.92 ± 1.44	29.32 ± 1.16	24.31 ± 1.51	46.38 ± 4.10	0.39 ± 0.04	0.63 ± 0.01	0.48 ± 0.01
		120	37.07 ± 1.67	28.38 ± 1.14	22.37 ± 0.89	48.73 ± 4.57	0.37 ± 0.02	0.63 ± 0.02	0.47 ± 0.02

Note. Values in bold indicate the best evaluation outcome for each metric in each learning scenario, and grey highlights denote the best model in each scenario based on the best-achieved evaluation metrics. Note. PID: patient identification; PH: prediction horizon; LL: lag length; RMSE: root mean square error; SD: standard deviation; MAE: mean absolute error; MAPE: mean absolute percentage error; r^2: coefficient of determination; MCC: Matthew's correlation coefficient; SE: surveillance error; ASE: average surveillance error.

Table A4. *The evaluation results for non-stacking models created by long short-term memory learners using Ohio 2020 dataset.*

PID	PH	LL	Evaluation Metric						
			RMSE ± SD (mg/dL)	MAE ± SD (mg/dL)	MAPE ± SD (%)	r^2 ± SD (%)	MCC ± SD (%)	SE < 0.5 ± SD (%)	ASE ± SD
540	30	30	25.76 ± 1.26	19.38 ± 0.62	14.84 ± 0.24	84.25 ± 1.55	0.67 ± 0.01	0.79 ± 0.00	0.31 ± 0.00
		60	24.84 ± 0.42	18.48 ± 0.70	13.81 ± 1.24	85.37 ± 0.49	0.67 ± 0.02	0.80 ± 0.01	0.29 ± 0.02
		90	28.02 ± 3.64	21.40 ± 2.68	15.98 ± 2.30	81.18 ± 4.68	0.63 ± 0.03	0.76 ± 0.03	0.33 ± 0.04
		120	27.92 ± 1.82	21.00 ± 1.99	15.38 ± 2.29	81.48 ± 2.40	0.63 ± 0.02	0.76 ± 0.02	0.32 ± 0.04
	60	30	42.60 ± 1.15	31.84 ± 0.41	23.25 ± 0.53	57.07 ± 2.32	0.48 ± 0.02	0.64 ± 0.01	0.45 ± 0.00
		60	41.36 ± 0.58	30.69 ± 0.37	22.40 ± 0.20	59.56 ± 1.12	0.50 ± 0.02	0.66 ± 0.00	0.44 ± 0.00
		90	43.78 ± 2.80	32.44 ± 2.02	23.51 ± 1.66	54.55 ± 5.78	0.50 ± 0.04	0.64 ± 0.02	0.45 ± 0.02
		120	48.17 ± 1.39	34.62 ± 2.09	24.69 ± 2.33	45.10 ± 3.15	0.48 ± 0.04	0.63 ± 0.03	0.48 ± 0.03
544	30	30	21.23 ± 0.53	15.00 ± 0.49	9.93 ± 0.35	82.45 ± 0.87	0.76 ± 0.01	0.89 ± 0.00	0.21 ± 0.01
		60	20.66 ± 0.31	14.71 ± 0.43	9.99 ± 0.53	83.40 ± 0.50	0.75 ± 0.01	0.88 ± 0.02	0.22 ± 0.01
		90	22.55 ± 0.45	15.56 ± 0.37	10.40 ± 0.27	80.21 ± 0.79	0.72 ± 0.01	0.88 ± 0.01	0.22 ± 0.00
		120	23.38 ± 2.94	16.49 ± 1.81	11.35 ± 1.30	78.51 ± 5.18	0.71 ± 0.04	0.84 ± 0.03	0.24 ± 0.03
	60	30	31.43 ± 0.05	23.19 ± 0.08	15.59 ± 0.16	61.46 ± 0.12	0.58 ± 0.01	0.76 ± 0.00	0.32 ± 0.00
		60	30.45 ± 0.12	22.09 ± 0.45	14.81 ± 0.52	63.83 ± 0.29	0.59 ± 0.02	0.78 ± 0.01	0.31 ± 0.01
		90	32.39 ± 0.61	22.91 ± 0.32	15.40 ± 0.39	59.04 ± 1.55	0.57 ± 0.01	0.76 ± 0.01	0.33 ± 0.01
		120	36.19 ± 1.38	25.61 ± 0.40	17.44 ± 0.10	48.85 ± 3.94	0.52 ± 0.04	0.74 ± 0.01	0.36 ± 0.01
552	30	30	16.72 ± 0.44	10.31 ± 0.24	8.04 ± 0.22	80.45 ± 1.01	0.71 ± 0.02	0.90 ± 0.01	0.16 ± 0.01
		60	21.54 ± 3.51	14.67 ± 3.62	11.21 ± 2.37	66.99 ± 10.53	0.59 ± 0.14	0.85 ± 0.04	0.22 ± 0.04
		90	18.81 ± 1.50	12.58 ± 1.52	9.73 ± 0.98	75.16 ± 3.97	0.69 ± 0.01	0.89 ± 0.01	0.19 ± 0.01
		120	20.91 ± 5.44	14.00 ± 4.23	11.01 ± 3.87	68.05 ± 17.09	0.69 ± 0.08	0.85 ± 0.10	0.22 ± 0.08
	60	30	25.47 ± 0.30	16.27 ± 0.24	13.02 ± 0.27	54.73 ± 1.05	0.61 ± 0.01	0.83 ± 0.01	0.24 ± 0.01
		60	27.15 ± 1.00	18.20 ± 0.92	15.02 ± 0.93	48.51 ± 3.76	0.58 ± 0.03	0.78 ± 0.02	0.28 ± 0.02
		90	27.51 ± 2.98	17.70 ± 1.96	14.55 ± 1.73	46.78 ± 11.78	0.56 ± 0.06	0.80 ± 0.04	0.27 ± 0.04
		120	40.75 ± 25.37	32.17 ± 26.99	26.17 ± 21.83	45.82 ± 170.04	0.33 ± 0.44	0.60 ± 0.38	0.53 ± 0.49
567	30	30	26.21 ± 1.00	18.74 ± 1.00	14.41 ± 1.01	79.74 ± 1.56	0.61 ± 0.01	0.77 ± 0.01	0.32 ± 0.02
		60	25.54 ± 0.32	18.38 ± 0.28	13.83 ± 0.55	80.78 ± 0.48	0.61 ± 0.01	0.78 ± 0.00	0.31 ± 0.01
		90	24.64 ± 0.97	17.85 ± 0.81	13.48 ± 0.66	82.10 ± 1.41	0.60 ± 0.01	0.78 ± 0.01	0.31 ± 0.01
		120	27.89 ± 3.45	20.96 ± 3.26	16.17 ± 2.94	76.86 ± 5.47	0.57 ± 0.05	0.74 ± 0.04	0.35 ± 0.06
	60	30	43.16 ± 1.27	32.69 ± 1.21	27.34 ± 1.23	45.19 ± 3.24	0.44 ± 0.02	0.60 ± 0.02	0.53 ± 0.02
		60	40.13 ± 1.22	30.57 ± 1.14	25.05 ± 1.96	52.61 ± 2.86	0.45 ± 0.01	0.62 ± 0.02	0.50 ± 0.03
		90	42.89 ± 2.29	32.84 ± 2.03	26.97 ± 2.57	45.79 ± 5.74	0.41 ± 0.01	0.60 ± 0.02	0.53 ± 0.03
		120	45.08 ± 4.52	34.30 ± 3.01	26.78 ± 0.56	39.83 ± 12.30	0.40 ± 0.06	0.58 ± 0.04	0.54 ± 0.04
584	30	30	26.87 ± 0.77	19.56 ± 0.72	13.10 ± 0.55	79.81 ± 1.16	0.72 ± 0.02	0.84 ± 0.01	0.26 ± 0.01
		60	25.31 ± 1.32	18.27 ± 0.95	11.49 ± 0.52	82.05 ± 1.89	0.75 ± 0.01	0.86 ± 0.01	0.23 ± 0.01
		90	25.93 ± 1.03	19.25 ± 0.82	13.00 ± 0.65	81.19 ± 1.47	0.74 ± 0.01	0.85 ± 0.01	0.26 ± 0.01
		120	27.62 ± 0.80	20.65 ± 1.21	13.36 ± 0.35	78.66 ± 1.24	0.72 ± 0.02	0.84 ± 0.01	0.27 ± 0.00
	60	30	41.45 ± 1.58	31.50 ± 1.91	21.43 ± 2.17	51.75 ± 3.64	0.55 ± 0.03	0.67 ± 0.04	0.42 ± 0.04
		60	42.14 ± 1.60	32.72 ± 1.78	23.12 ± 1.60	50.12 ± 3.74	0.55 ± 0.01	0.64 ± 0.04	0.45 ± 0.03
		90	41.75 ± 0.90	32.60 ± 0.83	22.86 ± 1.00	51.08 ± 2.11	0.56 ± 0.01	0.65 ± 0.02	0.44 ± 0.02
		120	47.83 ± 3.54	37.15 ± 4.34	25.97 ± 4.37	35.58 ± 9.66	0.46 ± 0.05	0.59 ± 0.07	0.50 ± 0.08
596	30	30	19.96 ± 0.28	14.31 ± 0.03	10.83 ± 0.18	83.78 ± 0.45	0.70 ± 0.01	0.87 ± 0.00	0.23 ± 0.00
		60	21.15 ± 0.65	15.31 ± 0.40	11.64 ± 0.41	81.77 ± 1.12	0.69 ± 0.01	0.86 ± 0.01	0.24 ± 0.01
		90	22.54 ± 0.82	16.38 ± 0.95	12.32 ± 0.90	79.29 ± 1.50	0.66 ± 0.04	0.85 ± 0.01	0.25 ± 0.01
		120	33.46 ± 10.29	25.29 ± 8.45	19.64 ± 6.92	51.54 ± 25.67	0.50 ± 0.16	0.75 ± 0.10	0.36 ± 0.11
	60	30	30.97 ± 0.19	22.79 ± 0.17	17.23 ± 0.22	61.02 ± 0.48	0.52 ± 0.01	0.78 ± 0.00	0.33 ± 0.00
		60	30.28 ± 0.72	22.17 ± 0.71	16.97 ± 0.45	62.72 ± 1.77	0.56 ± 0.02	0.79 ± 0.00	0.32 ± 0.01
		90	31.70 ± 1.25	23.44 ± 1.22	17.94 ± 1.21	59.12 ± 3.24	0.52 ± 0.03	0.78 ± 0.01	0.34 ± 0.02
		120	36.31 ± 9.68	27.21 ± 8.48	21.03 ± 6.87	43.87 ± 30.66	0.43 ± 0.21	0.71 ± 0.13	0.40 ± 0.11

Note. Values in bold indicate the best evaluation outcome for each metric in each learning scenario, and grey highlights denote the best model in each scenario based on the best-achieved evaluation metrics. Note. PID: patient identification; PH: prediction horizon; LL: lag length; RMSE: root mean square error; SD: standard deviation; MAE: mean absolute error; MAPE: mean absolute percentage error; r^2: coefficient of determination; MCC: Matthew's correlation coefficient; SE: surveillance error; ASE: average surveillance error.

References

1. DiMeglio, L.A.; Evans-Molina, C.; Oram, R.A. Type 1 Diabetes. *Lancet* **2018**, *391*, 2449–2462. https://doi.org/10.1016/S0140-6736(18)31320-5.
2. Melin, J.; Lynch, K.F.; Lundgren, M.; Aronsson, C.A.; Larsson, H.E.; Johnson, S.B.; Rewers, M.; Barbour, A.; Bautista, K.; Baxter, J.; et al. Is Staff Consistency Important to Parents' Satisfaction in a Longitudinal Study of Children at Risk for Type 1 Diabetes: The TEDDY Study. *BMC Endocr. Disord.* **2022**, *22*, 19. https://doi.org/10.1186/S12902-021-00929-W/FIGURES/2.
3. Khadem, H.; Nemat, H.; Elliott, J.; Benaissa, M. Interpretable Machine Learning for Inpatient COVID-19 Mortality Risk Assessments: Diabetes Mellitus Exclusive Interplay. *Sensors* **2022**, *22*, 8757. https://doi.org/10.3390/s22228757.
4. Yamada, T.; Shojima, N.; Noma, H.; Yamauchi, T.; Kadowaki, T. Sodium-Glucose Co-Transporter-2 Inhibitors as Add-on Therapy to Insulin for Type 1 Diabetes Mellitus: Systematic Review and Meta-Analysis of Randomized Controlled Trials. *Diabetes, Obes. Metab.* **2018**, *20*, 1755–1761. https://doi.org/10.1111/dom.13260.
5. Smith, A.; Harris, C. Type 1 Diabetes: Management Strategies. *Am. Fam. Physician* **2018**, *98*, 154–162.
6. Hamilton, K.; Stanton-Fay, S.H.; Chadwick, P.M.; Lorencatto, F.; de Zoysa, N.; Gianfrancesco, C.; Taylor, C.; Coates, E.; Breckenridge, J.P.; Cooke, D.; et al. Sustained Type 1 Diabetes Self-Management: Specifying the Behaviours Involved and Their Influences. *Diabet. Med.* **2021**, *38*, e14430. https://doi.org/10.1111/DME.14430.
7. Campbell, F.; Lawton, J.; Rankin, D.; Clowes, M.; Coates, E.; Heller, S.; De Zoysa, N.; Elliott, J.; Breckenridge, J.P. Follow-Up Support for Effective Type 1 Diabetes Self-Management (The FUSED Model): A Systematic Review and Meta-Ethnography of the Barriers, Facilitators and Recommendations for Sustaining Self-Management Skills after Attending a Structured Education Programme. *BMC Health Serv. Res.* **2018**, *18*, 898. https://doi.org/10.1186/S12913-018-3655-Z/TABLES/6.
8. Cummings, C.; Benjamin, N.E.; Prabhu, H.Y.; Cohen, L.B.; Goddard, B.J.; Kaugars, A.S.; Humiston, T.; Lansing, A.H. Habit and Diabetes Self-Management in Adolescents With Type 1 Diabetes. *Health Psychol.* **2022**, *41*, 13–22. https://doi.org/10.1037/hea0001097.
9. McCarthy, M.M.; Grey, M. Type 1 Diabetes Self-Management From Emerging Adulthood Through Older Adulthood. *Diabetes Care* **2018**, *41*, 1608–1614. https://doi.org/10.2337/DC17-2597.
10. Saoji, N.; Palta, M.; Young, H.N.; Moreno, M.A.; Rajamanickam, V.; Cox, E.D. The Relationship of

Type 1 Diabetes Self-Management Barriers to Child and Parent Quality of Life: A US Cross-Sectional Study. *Diabet. Med.* **2018**, *35*, 1523–1530. https://doi.org/10.1111/DME.13760.

11. Butler, A.M.; Weller, B.E.; Rodgers, C.R.R.; Teasdale, A.E. Type 1 Diabetes Self-Management Behaviors among Emerging Adults: Racial/Ethnic Differences. *Pediatr. Diabetes* **2020**, *21*, 979–986. https://doi.org/10.1111/PEDI.13061.
12. Dai, X.; Luo, Z.C; Zhai, L.; Zhao, W.P.; Huang, F. Artificial Pancreas as an Effective and Safe Alternative in Patients with Type 1 Diabetes Mellitus: A Systematic Review and Meta-Analysis. *Diabetes Ther.* **2018**, *9*, 1269–1277. https://doi.org/10.1007/S13300-018-0436-Y/FIGURES/3.
13. Bekiari, E.; Kitsios, K.; Thabit, H.; Tauschmann, M.; Athanasiadou, E.; Karagiannis, T.; Haidich, A.B.; Hovorka, R.; Tsapas, A. Artificial Pancreas Treatment for Outpatients with Type 1 Diabetes: Systematic Review and Meta-Analysis. *BMJ* **2018**, *361*, 1310. https://doi.org/10.1136/BMJ.K1310.
14. Zhang, Y.; Sun, J.; Liu, L.; Qiao, H. A Review of Biosensor Technology and Algorithms for Glucose Monitoring. *J. Diabetes Complications* **2021**, *35*, 107929. https://doi.org/10.1016/J.JDIACOMP.2021.107929.
15. Choudhary, P.; Amiel, S.A. Hypoglycaemia in Type 1 Diabetes: Technological Treatments, Their Limitations and the Place of Psychology. *Diabetologia* **2018**, *61*, 761–769. https://doi.org/10.1007/S00125-018-4566-6/FIGURES/1.
16. Tagougui, S.; Taleb, N.; Rabasa-Lhoret, R. The Benefits and Limits of Technological Advances in Glucose Management around Physical Activity in Patients Type 1 Diabetes. *Front. Endocrinol. (Lausanne).* **2019**, *10*, 818. https://doi.org/10.3389/FENDO.2018.00818/BIBTEX.
17. Laffel, L.M.; Kanapka, L.G.; Beck, R.W.; Bergamo, K.; Clements, M.A.; Criego, A.; Desalvo, D.J.; Goland, R.; Hood, K.; Liljenquist, D.; et al. Effect of Continuous Glucose Monitoring on Glycemic Control in Adolescents and Young Adults With Type 1 Diabetes: A Randomized Clinical Trial. *JAMA* **2020**, *323*, 2388–2396. https://doi.org/10.1001/JAMA.2020.6940.
18. Martens, T.; Beck, R.W.; Bailey, R.; Ruedy, K.J.; Calhoun, P.; Peters, A.L.; Pop-Busui, R.; Philis-Tsimikas, A.; Bao, S.; Umpierrez, G.; et al. Effect of Continuous Glucose Monitoring on Glycemic Control in Patients With Type 2 Diabetes Treated With Basal Insulin: A Randomized Clinical Trial. *JAMA* **2021**, *325*, 2262–2272. https://doi.org/10.1001/JAMA.2021.7444.
19. Pickup, J.C. Is Insulin Pump Therapy Effective in Type 1 Diabetes? *Diabet. Med.* **2019**, *36*, 269–278. https://doi.org/10.1111/DME.13793.
20. Ranjan, A.G.; Rosenlund, S.V.; Hansen, T.W.; Rossing, P.; Andersen, S.; Nørgaard, K. Improved Time in Range Over 1 Year Is Associated With Reduced Albuminuria in Individuals With Sensor-Augmented Insulin Pump–Treated Type 1 Diabetes. *Diabetes Care* **2020**, *43*, 2882–2885.

https://doi.org/10.2337/DC20-0909.

21. Mian, Z.; Hermayer, K.L.; Jenkins, A. Continuous Glucose Monitoring: Review of an Innovation in Diabetes Management. *Am. J. Med. Sci.* **2019**, *358*, 332–339. https://doi.org/10.1016/J.AMJMS.2019.07.003.
22. Aggarwal, A.; Pathak, S.; Goyal, R. Clinical and Economic Outcomes of Continuous Glucose Monitoring System (CGMS) in Patients with Diabetes Mellitus: A Systematic Literature Review. *Diabetes Res. Clin. Pract.* **2022**, *186*, 109825. https://doi.org/10.1016/J.DIABRES.2022.109825.
23. Burckhardt, M.A.; Smith, G.J.; Cooper, M.N.; Jones, T.W.; Davis, E.A. Real-World Outcomes of Insulin Pump Compared to Injection Therapy in a Population-Based Sample of Children with Type 1 Diabetes. *Pediatr. Diabetes* **2018**, *19*, 1459–1466. https://doi.org/10.1111/PEDI.12754.
24. Cardona-Hernandez, R.; Schwandt, A.; Alkandari, H.; Bratke, H.; Chobot, A.; Coles, N.; Corathers, S.; Goksen, D.; Goss, P.; Imane, Z.; et al. Glycemic Outcome Associated With Insulin Pump and Glucose Sensor Use in Children and Adolescents With Type 1 Diabetes. Data From the International Pediatric Registry SWEET. *Diabetes Care* **2021**, *44*, 1176–1184. https://doi.org/10.2337/DC20-1674.
25. Rytter, K.; Schmidt, S.; Rasmussen, L.N.; Pedersen-Bjergaard, U.; Nørgaard, K. Education Programmes for Persons with Type 1 Diabetes Using an Insulin Pump: A Systematic Review. *Diabetes. Metab. Res. Rev.* **2021**, *37*, e3412. https://doi.org/10.1002/DMRR.3412.
26. Vashist, S.K. Non-Invasive Glucose Monitoring Technology in Diabetes Management: A Review. *Anal. Chim. Acta* **2012**, *750*, 16–27. https://doi.org/10.1016/j.aca.2012.03.043.
27. Alrezj, O.; Benaissa, M.; Alshebeili, S.A. Digital Bandstop Filtering in the Quantitative Analysis of Glucose from Near-Infrared and Midinfrared Spectra. *J. Chemom.* **2020**, *34*, e3206. https://doi.org/10.1002/CEM.3206.
28. Khadem, H.; Nemat, H.; Elliott, J.; Benaissa, M. Signal Fragmentation Based Feature Vector Generation in a Model Agnostic Framework with Application to Glucose Quantification Using Absorption Spectroscopy. *Talanta* **2022**, *243*, 123379. https://doi.org/10.1016/j.talanta.2022.123379.
29. Khadem, H.; Eissa, M.R.; Nemat, H.; Alrezj, O.; Benaissa, M. Classification before Regression for Improving the Accuracy of Glucose Quantification Using Absorption Spectroscopy. *Talanta* **2020**, *211*, 120740. https://doi.org/10.1016/j.talanta.2020.120740.
30. Vettoretti, M.; Cappon, G.; Facchinetti, A.; Sparacino, G. Advanced Diabetes Management Using Artificial Intelligence and Continuous Glucose Monitoring Sensors. *Sensors* **2020**, *20*, 3870. https://doi.org/10.3390/S20143870.
31. Nemat, H.; Khadem, H.; Elliott, J.; Benaissa, M. Causality Analysis in Type 1 Diabetes Mellitus with Application to Blood Glucose Level Prediction. *Comput. Biol. Med.* **2023**, *153*, 106535.

https://doi.org/10.1016/j.compbiomed.2022.106535.

32. Xie, J.; Wang, Q. Benchmarking Machine Learning Algorithms on Blood Glucose Prediction for Type i Diabetes in Comparison with Classical Time-Series Models. *IEEE Trans. Biomed. Eng.* **2020**, *67*, 3101–3124. https://doi.org/10.1109/TBME.2020.2975959.
33. Nemat, H.; Khadem, H.; Elliott, J.; Benaissa, M. Data Fusion of Activity and CGM for Predicting Blood Glucose Levels. In *Knowledge Discovery in Healthcare Data 2020, Proceedings of the 5th International Workshop on Knowledge Discovery in Healthcare Data co-located with 24th European Conference on Artificial Intelligence (ECAI 2020), Santiago de Compostela, Spain (virtual), 29–30 Aug 2020*; Bach, K., Bunescu, R., Marling, C., Wiratunga, N., Eds.; CEUR Workshop Proceedings: 2020; Volume 2675, pp. 120–124.
34. Woldaregay, A.Z.; Årsand, E.; Botsis, T.; Albers, D.; Mamykina, L.; Hartvigsen, G. Data-Driven Blood Glucose Pattern Classification and Anomalies Detection: Machine-Learning Applications in Type 1 Diabetes. *J. Med. Internet Res.* **2019**, *21*, e11030. https://doi.org/10.2196/11030.
35. Khadem, H.; Nemat, H.; Elliott, J.; Benaissa, M. Multi-Lag Stacking for Blood Glucose Level Prediction. In *Knowledge Discovery in Healthcare Data 2020, Proceedings of the 5th International Workshop on Knowledge Discovery in Healthcare Data co-located with 24th European Conference on Artificial Intelligence (ECAI 2020), Santiago de Compostela, Spain (virtual), 29–30 Aug 2020*; Bach, K., Bunescu, R., Marling, C., Wiratunga, N., Eds.; CEUR Workshop Proceedings: 2020; Volume 2675, pp. 146–150.
36. Boughton, C.K.; Hovorka, R. Is an Artificial Pancreas (Closed-Loop System) for Type 1 Diabetes Effective? *Diabet. Med.* **2019**, *36*, 279–286. https://doi.org/10.1111/DME.13816.
37. Bremer, A.A.; Arreaza-Rubín, G. Analysis of "Artificial Pancreas (AP) Systems for People With Type 2 Diabetes: Conception and Design of the European CLOSE Project". *J. Diabetes Sci. Technol.* **2019**, *13*, 268–270. https://doi.org/10.1177/1932296818823770.
38. Woldaregay, A.Z.; Årsand, E.; Walderhaug, S.; Albers, D.; Mamykina, L.; Botsis, T.; Hartvigsen, G. Data-Driven Modeling and Prediction of Blood Glucose Dynamics: Machine Learning Applications in Type 1 Diabetes. *Artif. Intell. Med.* **2019**, *98*, 109–134. https://doi.org/10.1016/J.ARTMED.2019.07.007.
39. Nemat, H.; Khadem, H.; Eissa, M.R.; Elliott, J.; Benaissa, M. Blood Glucose Level Prediction: Advanced Deep-Ensemble Learning Approach. *IEEE J. Biomed. Health Inform.* **2022**, *26*, 2758–2769. https://doi.org/10.1109/JBHI.2022.3144870.
40. Felizardo, V.; Garcia, N.M.; Pombo, N.; Megdiche, I. Data-Based Algorithms and Models Using Diabetics Real Data for Blood Glucose and Hypoglycaemia Prediction—A Systematic Literature

Review. *Artif. Intell. Med.* **2021**, *118*, 102120. https://doi.org/10.1016/J.ARTMED.2021.102120.

41. Semenoglou, A.-A.; Spiliotis, E.; Assimakopoulos, V. Image-Based Time Series Forecasting: A Deep Convolutional Neural Network Approach. *Neural Netw.* **2023**, *157*, 39–53. https://doi.org/10.1016/J.NEUNET.2022.10.006.
42. Garg, A.; Zhang, W.; Samaran, J.; Savitha, R.; Foo, C.S. An Evaluation of Anomaly Detection and Diagnosis in Multivariate Time Series. *IEEE Trans. Neural Networks Learn. Syst.* **2022**, *33*, 2508–2517. https://doi.org/10.1109/TNNLS.2021.3105827.
43. De Oliveira, J.F.L.; Silva, E.G.; De Mattos Neto, P.S.G. A Hybrid System Based on Dynamic Selection for Time Series Forecasting. *IEEE Trans. Neural Netw. Learn. Syst.* **2022**, *33*, 3251–3263. https://doi.org/10.1109/TNNLS.2021.3051384.
44. Cichos, F.; Gustavsson, K.; Mehlig, B.; Volpe, G. Machine Learning for Active Matter. *Nat. Mach. Intell.* **2020**, *2*, 94–103. https://doi.org/10.1038/s42256-020-0146-9.
45. Lim, B.; Zohren, S. Time-Series Forecasting with Deep Learning: A Survey. *Philos. Trans. R. Soc. A* **2021**, *379*, 20200209. https://doi.org/10.1098/RSTA.2020.0209.
46. Ismail Fawaz, H.; Forestier, G.; Weber, J.; Idoumghar, L.; Muller, P.A. Deep Learning for Time Series Classification: A Review. *Data Min. Knowl. Discov.* **2019**, *33*, 917–963. https://doi.org/10.1007/S10618-019-00619-1.
47. Zhu, T.; Wang, W.; Yu, M. A Novel Blood Glucose Time Series Prediction Framework Based on a Novel Signal Decomposition Method. *Chaos Solitons Fractals* **2022**, *164*, 112673. https://doi.org/10.1016/J.CHAOS.2022.112673.
48. Tejedor, M.; Woldaregay, A.Z.; Godtliebsen, F. Reinforcement Learning Application in Diabetes Blood Glucose Control: A Systematic Review. *Artif. Intell. Med.* **2020**, *104*, 101836. https://doi.org/10.1016/J.ARTMED.2020.101836.
49. Aiello, E.M.; Lisanti, G.; Magni, L.; Musci, M.; Toffanin, C. Therapy-Driven Deep Glucose Forecasting. *Eng. Appl. Artif. Intell.* **2020**, *87*, 103255. https://doi.org/10.1016/J.ENGAPPAI.2019.103255.
50. Asad, M.; Qamar, U. A Review of Continuous Blood Glucose Monitoring and Prediction of Blood Glucose Level for Diabetes Type 1 Patient in Different Prediction Horizons (PH) Using Artificial Neural Network (ANN). *Adv. Intell. Syst. Comput.* **2020**, *1038*, 684–695. https://doi.org/10.1007/978-3-030-29513-4_51/COVER.
51. Li, K.; Daniels, J.; Liu, C.; Herrero, P.; Georgiou, P. Convolutional Recurrent Neural Networks for Glucose Prediction. *IEEE J. Biomed. Health Inform.* **2020**, *24*, 603–613. https://doi.org/10.1109/JBHI.2019.2908488.

52. Zhang, M.; Flores, K.B.; Tran, H.T. Deep Learning and Regression Approaches to Forecasting Blood Glucose Levels for Type 1 Diabetes. *Biomed. Signal Process. Control* **2021**, *69*, 102923. https://doi.org/10.1016/J.BSPC.2021.102923.
53. Tena, F.; Garnica, O.; Lanchares, J.; Hidalgo, J.I.; Cappon, G.; Herrero, P.; Sacchi, L.; Coltro, W. Ensemble Models of Cutting-Edge Deep Neural Networks for Blood Glucose Prediction in Patients with Diabetes. *Sensors* **2021**, *21*, 7090. https://doi.org/10.3390/S21217090.
54. Wadghiri, M.Z.; Idri, A.; El Idrissi, T.; Hakkoum, H. Ensemble Blood Glucose Prediction in Diabetes Mellitus: A Review. *Comput. Biol. Med.* **2022**, *147*, 105674. https://doi.org/10.1016/J.COMPBIOMED.2022.105674.
55. Daniels, J.; Herrero, P.; Georgiou, P. A Multitask Learning Approach to Personalized Blood Glucose Prediction. *IEEE J. Biomed. Health Inform.* **2022**, *26*, 436–445. https://doi.org/10.1109/JBHI.2021.3100558.
56. Yang, T.; Yu, X.; Ma, N.; Wu, R.; Li, H. An Autonomous Channel Deep Learning Framework for Blood Glucose Prediction. *Appl. Soft Comput.* **2022**, *120*, 108636. https://doi.org/10.1016/j.asoc.2022.108636.
57. Zhu, T.; Li, K.; Chen, J.; Herrero, P.; Georgiou, P. Dilated Recurrent Neural Networks for Glucose Forecasting in Type 1 Diabetes. *J. Healthc. Inform. Res.* **2020**, *4*, 308–324. https://doi.org/10.1007/s41666-020-00068-2.
58. Martinsson, J.; Schliep, A.; Eliasson, B.; Mogren, O. Blood Glucose Prediction with Variance Estimation Using Recurrent Neural Networks. *J. Healthc. Inform. Res.* **2020**, *4*, 1–18. https://doi.org/10.1007/S41666-019-00059-Y/FIGURES/8.
59. Rodríguez-Rodríguez, I.; Rodríguez, J.V.; Molina-García-Pardo, J.M.; Zamora-Izquierdo, M.Á.; Martínez-Inglés, M.T. A Comparison of Different Models of Glycemia Dynamics for Improved Type 1 Diabetes Mellitus Management with Advanced Intelligent Analysis in an Internet of Things Context. *Appl. Sci.* **2020**, *10*, 4381. https://doi.org/10.3390/APP10124381.
60. Marling, C.; Bunescu, R. The OhioT1DM Dataset for Blood Glucose Level Prediction: Update 2020. In *Proceedings of the International Workshop on Knowledge Discovery in Healthcare Data*; NIH Public Access: Bethesda, MD, USA, 2020; Volume 2675, pp. 71–74.
61. Kwiatkowski, D.; Phillips, P.C.B.; Schmidt, P.; Shin, Y. Testing the Null Hypothesis of Stationarity against the Alternative of a Unit Root: How Sure Are We That Economic Time Series Have a Unit Root? *J. Econom.* **1992**, *54*, 159–178. https://doi.org/10.1016/0304-4076(92)90104-Y.
62. Dickey, D.A.; Fuller, W.A. Distribution of the Estimators for Autoregressive Time Series with a Unit Root. *J. Am. Stat. Assoc.* **2012**, *74*, 427–431. https://doi.org/10.1080/01621459.1979.10482531.

63. Sagi, O.; Rokach, L. Ensemble Learning: A Survey. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2018**, *8*, e1249. https://doi.org/10.1002/WIDM.1249.
64. Breiman, L. Stacked Regressions. *Mach. Learn.* **1996**, *24*, 49–64. https://doi.org/10.1007/BF00117832.
65. Zhu, Q. On the Performance of Matthews Correlation Coefficient (MCC) for Imbalanced Dataset. *Pattern Recognit. Lett.* **2020**, *136*, 71–80. https://doi.org/10.1016/J.PATREC.2020.03.030.
66. Klonoff, D.C.; Lias, C.; Vigersky, R.; Clarke, W.; Parkes, J.L.; Sacks, D.B.; Kirkman, M.S.; Kovatchev, B. The Surveillance Error Grid. *J. Diabetes Sci. Technol.* **2014**, *8*, 658–672. https://doi.org/10.1177/1932296814539589.
67. Friedman, M. A Comparison of Alternative Tests of Significance for the Problem of m Rankings on JSTOR. *Ann. Math. Stat.* **1940**, *11*, 86–92.
68. Fisher, R. Statistical Methods and Scientific Induction. *J. R. Stat. Soc. Ser. B* **1955**, *17*, 69–78. https://doi.org/10.1111/J.2517-6161.1955.TB00180.X.
69. Nemenyi, P.B. *Distribution-Free Multiple Comparisons*; Princeton University: Princeton, NJ, USA, 1963.
70. Holm, S. A Simple Sequentially Rejective Multiple Test Procedure. *Scand. J. Stat.* **1979**, *6*, 65–70.
71. Demšar, J. Statistical Comparisons of Classifiers over Multiple Data Sets. *J. Mach. Learn. Res.* **2006**, *7*, 1–30.
72. Van Rossum, G.; Drake, F.L. *Python 3 Reference Manual*; CreateSpace: Scotts Valley, CA, USA, 2009; ISBN 1441412697.
73. Abadi, M.; Barham, P.; Chen, J.; Chen, Z.; Davis, A.; Dean, J.; Devin, M.; Ghemawat, S.; Irving, G.; Isard, M.; et al. Tensorflow: A System for Large-Scale Machine Learning. In Proceedings of the 12th Symposium on Operating Systems Design and Implementation, Savannah, GA, USA, 2–4 November 2016; pp. 265–283.
74. McKinney, W. Data Structures for Statistical Computing in Python. In Proceedings of the the 9th Python in Science Conference, Austin, TX, USA, 28 June–3 July *2010*; Volume 445, pp. 51–56.
75. Harris, C.R.; Millman, K.J.; van der Walt, S.J.; Gommers, R.; Virtanen, P.; Cournapeau, D.; Wieser, E.; Taylor, J.; Berg, S.; Smith, N.J.; et al. Array Programming with {NumPy}. *Nature* **2020**, *585*, 357–362. https://doi.org/10.1038/s41586-020-2649-2.
76. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
77. Virtanen, P.; Gommers, R.; Oliphant, T.E.; Haberland, M.; Reddy, T.; Cournapeau, D.; Burovski, E.; Peterson, P.; Weckesser, W.; Bright, J.; et al. SciPy 1.0: Fundamental Algorithms for Scientific

Computing in Python. *Nat. Methods* **2020**, *17*, 261–272. https://doi.org/10.1038/S41592-019-0686-2.

78. Seabold, S.; Perktold, J. Statsmodels: Econometric and Statistical Modeling with Python. In Proceedings of the 9th Python in Science Conference, Austin, TX, USA, 28 June–3 July *2010*.
79. Terpilowski, M. Scikit-Posthocs: Pairwise Multiple Comparison Tests in Python. *J. Open Source Softw.* **2019**, *4*, 1169. https://doi.org/10.21105/joss.01169.
80. Benavoli, A.; Corani, G.; Mangili, F. Should We Really Use Post-Hoc Tests Based on Mean-Ranks? *J. Mach. Learn. Res.* **2016**, *17*, 152–161.

Publication 3.

Classification before Regression for Improving the Accuracy of Glucose Quantification using Absorption Spectroscopy[3]

Abstract. This work contributes to the improvement of glucose quantification using near-infrared (NIR), mid-infrared (MIR), and a combination of NIR and MIR absorbance spectroscopy by classifying the spectral data prior to the application of regression models. Both manual and automated classification are presented based on three homogeneous classes defined following the clinical definition of the glycaemic ranges (hypoglycaemia, euglycaemia, and hyperglycaemia). For the manual classification, partial least squares and principal component regressions are applied to each class separately and shown to lead to improved quantification results compared to when applying the same regression models for the whole dataset. For the automatic classification, linear discriminant analysis coupled with principal component analysis is deployed, and regressions are applied to each class separately. The results obtained are shown to outperform those of regressions for the entire dataset.

Keywords. Glucose; Non-invasive; Near-infrared; Mid-infrared; Spectroscopy

1. Introduction

The importance of the development of non-invasive glucose monitoring in diabetes management has spurred research into the quantification of glucose through in vivo and in vitro experiments [1], [2]. The underlying modalities pursued in these studies can be listed as; near-infrared (NIR), mid-infrared (MIR), Raman and bio-impedance spectroscopy, electromagnetic sensing, fluorescence technology, optical

---

[3] This article was published in Talanta. 211 (2020). Authors: **H. Khadem**, M.R. Eissa, H. Nemat, O. Alrezj, M. Benaissa.

coherence tomography, optical polarimetry, reverse iontophoresis, and ultrasound technology [3]. Of all the techniques mentioned above, NIR and MIR spectroscopy are promising and commonly used methods [4], [5].

NIR and MIR spectroscopy use light beams in the wavelength range of 750–2500nm (13333–4000cm^{-1}) and 2500–10000nm (4000–1000cm^{-1}), respectively [6]. These technologies are not expensive for frequent measurements as it does not use any specific reagent [7]. One advantage of the MIR method is decreased scattering phenomena and increased absorption because of higher wavelengths compared with NIR spectroscopy [8]. Moreover, the peaks of glucose are sharper in the MIR region [9]. NIR light, on the other hand, possesses a deep penetration length [10] and could traverse through different skin layers to reach the subcutaneous area [11].

While travelling through a sample, some frequencies of NIR/MIR light are absorbed and scattered because of the interaction with the physiological compounds of the skin [12]. These absorption and scattering measures are used for the quantification analysis of the glucose or other chromophores in the sample [6].

For extracting glucose-related information from the NIR and MIR spectra, multivariate calibration methods such as partial least squares regression (PLSR), principal component regression (PCR), multiple linear regression, artificial neural networks, and support vector machine regression are typically applied to the recorded signals [13]. For improving the accuracy of the analyses, many pre-processing methods have also been proposed [14], such as multivariate scatter correction (MSC), smoothing, and digital band-pass filtering. However, accurate quantification results remain a challenge [15].

This paper proposes a classification-before-regression methodology to improve glucose measurement using NIR, MIR and a combination of NIR and MIR (hereafter referred to as NIR-NIR) spectroscopy. Both manual and automatic classification are carried out by classifying the dataset into three more homogeneous groups following the clinical definition of the glycaemic ranges (hypoglycaemia, euglycaemia, and hyperglycaemia). Partial least squares and principal component regressions are applied with the manual

classification; for the automatic classification, linear discriminant analysis coupled with principal component analysis in both cases is deployed, and regressions are created for each class. The results obtained using the same data for both cases are shown to outperform the results obtained when no classification-before-regression is used.

Classification of spectral data before regressions has been previously used in other research areas, such as rapid analysis of coal properties using NIR spectroscopy [16], [17]. However, this is the first paper to our knowledge that correlates spectral data using a pre-classification approach in order to improve the accuracy of glucose measurement.

2. Dataset

This study utilised absorbance spectroscopic data obtained from a synthesised blood sample set consisting of 100 samples with phosphate (0.01 M/dl), human serum albumin (5 g/dl), and glucose with concentrations ranging from 5mg/dl to 500mg/dl, at intervals of 5mg/dl. The recorded spectra encompassed a wavelength range of 2100–8000nm with a resolution of 1.7nm. Specifically, the wavelengths from 2100 to 2500nm corresponded to the NIR region, while those from 2500 to 8000nm belonged to the MIR region.

3. Methods

3.1. Quantification methods

As shown in Figure 1, three methods were developed in this work for glucose quantification from the collected NIR, MIR, and NIR-NIR spectral data. For data analysis in these quantification methods, we used Python (3.6.7), scikit-learn (0.15.2), and SciPy (0.12.0).

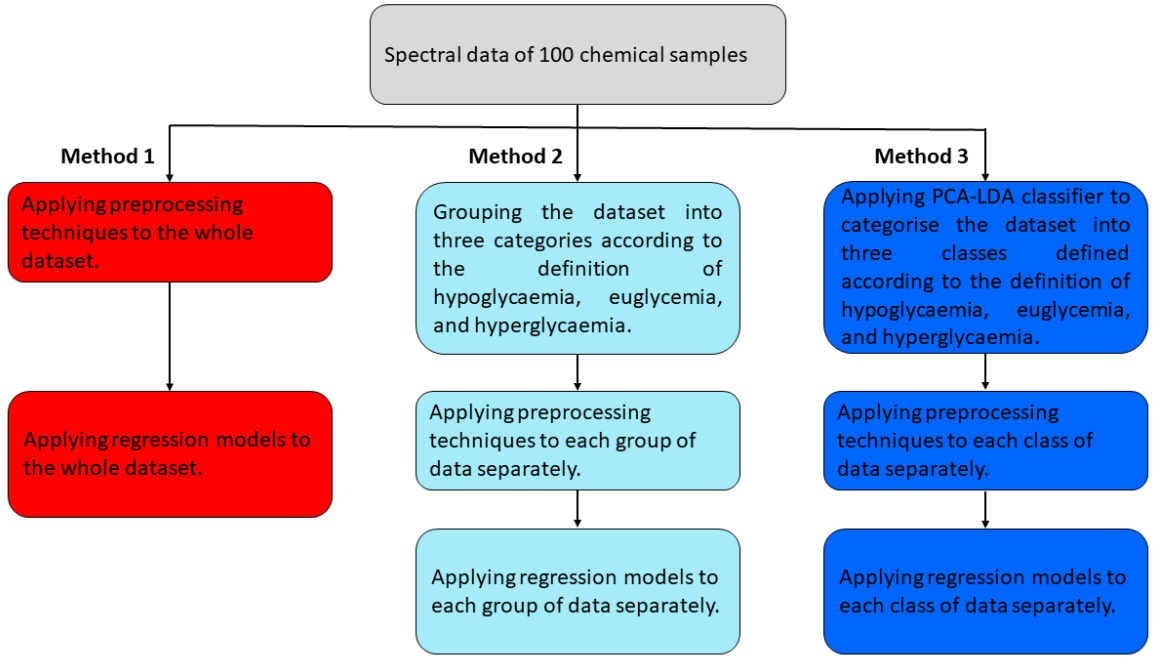

Figure 1. Quantification methods applied in this paper for glucose measurement using NIR, MIR, and NIR-MIR spectroscopy

3.1.1. Method 1

In this method, regression models were implemented for the raw and pre-processed spectra of the whole sample set to predict the relevant glucose concentrations.

3.1.2. Method 2

In this method, the spectral data were divided into the three groups shown in Table 1. Pre-processing and regression methods were then implemented for each group individually. By performing this method, the idea was to investigate the effect of categorising the dataset as being in the hypoglycaemic range (≤70mg/dl), euglycaemic range (70-180mg/dl), and hyperglycaemic range (180mg/dl≤) [19] on the accuracy of the measurements.

As shown later in the paper, Method 2 improved the measurement results as compared to Method 1. This improvement led us to proceed to Method 3, which as Method 1 used only the spectral data without further information (Method 2 used the labels in addition to the spectral data).

Table 1. Division of the dataset into three groups following the clinical definition of the glycaemic ranges

	Label		
	Class 1	Class 2	Class 3
Glucose concentration range	5–70 mg/dl	75–180 mg/dl	185–500 mg/dl
Corresponding glycaemic range	Hypoglycaemia	Euglycaemia	Hyperglycaemia
Quantity of data in the class	14	22	64

3.1.3. Method 3

This method had the same principles as Method 2 but automated the grouping process using a PCA-LDA classifier.

3.2. Pre-processing methods

In spectroscopic analyses, pre-processing techniques are generally applied to the raw spectra to curtail adverse effects from elements other than the analyte of interest and environmental conditions [20]. In this work, three pre-processing techniques are applied to assess the effectiveness of the proposed pre-classification approach when it is coupled with conventional pre-processing methods.

3.2.1. Smoothing (S)

Savitzky-Golay smoothing is a pre-processing method to diminish the effect of noise on the raw spectra [17]. The method is an averaging algorithm that fits a polynomial with successive subsets of adjacent data points based on the least-squares [21]. For the Savitzky-Golay filter in this work, a five-point window and a second-order polynomial were implemented.

3.2.2. Multivariate scatter correction (MSC)

The scattering phenomenon is the most significant obstacle when attempting quantitative measurements using NIR spectroscopy [22]. MSC is a useful pre-processing technique to eliminate the effect of light scattering [23]. In this method, the scattering of each spectrum is estimated relative to a reference spectrum [14]. In this work, the average of all signals in the calibration set was considered as the reference spectrum; each signal was then adjusted using the reference so that they all had the same scatter level [17].

3.2.3. Smoothing coupled with MSC (S-MSC)

Applying different pre-processing techniques together is a common approach to dealing with spectroscopic data [15]. In this paper, a combination of the smoothing and MSC methods was also applied as a third pre-processing technique [24].

3.3. Regression methods

For constructing predictive models for selected analytes such as glucose, multivariate calibration methods are applied to spectroscopic data [12], [25]. Linear models such as PLSR and PCR are generally preferred since they are easy to apply and amenable to Physicochemical interpretation [26]. Likewise, in this paper, PLSR and PCR are selected for glucose quantification using absorbance spectroscopy.

For optimising the PLS and PCR components quantity, different numbers ranging from 1 to 10 were examined; and each time, the sum of squares of differences between reference and predicted glucose levels, based on ten-fold cross-validation analysis, was calculated to form the predicted residual sum of squares (*PRESS*). The model minimising the value of *PRESS/(N-A-1)* is then selected; where *N* is the size of the calibration set and *A* is the number of components [27].

3.4. Classification method

Principal component analysis (PCA) reduces the dimensionality of data while retaining most of the information present in the dataset [28]. Linear discriminant analysis (LDA) is a technique that maximises the variance between groups while minimising the variance within groups based on the determination of linear discriminant functions [29]. In this work, PCA coupled with LDA (PCA-LDA) is employed to classify the dataset, a method which was shown to be useful in this regard [30]. Different values from 1 to 10 were considered as the number of PCA components and the model resulting in the best classification accuracy, based on ten-fold cross-validation results, was then selected [17].

3.5. Evaluation method

The maximisation of the training data size is a basic approach for dealing with small datasets, and cross-validation is suitable for this purpose [31]. Ten-folds cross-validation was applied in this work to evaluate the regression and classification models [32].

3.6. Evaluation Metrics

3.6.1. Root mean square error of prediction (RMSEP)

In this work, RMSE was calculated as follows to measure the actual error of quantifications [33].

$$RMSEP = \sqrt{(\sum_{i=1}^{N}(y_i - \hat{y}_i)^2)/N}$$

N : the size of the calibration set

y_i : reference value

$\hat{y}_i$: predicted value

3.6.2. Percentage error around the mean (PEM)

PEM was used to analyse the performance of the quantification methods for each class of data [17] (“quantification methods” and “classes” are discussed in section 3.2).

$$PEM = (RMSECV / \bar{Y}) \times 100$$

$\bar{Y}$: the average of reference values

3.6.3. Correlation coefficients (r)

r is a statistical measure indicating correlations between the reference and predicted glucose concentrations [34].

$$r = Cov\,(Y,\ \hat{Y}) / \sigma_Y \sigma_{\hat{Y}}$$

Y: referencevalues; $\hat{Y}$: predicted values; Cov (Y, $\hat{Y}$): covariance between Y and $\hat{Y}$; σ_Y: standard deviation of Y; $\sigma_{\hat{Y}}$: standard deviation of $\hat{Y}$

3.6.4. Clarke error grid analysis (EGA)

EGA considers the relative difference between reference and predicted glucose levels and the clinical significance of this difference [35]. In this paper, EGA was performed to assess the clinical accuracy of the measurements, a method which can be used to evaluate in vitro quantitative analysis of glucose[36].

4. Results

4.1. Classification results

It was mentioned earlier that the proper number of PCA components in the PCA-LDA classifier, implemented in the third quantification method, was chosen based on the examination of varying values. Classification results based on ten-fold cross-validation for a different number of PCA components ranging from 1 to 10 are illustrated in Figure 2. As the figure shows, in the NIR region, the classification accuracy improved significantly when the number of components rose from 1 to 5 but remained steady afterwards. Therefore, we set the number of PCA components at 5 for this region. Similarly, the number of PCA elements when using MIR and IR spectral data were both set at 4.

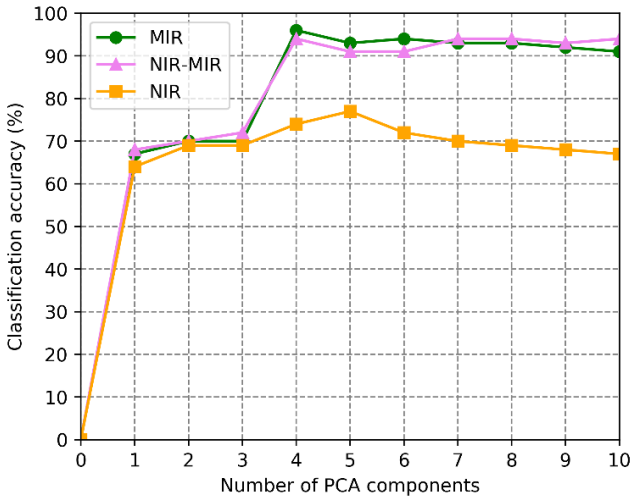

Figure 2. The Accuracy of the PCA-LDA classifier for different numbers of PCA components

The detailed classification results after setting the number of PCA components at the values mentioned above are presented in Table 2. Overall, the best and the lowest classification accuracy were found when using the MIR and NIR spectral data, respectively.

Table 2. The PCA-LDA classification results based on ten-fold cross-validation in different spectral regions

Spectral region	Class 1 data (Hypoglycaemia range)		Class 2 data (Euglycaemia range)		Class 3 data (Hyperglycaemia range)		All data together (whole glycaemic range)	
	No. of Errors	Classification Accuracy (%)	No. of Errors	Classification Accuracy (%)	No. of Errors	Classification Accuracy (%)	No. of Errors	Classification Accuracy (%)
NIR	3	78.5	15	31.8	5	92.1	23	77
MIR	0	100	4	81.8	0	100	4	96
NIR-MIR	0	100	4	81.8	2	96.8	6	94

4.2. Quantification results

This section is partitioned into three parts, each of which reports the quantification results belonging to the analyses in one of the spectral regions (NIR, MIR, and NIR-MIR region). In this way, the capability of the proposed method to improve the analysis precision in either of the three spectral regions could be shown more effectively.

4.2.1. Quantification results in the NIR region

Table 3 lists the results of RMSE, PEM and correlation coefficient (r) of the three quantification methods in the NIR region, and in addition, the improvement of RMSE for Methods 2 and 3 compared to Method 1. The values in bold indicate the best results of each quantification method based on the lowest

RMSE of the whole dataset; these results are considered for EGA analysis, as discussed later. For comparison purposes, the results of each class and also for that of all classes together are presented separately.

Methods 2 and 3 possessed smaller calibration sets than Method 1, a characteristic that might have harmed the results of these methods. However, they provided more accurate quantification results than Method 1. The accuracy of predictions obtained by Method 2 outweighed those of Method 3. The reason is that the weak classification accuracy in the NIR region (discussed in section 4.1) negatively affected the measurements of Method 3. For Method 2, the improvements in the quantification results in comparison to Method 1 were more pronounced for data in Classes 1 and 2; for Method 3, it happened for data in Class 1.

Table 3. Results of ten-fold cross-validation for the quantification methods in the NIR region

PM	RM	QM	Class 1 data (Hypoglycaemia range)			Class 2 data (Euglycaemia range)			Class 3 data (Hyperglycaemia range)			All data together (Whole glycaemic range)		
			RMSEP (mg/dl)	Im (%)	PEM (%)	RMSEP (mg/dl)	Im (%)	PEM (%)	RMSEP (mg/dl)	Im (%)	PEM (%)	RMSEP (mg/dl)	Im (%)	r
NP	PLSR	1	63.5	—	181.4	100.0	—	80.0	76.6	—	22.3	81.1	—	0.82
		2	11.8	+81.4	31.5	19.4	+80.6	15.2	53.2	+30.5	15.5	43.7	+46.1	0.95
		3	39.5	+37.7	113.4	97.0	+3.0	77.6	69.7	+9.0	20.3	74.0	+8.7	0.85
	PCR	1	64.3	—	183.9	98.4	—	78.7	77.7	—	22.6	81.4	—	0.82
		2	**12.4**	**+80.7**	**33.0**	**19.3**	**+80.3**	**15.1**	**53.0**	**+31.7**	**15.4**	**43.6**	**+46.4**	**0.95**
		3	42.4	+34.0	121.3	98.1	+0.3	78.5	69.2	+10.9	20.2	74.2	+8.8	0.85
S	PLSR	1	60.5	—	173.0	100.8	—	80.7	76.3	—	22.2	80.9	—	0.82
		2	11.9	+80.3	31.8	19.4	+80.7	15.2	53.0	+30.5	15.4	43.6	+46.1	0.95
		3	**39.1**	**+35.3**	**11.9**	**97.1**	**+3.6**	**77.6**	**68.8**	**+9.8**	**20.0**	**73.4**	**+9.2**	**0.83**
	PCR	**1**	**60.8**	**—**	**173.9**	**97.9**	**—**	**78.3**	**76.3**	**—**	**22.2**	**80.0**	**—**	**0.83**
		2	12.2	+79.9	32.6	19.3	+80.2	15.1	52.8	+30.7	15.4	43.4	45.7	0.95
		3	42.1	+30.7	120.4	98.1	-0.2	78.5	68.7	+9.9	20.6	73.9	+7.6	0.85
MSC	PLSR	1	63.5	—	181.4	100.0	—	80.0	76.6	—	22.3	81.1	—	0.82
		2	11.8	+81.4	31.5	19.4	+80.6	15.2	53.2	+30.5	15.5	43.7	+46.1	0.95
		3	39.5	+37.7	113.0	97.0	+3.0	77.6	69.7	+9.0	20.3	74.0	+8.7	0.85
	PCR	1	64.3	—	183.9	98.4	—	78.7	77.7	—	22.6	81.4	—	0.82
		2	12.4	+80.7	33.0	19.3	+80.3	15.1	53.0	+31.7	15.4	43.6	+46.4	0.95
		3	42.4	+34.0	121.3	98.1	+0.3	78.5	69.2	+10.9	20.3	74.2	+8.8	0.85
S-MSC	PLSR	1	175.0	—	500.0	136.2	—	109.0	95.4	—	27.8	118.6	—	0.57
		2	12.1	+93.0	32.3	32.6	+76.0	25.6	56.2	+41.0	16.4	47.7	+59.7	0.94
		3	79.3	+54.6	226.6	115.4	+15.2	92.3	71.0	+25.5	20.7	84.3	+28.9	0.81
	PCR	1	179.5	—	512.9	137.7	—	110.2	96.2	—	28.1	120.3	—	0.55
		2	11.3	+93.7	30.2	38.8	+71.8	30.4	55.7	+42.0	16.2	48.3	+59.8	0.94
		3	78.1	+56.4	223.3	114.8	+16.6	91.8	71.0	+26.1	20.7	83.9	+30.2	0.81

Abbreviations: PM = pre-processing method; RM = regression model; QM = quantification method; RMSEP = root mean square error of prediction; Im = improvement of RMSEP in comparison to that of Method 1; PEM = percentage error around the mean; r = correlation coefficient; NP = no pre-processing; S = smoothing; MSC = multivariate scatter correction, S-MSC = smoothing couples with multivariate scatter correction.

As mentioned earlier, EGA was performed to evaluate the accuracy of the measurements further. The EGA comparison between the best prediction results of the three quantification methods in the NIR region

(results in bold in Table 3) is shown in Figure 3(A); and the percentage of predictions located in Zone A—the most clinically desired measurement—is presented in Fig(B). As shown in the figures, using Methods 2 and 3, a higher ratio of predictions is located in zone A, especially for data in Classes 1 and 2.

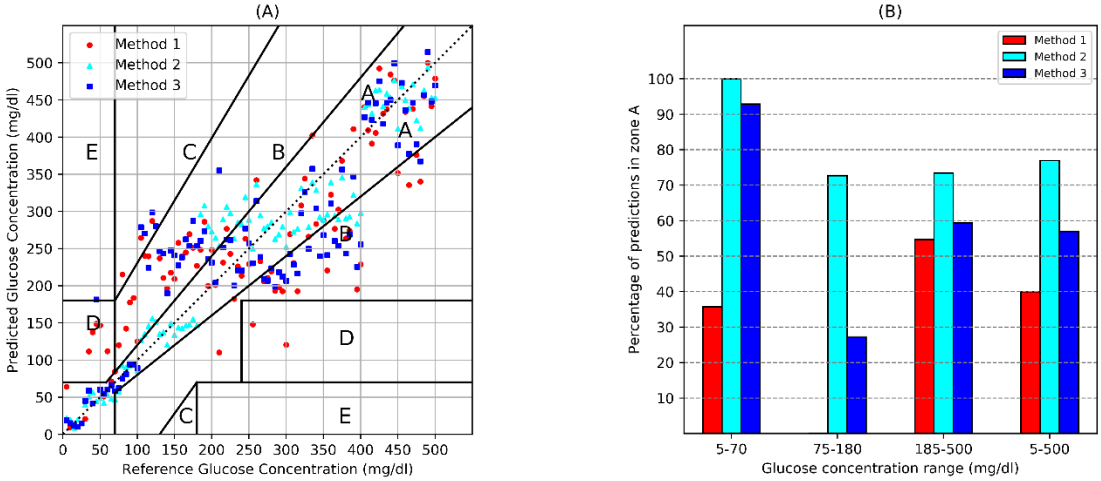

Figure 3. (A) EGA of the quantification methods in the NIR region (Some predictions of the first calibration method had negative values, so have not appeared in the graph.), (B) the statistics of the EGA graphQuantification results in the MIR region

4.2.2. Quantification results in the MIR region

Table 4 presents the quantification results of the three methods in the MIR region. Methods 2 and 3 were more accurate than Method 1, especially for data with lower glucose concentrations (Classes 1 and 2 data). The results of Methods 2 and 3 were comparable in the MIR region, which was due to the acceptable classification accuracy in this region.

EGA for the best result of the quantification methods in the MIR region and the percentage of measurements distributed in Zone A for each method are displayed in Figure 4. As the figures show, more accurate prediction results were obtained using Methods 2 and 3 rather than Method 1, notably for lower glucose levels.

Table 4. Results of ten-fold cross-validation for the quantification methods in the MIR region

PM	RM	QM	Class 1 data (Hypoglycaemia range)			Class 2 data (Euglycaemia range)			Class 3 data (Hyperglycaemia range)			All data together (Whole glycaemic range)		
			RMSEP (mg/dl)	Im (%)	PEM (%)	RMSEP (mg/dl)	Im (%)	PEM (%)	RMSEP (mg/dl)	Im (%)	PEM (%)	RMSEP (mg/dl)	Im (%)	r
NP	PLSR	1	73.8	—	211.0	38.2	—	30.5	32.5	—	9.5	41.5	—	0.95
		2	17.0	+76.9	45.4	14.4	+62.3	11.3	31.5	+3.0	9.2	26.6	+35.9	0.98
		3	15.0	79.6	42.9	23.3	+39.0	18.6	31.2	+4.0	9.1	27.9	+32.7	0.98
	PCR	1	81.1	—	231.9	33.2	—	26.6	32.3	—	9.4	42.1	—	0.95
		2	7.1	+91.2	19.1	13.1	+60.5	10.3	27.2	+18.3	7.9	22.8	+45.8	0.98
		3	16.2	+80.0	46.3	21.4	+35.5	17.19	28.5	+14.4	8.3	25.7	+38.9	0.98
S	PLSR	1	73.9	—	211.2	37.9	—	30.3	32.5	—	9.5	41.4	—	0.95
		2	17.0	+76.9	45.4	14.3	+62.2	11.2	31.4	+3.3	9.1	26.8	+35.2	0.98
		3	15.0	+79.7	42.9	23.2	+38.7	18.5	31.2	+4.0	9.1	27.9	+32.6	0.98
	PCR	1	81.2	—	232.1	33.2	—	26.6	32.3	—	9.4	42.2	—	0.95
		2	7.1	+91.2	19.0	13.1	+60.5	10.2	27.2	+15.7	7.9	22.8	+45.9	0.98
		3	16.2	+80.0	46.3	21.4	+35.5	17.1	28.5	+11.7	8.3	25.7	+39.0	0.98
MSC	PLSR	1	73.8	—	211.0	38.2	—	30.5	32.5	—	9.5	41.5	—	0.95
		2	17.0	+76.9	45.4	14.4	+62.3	11.3	31.5	+3.0	9.2	26.9	+35.1	0.98
		3	15.0	+79.6	42.9	23.2	+39.2	18.6	31.2	+4.0	9.1	27.9	+32.7	0.98
	PCR	1	81.1	—	231.9	33.2	—	26.6	32.3	—	9.4	42.1	—	0.95
		2	**7.1**	**+91.2**	**19.1**	**13.1**	**+60.5**	**10.3**	**27.2**	**+15.7**	**7.9**	**22.8**	**+45.8**	**0.98**
		3	16.2	+80.0	46.3	21.4	+35.5	17.1	28.5	+11.7	8.3	25.7	+38.9	0.98
S-MSC	PLSR	**1**	**65.9**	**—**	**188.3**	**37.2**	**—**	**29.8**	**32.5**	**—**	**9.4**	**39.5**	**—**	**0.96**
		2	9.3	+85.8	24.9	11.9	+68.0	9.4	27.8	+14.4	8.1	23.2	+41.2	0.98
		3	**10.2**	**+84.5**	**29.2**	**17.9**	**+51.8**	**14.3**	**26.6**	**+18.1**	**7.7**	**23.2**	**+41.2**	**0.98**
	PCR	1	77.0	—	220.0	36.3	—	29.0	30.6	—	8.9	40.9	—	0.95
		2	8.2	+89.3	21.9	9.5	+73.8	7.4	28.7	+6.2	8.3	23.6	+42.2	0.98
		3	13.1	+82.9	37.5	16.8	+53.7	13.4	26.8	+12.4	7.8	23.4	+42.7	0.98

Abbreviations: PM = pre-processing method; RM = regression model; QM = quantification method; RMSEP = root mean square error of prediction; Im = improvement of RMSEP in comparison to that of Method 1; PEM = percentage error around the mean; r = correlation coefficient; NP = no pre-processing; S = smoothing; MSC = multivariate scatter correction, S-MSC = smoothing couples with multivariate scatter correction.

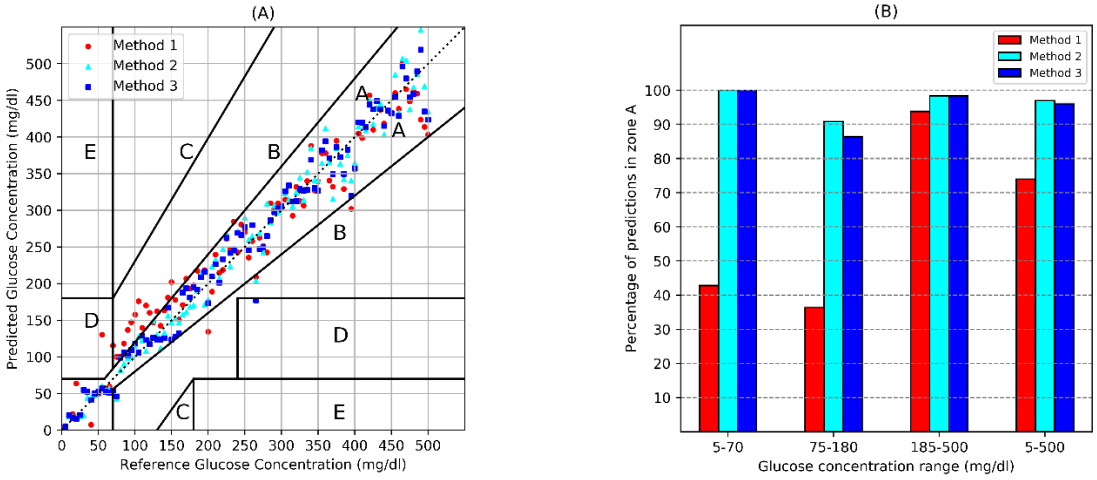

Figure 4. (A) EGA of the quantification methods in the MIR region, (B) the statistics of the EGA graphQuantification results in the NIR-MIR region

4.2.3. Quantification results in the NIR-MIR region

Table 5 reports the prediction results in the NIR-MIR region for all the quantification methods. Overall, in all cases, Methods 2 and 3 provided more accurate prediction results. All results obtained in the NIR-MIR region were generally comparable to those of the MIR region, which indicates that our proposed classification-before-regression methodology, still maintains its effectiveness over a wider range of spectra.

Table 5. Results of ten-fold cross-validation for the quantification methods in the NIR-MIR region

PM	RM	QM	Class 1 data (Hypoglycaemia range)			Class 2 data (Euglycaemia range)			Class 3 data (Hyperglycaemia range)			All data together (Whole glycaemic range)		
			RMSEP (mg/dl)	Im (%)	PEM (%)	RMSEP (mg/dl)	Im (%)	PEM (%)	RMSEP (mg/dl)	Im (%)	PEM (%)	RMSEP (mg/dl)	Im (%)	r
NP	PLSR	1	67.5	—	193.1	40.9	—	32.7	32.0	—	9.3	40.4	—	0.95
		2	18.0	+73.3	48.1	13.2	+67.7	10.3	28.3	+11.5	8.2	24.4	+39.6	0.98
		3	16.5	+75.5	47.2	24.6	+39.8	19.7	28.2	+11.8	8.2	26.2	+35.1	0.98
	PCR	1	82.1	—	234.8	37.6	—	30.0	33.1	—	9.6	43.6	—	0.95
		2	8.1	+90.1	21.6	12.3	+67.2	9.7	26.2	+20.8	7.6	21.9	+49.7	0.98
		3	14.1	+82.8	40.5	14.7	+60.9	11.7	24.9	+24.7	7.2	21.7	+50.2	0.98
S	PLSR	1	67.6	—	193.2	41.0	—	32.8	32.1	—	9.3	40.5	—	0.95
		2	18.0	+73.3	48.1	13.0	+68.2	10.2	28.2	+12.1	8.2	24.3	+40.0	0.98
		3	16.5	+75.5	47.3	24.5	+40.2	19.6	28.1	+12.4	8.2	26.1	+35.5	0.98
	PCR	1	82.0	—	234.3	37.6	—	30.0	33.1	—	9.6	43.6	—	0.95
		2	8.0	+90.2	21.3	12.3	+67.2	9.6	26.2	+20.8	7.6	21.9	+49.7	0.98
		3	**14.2**	**+82.6**	**40.5**	**14.6**	**+61.1**	**11.6**	**24.9**	**+24.7**	**7.2**	**21.7**	**+50.2**	**0.98**
MSC	PLSR	**1**	**67.5**	**—**	**193.1**	**40.9**	**—**	**32.7**	**32.0**	**—**	**9.3**	**40.4**	**—**	**0.95**
		2	18.0	+73.3	48.1	13.2	+67.7	10.3	28.3	+11.5	8.2	24.4	+39.6	0.98
		3	16.5	+75.5	47.2	24.6	+39.8	19.7	28.2	+11.5	8.2	26.2	+35.1	0.98
	PCR	1	82.1	—	234.8	37.6	—	30.0	33.1	—	9.6	43.6	—	0.95
		2	8.1	+90.1	21.6	12.3	+67.2	9.7	26.2	+20.8	7.6	21.9	+49.7	0.98
		3	14.1	+82.8	40.5	14.7	+60.9	11.7	24.9	+24.7	7.2	21.7	+50.2	0.98
S-MSC	PLSR	1	75.6	—	216.0	37.1	—	29.7	32.8	—	9.5	41.8	—	0.95
		2	**9.7**	**+87.1**	**26.0**	**11.5**	**+69.0**	**9.0**	**26.0**	**+20.7**	**7.6**	**21.8**	**+47.8**	**0.98**
		3	10.6	+85.9	30.4	16.5	+55.5	13.2	26.9	+19.9	7.8	23.28	+44.3	0.98
	PCR	1	76.6	—	218.8	37.8	—	30.2	34.6	—	10.1	43.1	—	0.95
		2	8.5	+88.9	22.8	9.1	+75.9	7.1	26.4	+28.9	7.7	21.8	+49.4	0.98
		3	16.0	+79.1	45.8	13.2	+65.0	70.6	26.4	+28.9	7.7	22.8	+47.0	0.98

Abbreviations: PM = pre-processing method; RM = regression model; QM = quantification method; RMSEP = root mean square error of prediction; Im = improvement of RMSEP in comparison to that of Method 1; PEM = percentage error around the mean; r = correlation coefficient; NP = no pre-processing; S = smoothing; MSC = multivariate scatter correction, S-MSC = smoothing couples with multivariate scatter correction.

The EGA graph for the best result of the quantification methods in the NIR-MIR region, and a comparison of the predictions that occurred in Zone A for each method are presented in Figure 5. Based on the figure, it is clear that Methods 2 and 3 are more accurate than Method 1, also in the NIR-MIR region.

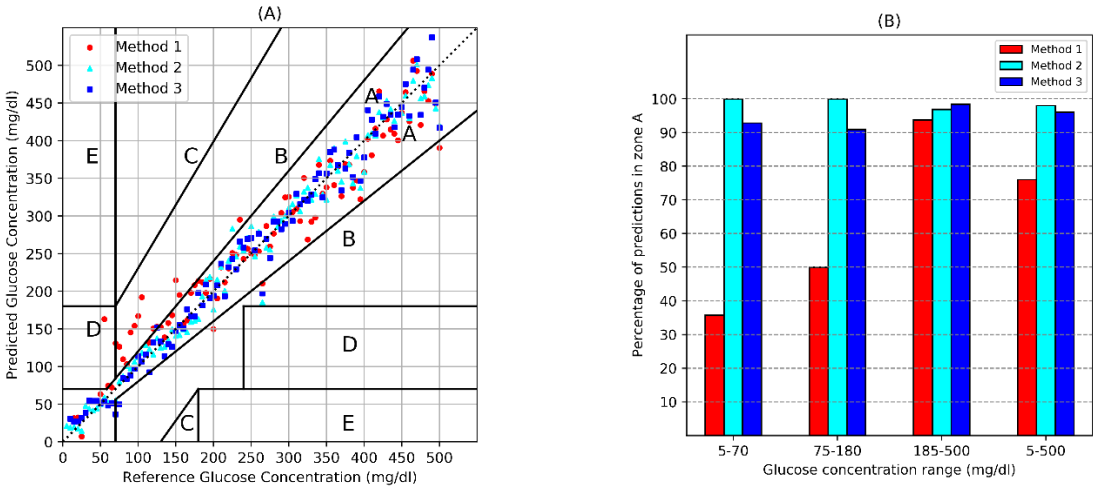

Figure 5. (A) EGA of the quantification methods in the NIR_MIR region, (B) the statistics of the EGA graphDiscussion

5. Discussion

All the quantification methods, as well as the PCA-LDA classifier applied in the third quantification method, showed a significantly better performance in the MIR and NIR-MIR regions than in the NIR region. The reason is that the MIR and NIR-MIR signals contained a more extensive range of wavelengths; these can possess more informative wavebands for glucose monitoring compared to the NIR spectra in the experiment.

Methods 2 and 3 resulted in more accurate measurements than Method 1. Riley et al. showed that informative wavebands for the quantitative analysis of four chemical components, including glucose, become narrower with a decrease in the concentration range of these analytes [37]. It can be inferred that, in our dataset, informative wavebands for glucose measurement are possibly homogenous for signals in each class, and are different from the optimal spectral range of data in other classes. These similarities between spectra in each calibration set, using Methods 2 and 3, could improve the accuracy of the regression analyses.

As shown in Table, in our dataset, the 3 classes occupy different glycemic ranges dominated by samples within the hyperglycaemia range. However, datasets from in vivo experiments on humans generally tend

to have the majority of data in the euglycaemia range; to show that our proposed methodology still applies in this case, the same analysis is repeated after adapting the step size in the hyperglycaemia range to allow for the majority of the data to be placed in the euglycaemia range. The step size for the hyperglycaemia range was increased from 5mg/dl to 20mg/dl; this lowered the number of samples by three-quarters for this class as illustrated in Table 6, and the other two classes remained unchanged. The new corresponding analysis results are shown in Table 7, which confirms the effectiveness of the pre-classification methodology for this distribution too.

Table 6. Division of the dataset after the step size adaptation

	Label		
	Class 1	Class 2	Class 3
Glucose concentration range	5–70 mg/dl	75–180 mg/dl	185–500 mg/dl
Corresponding glycaemic range	Hypoglycaemia	Euglycaemia	Hyperglycaemia
Quantity of data in the class	14	22	16

Table 7. Best results of ten-fold cross-validation for the quantification methods in all spectral region for modified data distribution

SR	QM	Class 1 data (Hypoglycaemia range)			Class 2 data (Euglycaemia range)			Class 3 data (Hyperglycaemia range)			All data together (whole glycaemic range)		
		RMSEP (mg/dl)	Im (%)	PEM (%)	RMSEP (mg/dl)	Im (%)	PEM (%)	RMSEP (mg/dl)	Im (%)	PEM (%)	RMSEP (mg/dl)	Im (%)	r
NIR	1	115.1	—	328.9	72.2	—	57.8	164.8	—	49.1	118.2	—	0.42
	2	12.4	+89.2	33.0	19.3	+73.2	15.1	60.6	+63.2	18.1	29.9	+74.7	0.94
	3	54.1	+52.9	154.6	76.1	-5.1	60.9	129.8	+21.2	38.7	92.1	+4.1	0.72
MIR	1	78.1	—	223.1	34.0	—	27.2	50.8	—	14.7	53.1	—	0.92
	2	7.1	+90.0	19.1	13.1	+61.4	10.3	49.5	+2.5	14.3	22.6	+57.4	0.97
	3	11.0	+85.9	31.7	20.6	+39.4	16.5	74.3	-46.2	21.5	44.6	+16.0	0.94
NIR-MIR	1	62.8	—	179.5	46.7	—	37.4	51.1	—	14.5	52.1	—	0.92
	2	9.7	+84.5	26.0	11.5	+75.3	9.0	50.8	+0.5	15.0	23.1	+55.6	0.97
	3	15.8	+74.8	45.1	20.2	+56.7	16.2	53.6	-4.8	15.5	34.1	+34.5	0.96

Abbreviations: SR = spectral region QM = quantification method; RMSEP = root mean square error of prediction; Im = improvement of RMSEP in comparison to that of Method 1; PEM = percentage error around the mean; r = correlation coefficient

6. Conclusion

Glucose measurement using NIR and MIR absorbance spectroscopy was improved by manually grouping the dataset into three categories according to the clinical definition of the glycaemic ranges and then applying regressions for each class separately. A PCA-LDA classifier was therefore implemented to assign each spectrum to the respective class automatically. The creation of regression models for different classes improved the results of glucose prediction as compared to regressions for the whole dataset. The improvements in the prediction results were more significant for lower glucose concentrations.

The performance of the proposed pre-classification approach was evaluated for two common regression methods, three pre-processing techniques, and also for a broader range of spectra by merging the NIR and MIR data. A primary evaluation of the proposed methodology was carried out by repeating the analysis for a modified version of the dataset to account for the distribution of data that is more representative of human in vivo experiments scenarios. For future work, the determination of informative wavebands for glucose measurement could be investigated in each glycaemic range individually.

Benaissa: conceptualisation, methodology, validation, investigation, resources, review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] N. S. Oliver, C. Toumazou, A. E. G. Cass, and D. G. Johnston, 'Glucose sensors: A review of current and emerging technology', *Diabet. Med.*, vol. 26, no. 3, pp. 197–210, 2009.

[2] T. Vahlsing, G. Steiner, H. M. Heise, S. Delbeck, and S. Leonhardt, 'Non-invasive monitoring of blood glucose using optical methods for skin spectroscopy—opportunities and recent advances', *Anal. Bioanal. Chem.*, vol. 411, no. 1, pp. 63–77, 2018.

[3] J. Chung, H. So, Choi, and T. K. S. Wong, 'Recent advances in noninvasive glucose monitoring', *Med. Devices Evid. Res.*, p. 45, 2012.

[4] A. Al-Mbaideen and M. Benaissa, 'Coupling subband decomposition and independent component regression for quantitative NIR spectroscopy', *Chemom. Intell. Lab. Syst.*, vol. 108, no. 2, pp. 112–122, 2011.

[5] J. Haas and B. Mizaikoff, 'Advances in Mid-Infrared Spectroscopy for Chemical Analysis', *Annu. Rev. Anal. Chem.*, vol. 9, no. 1, pp. 45–68, 2016.

[6] S. K. Vashist, 'Non-invasive glucose monitoring technology in diabetes management: A review', *Anal. Chim. Acta*, vol. 750, pp. 16–27, 2012.

[7] D. A. Burns and E. W. Ciurczak, *Handbook of near-infrared analysis*. CRC press, 2007.

[8] H. von Lilienfeld-Toal, M. Weidenmüller, A. Xhelaj, and W. Mäntele, 'A novel approach to non-invasive glucose measurement by mid-infrared spectroscopy: The combination of quantum cascade

lasers (QCL) and photoacoustic detection', *Vib. Spectrosc.*, vol. 38, no. 1–2, pp. 209–215, 2005.

[9] C.-F. So, K.-S. Choi, T. K. S. Wong, and J. W. Y. Chung, 'Recent advances in noninvasive glucose monitoring', *Med. Devices (Auckland, NZ)*, vol. 5, p. 45, 2012.

[10] B. Rabinovitch, W. F. March, and R. L. Adams, 'Noninvasive glucose monitoring of the aqueous humor of the eye: Part I. Measurement of very small optical rotations', *Diabetes Care*, vol. 5, no. 3, pp. 254–258, 1982.

[11] J. Yadav, A. Rani, V. Singh, and B. M. Murari, 'Prospects and limitations of non-invasive blood glucose monitoring using near-infrared spectroscopy', *Biomed. Signal Process. Control*, vol. 18, pp. 214–227, 2015.

[12] A. Tura, A. Maran, and G. Pacini, 'Non-invasive glucose monitoring: Assessment of technologies and devices according to quantitative criteria', *Diabetes Res. Clin. Pract.*, vol. 77, no. 1, pp. 16–40, 2007.

[13] J. Tenhunen, H. Kopola, and R. Myllylä, 'Non-invasive glucose measurement based on selective near infrared absorption; requirements on instrumentation and spectral range', *Meas. J. Int. Meas. Confed.*, vol. 24, no. 3, pp. 173–177, 1998.

[14] Å. Rinnan, F. van den Berg, and S. B. Engelsen, 'Review of the most common pre-processing techniques for near-infrared spectra', *TrAC - Trends Anal. Chem.*, vol. 28, no. 10, pp. 1201–1222, 2009.

[15] K. C. Patchava, O. Alrezj, M. Benaissa, and H. Behairy, 'Savitzky-golay coupled with digital bandpass filtering as a pre-processing technique in the quantitative analysis of glucose from near infrared spectra', *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. EMBS*, vol. 2016-Octob, pp. 6210–6213, 2016.

[16] J. M. Andrés and M. T. Bona, 'ASTM clustering for improving coal analysis by near-infrared spectroscopy', *Talanta*, vol. 70, no. 4, pp. 711–719, 2006.

[17] Y. Wang, M. Yang, G. Wei, R. Hu, Z. Luo, and G. Li, 'Improved PLS regression based on SVM classification for rapid analysis of coal properties by near-infrared reflectance spectroscopy', *Sensors Actuators, B Chem.*, vol. 193, pp. 723–729, 2014.

[18] H. Chen, Z. Liu, J. Gu, W. Ai, J. Wen, and K. Cai, 'Quantitative analysis of soil nutrition based on FT-NIR spectroscopy integrated with BP neural deep learning', *Anal. Methods*, vol. 10, no. 41, pp. 5004–5013, 2018.

[19] J. Kropff *et al.*, 'Accuracy of two continuous glucose monitoring systems: To-Head Comparison Under Clinical Research Centre and Daily Life', *Diabetes, Obes. Metab.*, vol. 2015, no. 17, pp. 343–349, 2015.

[20] L. C. Lee, C. Y. Liong, and A. A. Jemain, 'A contemporary review on Data Preprocessing (DP) practice

strategy in ATR-FTIR spectrum', *Chemom. Intell. Lab. Syst.*, vol. 163, no. December 2016, pp. 64–75, 2017.

[21] A. Savitzky and M. J. E. Golay, 'Smoothing and Differentiation of Data by Simplified Least Squares Procedures', *Anal. Chem.*, vol. 36, no. 8, pp. 1627–1639, 1964.

[22] J. Yadav, A. Rani, V. Singh, and B. M. Murari, 'Prospects and limitations of non-invasive blood glucose monitoring using near-infrared spectroscopy', *Biomed. Signal Process. Control*, vol. 18, pp. 214–227, 2015.

[23] Y. Mou, X. You, D. Xu, L. Zhou, W. Zeng, and S. Yu, 'Regularized multivariate scatter correction', *Chemom. Intell. Lab. Syst.*, vol. 132, pp. 168–174, 2014.

[24] D. Wu, J. Chen, B. Lu, L. Xiong, Y. He, and Y. Zhang, 'Application of near infrared spectroscopy for the rapid determination of antioxidant activity of bamboo leaf extract', *Food Chem.*, vol. 135, no. 4, pp. 2147–2156, Dec. 2012.

[25] Ian T. Jolliffe, 'A Note on the Use of Principal Components in Regression', *J. R. Stat. Soc. Ser. C (Applied Stat.*, vol. 31, no. 3, pp. 300–303, 1982.

[26] G. M. Escandar, P. C. Damiani, H. C. Goicoechea, and A. C. Olivieri, 'A review of multivariate calibration methods applied to biomedical analysis', *Microchem. J.*, vol. 82, no. 1, pp. 29–42, 2006.

[27] and L. E. Wold, Svante, Michael Sjöström, 'PLS-regression: a basic tool of chemometrics.', *Chemom. Intell. Lab. Syst.*, vol. 58, no. 2, pp. 109–130, 2001.

[28] L. A. Berrueta, R. M. Alonso-Salces, and K. Héberger, 'Supervised pattern recognition in food analysis', *J. Chromatogr. A*, vol. 1158, no. 1–2, pp. 196–214, 2007.

[29] S. A. Drivelos and C. A. Georgiou, 'Multi-element and multi-isotope-ratio analysis to determine the geographical origin of foods in the European Union', *TrAC - Trends Anal. Chem.*, vol. 40, pp. 38–51, 2012.

[30] J. Liu and S. Chen., 'Resampling LDA/QR and PCA+ LDA for face recognition', *Australas. Jt. Conf. Artif. Intell. Springer, Berlin, Heidelb.*, pp. 1221–1224, 2005.

[31] A. Pasini, 'Artificial neural networks for small dataset analysis', *J. Thorac. Dis.*, vol. 7, no. 5, pp. 953–960, 2015.

[32] J. Shao, 'Linear model selection by cross-validation', *J. Am. Stat. Assoc.*, vol. 88(422), pp. 486–492, 1993.

[33] M. L. F. Simeone, R. A. C. Parrella, R. E. Schaffert, C. M. B. Damasceno, M. C. B. Leal, and C. Pasquini, 'Near infrared spectroscopy determination of sucrose, glucose and fructose in sweet sorghum juice', *Microchem. J.*, vol. 134, pp. 125–130, 2017.

[34] and W. A. N. Lee Rodgers, Joseph, 'Thirteen ways to look at the correlation coefficient.', *Am. Stat.*,

vol. 42, no. 1, pp. 59–66, 1988.

[35] W. L. Clarke, 'The original Clarke error grid analysis (EGA)', *Diabetes Technol. Ther.*, vol. 7, no. 5, pp. 776–779, 2005.

[36] A. Al-Mbaideen and M. Benaissa, 'Frequency self deconvolution in the quantitative analysis of near infrared spectra', *Anal. Chim. Acta*, vol. 705, no. 1–2, pp. 135–147, 2011.

[37] M. R. Riley and H. M. Crider, 'The effect of analyte concentration range on measurement errors obtained by NIR spectroscopy', *Talanta*, vol. 52, no. 3, pp. 473–484, 2000.

Publication 4.

Signal fragmentation based feature vector generation in a model agnostic framework with application to glucose quantification using absorption spectroscopy[4]

Abstract. This paper proposes feature vector generation based on signal fragmentation equipped with a model interpretation module to enhance glucose quantification from absorption spectroscopy signals. For this purpose, near-infrared (NIR) and mid-infrared (MIR) spectra collected from experimental samples of varying glucose concentrations are scrutinised. Initially, a given spectrum is optimally dissected into several fragments. A base-learner then studies the obtained fragments individually to estimate the reference glucose concentration from each fragment. Subsequently, the resultant estimates from all fragments are stacked, forming a feature vector for the original spectrum. Afterwards, a meta-learner studies the generated feature vector to yield a final estimation of the reference glucose concentration pertaining to the entire original spectrum. The reliability of the proposed approach is reviewed under a set of circumstances encompassing modelling upon NIR or MIR signals alone and combinations of NIR and MIR signals at different fusion levels. In addition, the compatibility of the proposed approach with an underlying preprocessing technique in spectroscopy is assessed. The results substantiate the utility of incorporating the designed feature vector generator into standard benchmarked modelling procedures under all considered scenarios. Finally, to promote the transparency and adoption of the propositions, SHapley additive exPlanations (SHAP) is leveraged to interpret the quantification outcomes.

Keywords. Glucose quantification; Near-infrared spectroscopy; Mid-infrared spectroscopy; Machine learning, SHAP

---

[4] This article was published in Talanta. 243 (2022). Authors: **H. Khadem**, H. Nemat, J. Elliott, M. Benaissa.

1. Introduction

In vitro glucose quantification has practical applications in a variety of areas, e.g., food science, biology, and botany [1–4] Consequently, continued research is underway to expand this area of knowledge [5].

In this context, two optical modalities of near-infrared (NIR) and mid-infrared (MIR) have been broadly pursued in glucose quantification studies [6,7]. NIR and MIR signals are within the wavelength range of 750–2500 nm and 2500–10000 nm, respectively [8]. One advantage of using these technologies for glucose sensing is that the absence of reagents makes them economically appropriate for regular measurements [9].

NIR light possesses a high penetration rate enabling it to enter deeper parts of opaque compounds to seek the glucose trace [10–12]. On the other hand, the MIR region includes sharp peaks of glucose [13]. Of other merits of MIR spectroscopy for glucose sensing are the attenuated scattering phenomena and intensified absorption due to longer wavelengths [14]. Hence, there are stimuli to investigate glucose quantification from the combination of NIR and MIR spectra, as well.

As NIR/MIR light is traversing through an object and as a result of the interaction with physiological compounds of the object, some beam frequencies get scattered and absorbed [15,16]. These absorption and scattering patterns could be scrutinised using appropriate tools to derive information concerning the analyte(s) of interest [8]. Specifically, machine learning (ML) multivariate calibration algorithms, in particular, partial least squares regression (PLSR), are typically suggested for quantifying glucose from recorded NIR/MIR spectra [17,18].

Notwithstanding the general suitability of such algorithms, further advancements in the analysis are necessary towards achieving decisive glucose quantifications from NIR/MIR spectra [19,20]. In this regard, scopes exist to enhance the accuracy of the analysis by exploiting state-of-the-art ML techniques such as stack learning. Stack learning is an ensemble method for improving the competence of ML models in which a meta-learner integrates the outputs of multiple base-learner to produce a final output [21].

In conjunction with algorithms like stack learning, model interpretation frameworks could also be incorporated to expand the clarity of the analysis and further support the findings [22,23]. In this respect,

SHapley additive exPlanations (SHAP) is an elaborate game-theoretic model agnostic approach for interpreting ML models by attributing the contribution of each feature to a particular prediction [24]. SHAP joins optimal credit allocation with local explanations via the concept of Shapley values from cooperative game theory [25]. Resultant SHAP values designate the contribution of attributes to deviations from average estimations, a measurement to elucidate the effect of individual features on models' outputs [24].

This article suggests signal fragmentation based feature vector generation (SFFVG) dressed with model interpretations for in vitro glucose estimation upon absorbance spectroscopy data. First, a given signal was efficiently segmented into a number of sub-signals. The sub-signals were then autonomously investigated using a base-learner to estimate the reference glucose concentration. These fragmentary estimations were thereafter concatenated, forming a feature vector for the given signal. A meta-learner, utilising the concept of stack learning, later aggregated the generated feature vector's elements, creating an estimation related to the entire signal. The flexibility of the proposed approach was monitored by implementing it on NIR signals, MIR signals, and the fusion of NIR and MIR signals. Furthermore, the compatibility of the method with a conventional preprocessing technique in spectroscopy was examined. Finally, to spur the adoption of the propositions by increasing the clarity of the analysis, SHAP was carried out to delineate the influence of constructed features on the formation of final estimations.

2. Material and methods

SFFVG was proposed to advance glucose quantification from these absorption spectroscopic data. The effectiveness of the proposed method was examined within six different modelling strategies with and without including a classical preprocessing technique. Finally, to extend the transparency of the proposed method, SHAP was deployed to interpret the created models. The dataset and details of implementation steps are described in this section.

2.1. Dataset

For glucose quantification, this research investigated absorbance spectroscopic data collected from a set of 100 synthesised blood samples composed of phosphate (0.01 M/dl), human serum albumin (5 g/dl) and glucose. The samples covered a range of glucose concentrations from 5mg/dl to 500mg/dl, with intervals of 5mg/dl. The recorded spectra spanned a wavelength range of 2100–8000nm, with a precision of 1.7nm. Notably, the wavelengths from 2100 to 2500nm were classified as part of the NIR region, while those ranging from 2500 to 8000nm belonged to the MIR region. [26]

2.2. Calibration-validation split

For creating quantitative models, 80% of the data points were randomly selected and allocated as the calibration set, and the remaining 20% were considered as the validation set. Table 1summarises some statistical characteristics of the calibration and validation set. All subsequent model training and hyperparameter tuning operations were carried out using only the calibration set, whereas the validation set remained unseen for evaluation and model interpretation analysis.

Table 1. Characteristics of the calibration and validation set.

	Samples	Mean (mg dL-1)	Standard Deviation (mg dL-1)
Calibration set	80	250.3	146.4
Validation set	20	261.2	135.2

2.3. Feature vector generation

Figure 1 depicts the block diagram of SFFVG consisting of a signal fragmentation, regression, and concatenation unit. In the first step, the fragmentation unit efficiently breaks signals into several intervals. After that, the regression unit studies the obtained fragments independently to produce a corresponding fragmentary estimation of the reference glucose concentration. It should be noted that this regression block is trained separately for each interval using the corresponding fragments from the calibration set. Finally, the concatenation unit stacks the outputs of the regression unit, forming a feature vector for the original input signal.

The fragmentation unit was optimised for three separate scenarios depending on input data: NIR signals, MIR signals, or concatenation of NIR and MIR signals (hereafter referred to as NIR-MIR signals). For simplicity, equidistant fragmentation was considered, and signals were inputted in raw form. Values of 1 to 20 were explored as the number of intervals, and the one resulting in estimations (by the regression unit) with the lowest root mean square error (*RMSE*) of five-fold cross-validation on the calibration set was selected.

For the regressors block, PLSR was assigned, which previously has been demonstrated to be an excellent method in spectroscopic data analysis [27,28]. For tunning the number of PLSR components, values of 1 to max (10, the length of the input variable) were sought, and the one delivering the minimum *RMSE* of glucose quantification based on five-fold cross-validation on the calibration set was decided.

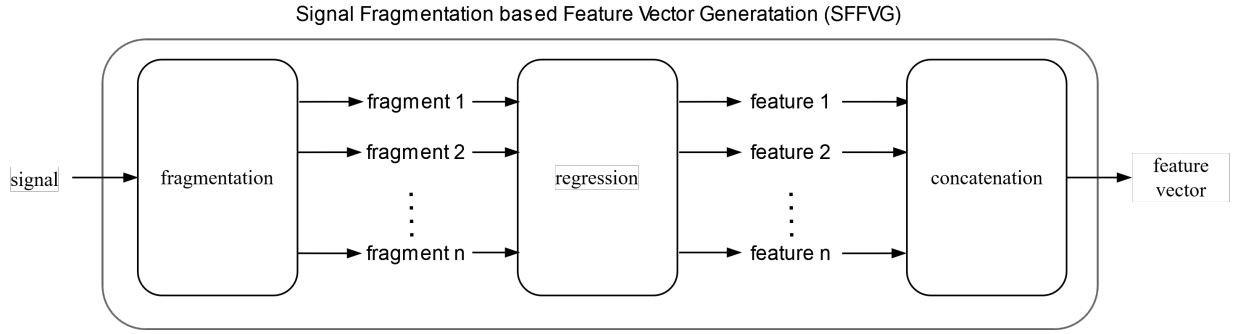

Figure 1. The general scheme of the proposed signal fragmentation based feature vector generation (SFFVG) method consists of signal fragmentation, regression, and concatenation. The input spectrum is optimally divided into a number of fragments. Next, each fragment is used as the input of a regressor (partial least square regression) to estimate the glucose concentration. Outputs of regressors were then stacked according to the order of the relevant fragments to form a generated feature vector.

2.4. Chemometric

In this work, for creating glucose quantification models, we assigned six different modelling strategies with the general block diagrams exhibited in Figure 2 [29–32]. As can be observed, SFFVG was a building block of all considered strategies.

For preprocessing units in the structure of the strategies shown in the figure, Savitzky-Golay (SG) smoothing filter was considered with a second-order polynomial and a five-point window [33,34]. Preprocessing analysis was to examine the compatibility of SFFVG with this prominent stage in

spectroscopy. For regression units, PLSR was appointed with the same tuning process described in subsection 2.2.

It is distinguishable from the block diagram that the first two modelling strategies were unimodal, where only NIR or MIR spectra took part in the modelling process. In contrast, the other four strategies were bimodal, utilising both NIR and MIR signals. Also, each dashed block in Figure 2 signifies two possible model creation scenarios for the associated modelling strategy by incorporating or not incorporating that particular block. Therefore, the working mechanism of modelling strategies was as follows.

a) NIR Modelling (Figure 2a): the raw or preprocessed form of a given NIR signal or their feature vector were input to a regression unit for making a final glucose estimation.
b) MIR Modelling (Figure 2b): this strategy was akin to NIR Modelling, except MIR signals were studied instead of NIR.
c) Raw Spectra Fusion Modelling (Figure 2c): a given NIR and MIR signal was first concatenated, forming an NIR-MIR signal. Raw or preprocessed form of the NIR-MIR signal or their feature vector was then fed to a regressor, creating a final estimation.
d) Preprocessed Spectra Fusion Modelling (Figure 2d): initially, a given NIR and MIR signal were separately preprocessed and then mixed. The resultant NIR-MIR signal or its feature vector were then given to a regressor to accomplish a final estimation.
e) Feature Fusion Modelling (Figure 2e): first, feature vectors were generated distinctly from the raw or preprocessed form of a given NIR and MIR signal and thereafter coupled. The obtained combined feature vector was then input to a regressor, making a final quantification.
f) Decision Fusion Modelling (Figure 2f): estimations created individually using a given NIR and MIR signal were ensembled by a regressor to generate a final estimation.

The goal of including different strategies was to comprehensively investigate the robustness of SFFVG under diverse circumstances. The idea was to generate quantitative models according to all possible

permutations for each strategy and later perform intra-strategy comparisons between models with SFFVG and those without SFFVG as benchmarks. Thus, by incorporating or skipping preprocessing and SFFVG blocks each, four models were constructed using NIR Modelling, MIR Modelling, Raw Spectra Fusion Modelling, and Decision Fusion Modelling. On the other hand, two models were created through each of Preprocessed Spectra Fusion Modelling and Feature Fusion Modelling by incorporating or not incorporating their sole dashed unit. It is worth clarifying that the preprocessing unit in Preprocessed Spectra Fusion Modelling and SFFVG unit in Feature Fusion Modelling was not skippable due to the essence of these strategies.

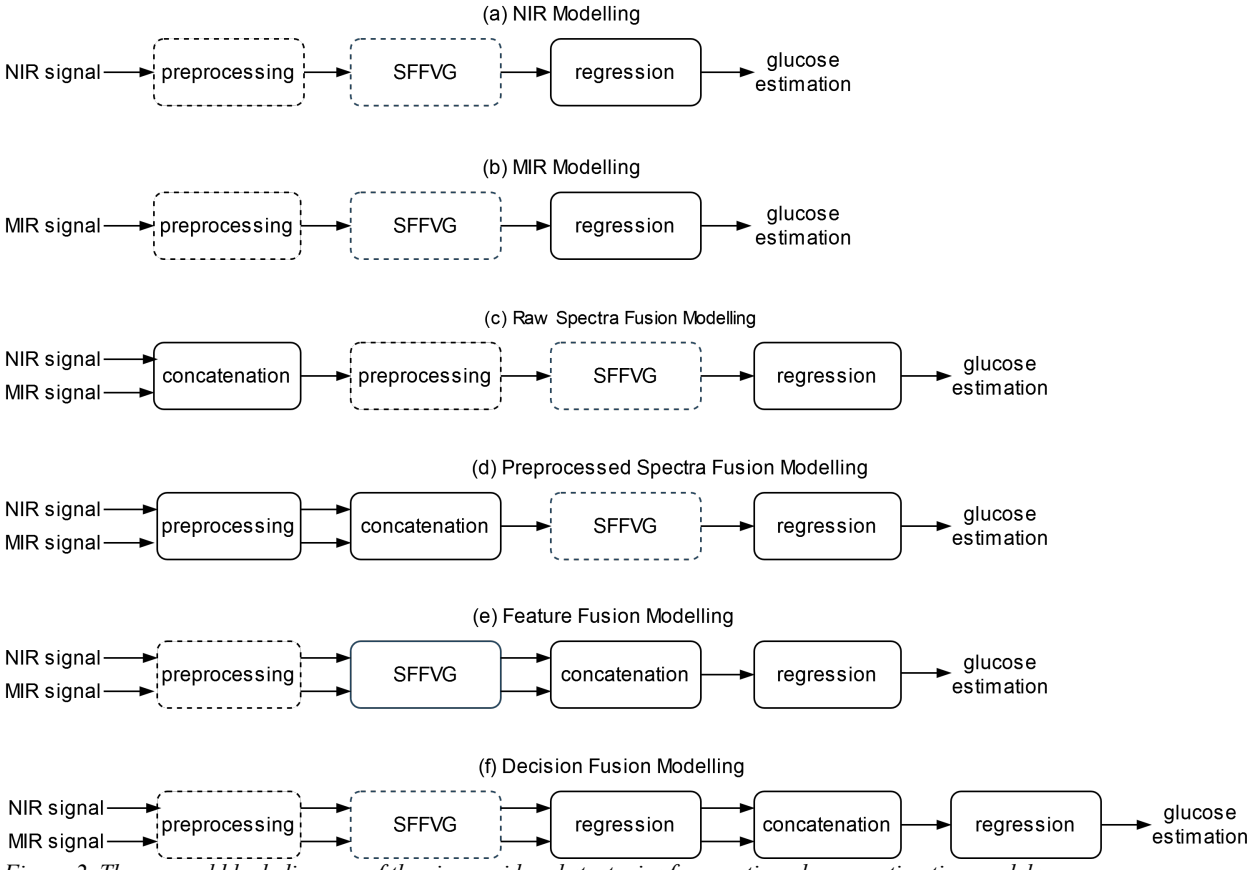

Figure 2. The general block diagram of the six considered strategies for creating glucose estimation models.
Note. For the preprocessing and regression blocks, the Savitzky-Golay filter and partial least square regression were used, respectively. SFFVG (signal fragmentation based feature vector generation) block's internal architecture is shown in Figure 1. The dashed blocks indicate that both conditions with or without including the block were investigated separately. (a, b) NIR Modelling and MIR Modelling, two unimodal strategies where glucose concentrations were estimated from NIR or MIR signals alone. (c) Raw Spectra Fusion Modelling where NIR and MIR data were fused in their raw format and then used to create quantitative models, (d) Preprocessed Spectra Fusion Modelling where NIR and MIR signals were fused after the preprocessing

and then used for constructing quantitative models, (e) Feature Fusion Modelling where features generated from NIR and MIR signal were fused and used to create quantitative models, and (f) Decision Fusion Modelling where quantitative models created from NIR and MIR signals were ensembled to form a combined model.

2.5. Model evaluation

The developed models were evaluated considering three frequently used regression metrics for estimations one the evaluation set; *RMSE* as Eq. (1) and mean absolute percentage deviation (*MAPD*) as Eq. (2) to reflect the error of quantifications [35], [36], and coefficient of determination (r^2) as Eq. (3) as a statistical measure to indicate correlations between the reference and estimated values [37].

$$RMSE = \sqrt{(\sum_{i=1}^{N}(y_i - f_i)^2)/N} \quad (1)$$

$$MAPD = ((\sum_{i=1}^{N}|(y_i - f_i)/y_i|)/N) \times 100 \quad (2)$$

$$r^2 = 1 - (RSS/TSS) \quad (3)$$

where, in Eqs. (1) and (2), N, y_i, and f_i are respectively the size of the evaluation set, actual value, and estimated value; and in Eq. (3), *RSS* and *TSS* respectively represent the residual sum of squares and the total sum of squares.

2.6. Model interpretation

SHAP is a game-theoric ML explainability technique. It stimulates how an ML model produces an estimation for a data instance as a game between input variables. Then, using the Shapley value concept from game theory [25], each input variable's contribution to generated estimation for the data instance is quantified as Eq. (4) [24].

$$SHAP_x(f) = \sum_{F:f\in F}\left(|F| \times \binom{f}{|F|}\right)^{-1} \times \left(\hat{x}_F - \hat{x}_{F\setminus f}\right) \quad (4)$$

where f is a given input variable; x is a given instance of data; $SHAP_x(f)$ represents the quantified contribution level of variable f in the generated estimation for x (SHAP value of variable f for x); F

represents all possible subsets of variables with *f* included; $|F|$ is the size of *F* (number of variables in *F*); $\hat{x}_F$ represents the model's estimation for *x* from *F;* $\hat{x}_{F\setminus f}$ is the model's estimation for *x* from *F* excluding *f*

Following the evaluation analysis, SHAP was deployed to globally interpret models assimilating SFFVG, i.e., explaining the impact of the features generated from sub-signals in producing the estimations across the entire validation set. For this purpose, the mean absolute of features' SHAP values presented in Eq. (4) was used.

This analysis allows analogies to be drawn between the importance of the segregated intervals, thereby increasing the transparency of investigations utilising SFFVG. For conciseness, interpretation was undertaken only for the best model generated by each modelling strategy according to their evaluation results presented later in subsection 4.1.

3. Results and discussion

This section reports the evaluation and model interpretation results alongside the corresponding discussion.

3.1. Signal fragmentation

Based on optimisation for the number of intervals, NIR signals were divided into four equal fragments (100 nm wide apiece), MIR signals into six equal fragments (≈ 916 nm wide apiece), and NIR-MIR signals into ten equal fragments (590 nm wide apiece). The results of spectra fragmentation are summarised in Table 2.

Consequently, for NIR Modelling, the fragmentation module divided signals into the four NIR intervals represented in the table and then the corresponding NIR features were extracted from these intervals. Similarly, for MIR Modelling, the signals were divided into the six MIR intervals shown in the table, and then the corresponding MIR features were generated. For Raw Spectra Fusion Modelling and Preprocessed Spectra Fusion Modelling, signals were divided into the ten NIR-MIR intervals shown in the table and then

the associated NIR-MIR features were constructed. Finally, for Feature Fusion Modelling and Decision Fusion Modelling, NIR and MIR signals were separately fragmented into respectively the four NIR intervals and the six MIR intervals presented in the table, and then the relevant features were created.

Table 2. Signal fragmentation process outcomes including the generated intervals their associated name and feature.

Region	Interval	Interval name	Generated feature
NIR	2100–2200 nm	NIR interval 1	NIR feature 1
	2200–2300 nm	NIR interval 2	NIR feature 2
	2300–2400 nm	NIR interval 3	NIR feature 3
	2400–2500 nm	NIR interval 4	NIR feature 4
MIR	2500–3416 nm	MIR interval 1	MIR feature 1
	3416–4333 nm	MIR interval 2	MIR feature 2
	4334–5250 nm	MIR interval 3	MIR feature 3
	5250–6166 nm	MIR interval 4	MIR feature 4
	6166–7084 nm	MIR interval 5	MIR feature 5
	7084–8000 nm	MIR interval 6	MIR feature 6
NIR-MIR	2100–2690 nm	NIR-MIR interval 1	NIR-MIR feature 1
	2690–3280 nm	NIR-MIR interval 2	NIR-MIR feature 2
	3280–3870 nm	NIR-MIR interval 3	NIR-MIR feature 3
	3870–4460 nm	NIR-MIR interval 4	NIR-MIR feature 4
	4460–5050 nm	NIR-MIR interval 5	NIR-MIR feature 5
	5050–5640 nm	NIR-MIR interval 6	NIR-MIR feature 6
	5640–6230 nm	NIR-MIR interval 7	NIR-MIR feature 7
	6230–6820 nm	NIR-MIR interval 8	NIR-MIR feature 8
	6820–7410 nm	NIR-MIR interval 9	NIR-MIR feature 9
	7410–8000 nm	NIR-MIR interval 10	NIR-MIR feature 10

Note. NIR: near-infrared; MIR: mid-infrared; NIR-MIR: the combination of near- and mid-infrared.

3.2. Model evaluation

Table 3 lists the results of *RMSE*, *MAPD* and r^2 for all created models. The table is compartmentalised with the results of the modelling strategies to facilitate intra-strategy comparisons of SFFVG-included models versus non-SFFVG models.

Each improvement ratio in Table 3 compares the result of an evaluation metric achieved by an SFFVG-included model versus the model with the same strategy and preprocessing but without SFVG (benchmarked model), reported in the raw above. According to the table, in all pair-wise comparisons, the majority of improvement ratios convey the efficacy of SFFVG-included models over non-SFFVG counterparts.

The values in bold in Table 3 are the best result(s) obtained for evaluation metrics through each modelling strategy. Grey cells in the table highlight the model(s) with the highest number of best values for

the evaluation metrics amongst the models created using the same modelling strategy. These highlights indicate that the best model of all strategies was SFFVG-included. Explicitly, applying SFFVG without preprocessing granted the highest overall performance with the best *MAPD* and r^2 values for the NIR Modelling, whilst applying SFFVG with preprocessing gave the lowest *RMSE* in this case. In MIR Modelling, SFFVG without preprocessing yielded the best overall results for all evaluation metrics. Moreover, for Raw Spectra Fusion Modelling, Preprocessed Data Fusion Modelling, and Feature Fusion Modelling, SFFVG joined with preprocessing conferred the best performance overall and according to each evaluation criterion. Finally, the best performance for Decision Fusion Modelling was achieved by SFFVG without preprocessing, the best *RMSE* and r^2, whilst the best *MAPD* was for non-SFFVG with preprocessing.

Overall, incorporating SFFVG in the six modelling strategies enhanced the accuracy of glucose estimation. Moreover, SFFVG maintained its effectiveness when a preprocessing step was also present in the modelling process. Such attainments underpin the functionality and flexibility of the proposed SFFVG approach. The coordinating power of stack learning could justify such fulfilments; deriving glucose information from fragments of a signal and then aggregating the outcomes dominated studying the whole signal at once. Finally, it is noteworthy that pre-partitioning procedures have recently found successful applications in image processing tasks, supporting the relevance of the core idea involved in this work to other areas where further exploration would be desirable [38,39].

Comparing bimodal strategies with MIR Modelling reveals that Raw Spectra Fusion Modelling, Preprocessed Data Fusion Modelling, and Feature Fusion Modelling produced results on par with MIR modelling whilst not conclusively outperforming it. Nevertheless, rather than taking advantage of synergistic effects, the object of including bimodal strategies in this work was to test SFFVG's capability under a broader range of spectra with different data fusion strategies.

Table 3. Evaluation results for all generated quantitative models.

Strategy	Preprocessing	SFFVG	RMSE (mg dL^{-1})	RMSE IR (%)	MAPD (%)	MAPD IR (%)	r^2	r^2 IR (%)
NIR Modelling	No	No	98.1	—	66.6	—	0.47	—
		Yes	91.0	+7.2	**46.5**	+30.1	**0.58**	+23.4
	Yes	No	98.7	—	67.3	—	0.46	—
		Yes	**89.7**	+9.1	48.0	+28.6	0.55	+19.5
MIR Modelling	No	No	36.3	—	28.0	—	0.92	—
		Yes	**24.5**	+32.5	**24.4**	+12.8	**0.96**	+4.3
	Yes	No	36.2	—	27.8	—	0.92	—
		Yes	24.6	+32.0	24.7	+11.1	**0.96**	+4.3
Raw Spectra Fusion Modelling	No	No	34.4	—	28.8	—	0.93	—
		Yes	32.3	+6.1	26.8	+6.9	**0.94**	+1.0
	Yes	No	34.5	—	29.0	—	0.93	—
		Yes	**32.1**	+6.6	**25.6**	+11.7	**0.94**	+1.0
Preprocessed Spectra Fusion Modelling	Yes	No	34.2	—	28.7	—	0.93	—
		Yes	**32.2**	+5.8	**25.7**	+10.4	**0.94**	+1.0
Feature Fusion Modelling	No	Yes	27.7	—	25.1	—	0.95	—
	Yes	Yes	**26.6**	+3.9	**24.1**	+3.9	**0.96**	+1.0
Decision Fusion Modelling	No	No	60.1	—	35.2	—	0.80	—
		Yes	**47.0**	+21.7	37.0	-5.1	**0.87**	+8.7
	Yes	No	60.0	—	**35.0**	—	0.81	—
		Yes	49.6	+17.3	38.3	-8.6	0.86	+5.8

Note. SFFVG: signal fragmentation based feature vector generation; RMSE: root mean square error; IR: improvement ratio (comparing the results of an SFFVG-included model versus the benchmarked non-SFFVG model reported in the raw above.); MAPD: mean absolute percentage deviation; r2: coefficient of determination; NIR: near-infrared; MIR: mid-infrared. The values in bold font indicate the best result for each evaluation metric in each strategy. The grey cells indicate the model(s) with the highest number of best-obtained evaluation metrics amongst models developed using the same modelling strategy.

3.3. Model interpretation

Figure 3 represents the variable importance graphs for the best model of each strategy (marked with grey cells in Table 3). The length of each bar in the graphs expresses the importance rate of the corresponding feature according to mean absolute SHAP values over the entire validation set.

As presented in Figure 3a, NIR feature 4 (associated with interval 2400–2500 nm) was the most informative variable for the best model from NIR Modelling. NIR features 1, 3, and 2 (intervals 2100–2200 nm, 2300–2400 nm, and 2200–2300 nm, respectively) in order placed in ranks 2 to 4.

According to Figure 3b, MIR feature 1 (interval 2500–3416 nm) had the dominant influence on the best model of MIR Modelling with a mean absolute SHAP value remarkably superior to others. In contrast, MIR feature 2 (interval 3416–4333 nm) carried the most inferior influence with a mean absolute SHAP value considerably lower than others. In comparison, MIR features 3, 4, 5, and 6 (intervals 4333–5250 nm, 5250–6166 nm, 6166–7084 nm, and 7084–8000 nm, respectively) induced comparable and medium impacts on the model.

For the best model of Raw Spectra Fusion Modelling (Figure 3c) and Preprocessed Spectra Fusion Modelling (Figure 3d), NIR-MIR feature 1 (interval 2100–2690 nm) supplied the maximum impact on the model with a mean absolute SHAP value appreciably higher than others. NIR-MIR features 2 and 8 (intervals 2690–3280 nm and 6230–6820 nm, respectively) placed the second and third rank. Other features had relatively subordinate effects.

Based on Figure 3e, for the best model of Feature Fusion Modelling, the impact of MIR feature 1 (interval 2500–3416 nm) outweighed that of other MIR and NIR features with a mean absolute SHAP value markedly higher than others. MIR feature 3 (interval 4333–5250 nm) and NIR feature 4 (interval 2400–2500 nm) placed in the second and third rank. The other three NIR features (1, 2, and 3) had the weakest impact on the model.

Figure 3f displays the variable importance plot for the best model of Decision Fusion Modelling. In this case, since decisions of NIR and MIR models were combined at the final stage, the effect of NIR and MIR decisions on the final estimations were compared. The results illustrate that the influence of MIR decisions on the models' outcomes surpassed NIR decisions.

According to the interpretation analysis, potential associations between the most informative intervals detected and the nearest glucose-informative bands according to the ordinary glucose signature in NIR and MIR regions could be inferred [40]. For instance, information possessed by the most influential features in different regions was potentially connected to the following vibrations in glucose molecule bonds: a combination of vibrations in CH and CH2 bonds for NIR feature 4, stretching vibrations in OH and CH bonds for MIR feature 1, a combination of vibrations in OH, CH, and CH2 bonds for NIR-MIR feature 1.

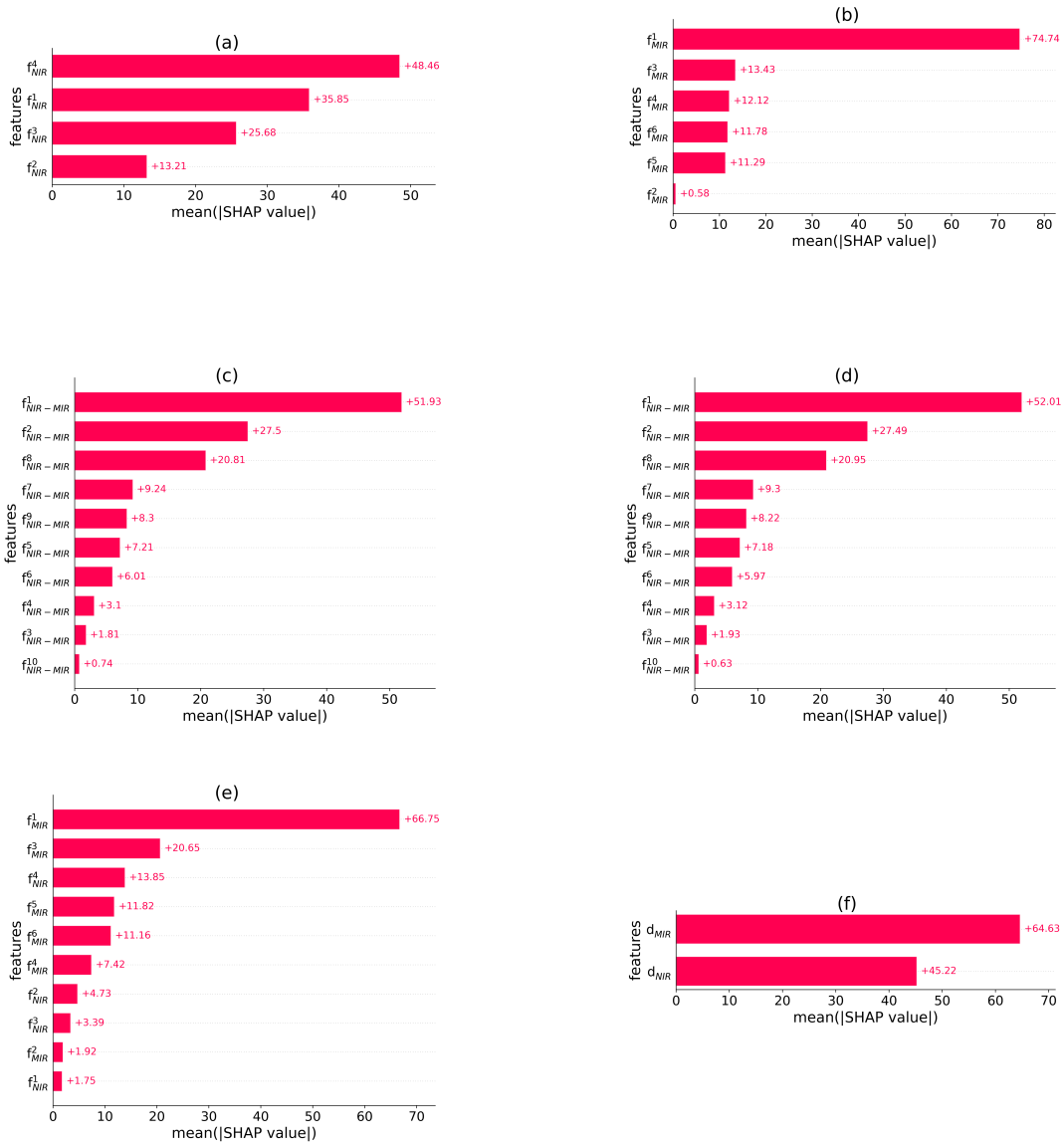

Figure 3. Feature importance plots which indicate the influence of variables upon collective absolute (SHapley Additive exPlanations) SHAP values for the best model from each modelling strategy (a) NIR modelling, (b) MIR modelling, (c) raw spectra fusion modelling, (d) preprocessed spectra fusion modelling, (e) feature fusion modelling, and (f) decision fusion modelling.
Note. f_s^i: feature generated from the ith fragment of s signals. d_s: decision from s signals.

3.4. Complementary analysis

3.4.1. Comparative analysis

Interval partial least squares (iPLS) is a well-known variable selection technique in spectroscopy data analysis [40]. The technique starts with breaking signals into several intervals. Then, some of these intervals

are selected for subsequent modelling analysis, where the selected intervals are stacked and inputted into a prediction model.

We conducted a basic comparative analysis between iPLS and the proposed SFFVG method. To this end, using the same block diagram shown in Figure 2, for each model with SFFVG, a comparator model with iPLS was constructed. The building blocks of each comparator model were similar to its reference model, except the SFFVG unit was replaced with an iPLS unit. For the sake of fair comparisons, the units of each comparator model underwent the same optimisation process performed for its reference model's units. As a result of identical fragmentation optimisation, the same intervals represented in Table 2 were utilised for comparator models with iPLS.

In iPLS analysis, first, an autonomous glucose quantification model was created for each interval. The interval providing the lowest RMSE of five-fold cross-validation on the calibration set was selected. Then, the selected interval combined with the remaining intervals, one at a time, were used to build quantitative models. The combination of intervals that produced the model with the lowest RMSE of five-fold cross-validation on the calibration set was selected. This successive interval selection cycle was repeated until adding a new interval could not lower the RMSE of five-fold cross-validation on the calibration set.

Table 4 presents the results of the comparative analysis of SFFVG and the iPLS. Values in bold and grey cells indicate the same information as explained in subsection 3.2 for Table 3. According to the grey cells in the table, four of the models with the dominant number of best-obtained evaluation metrics (amongst the models generated using each modelling strategy) were with SFFVG. These outcomes further support the capability of SFFVG.

Table 4. Evaluation results for the comparision analysis between SFFVG and iPLS.

Strategy	Preprocessing	Feature enineeing	RMSE (mg dL^{-1})	MAPD (%)	r^2
NIR Modelling	No	SFFVG	91.0	**46.5**	**0.58**
		iPLS	99.9	69.7	0.45
	Yes	SFFVG	**89.7**	48.0	0.55
		iPLS	99.9	71.3	0.44
MIR Modelling	No	SFFVG	**24.5**	24.4	**0.96**
		iPLS	29.3	18.5	0.95
	Yes	SFFVG	24.6	24.7	**0.96**
		iPLS	28.7	**18.4**	**0.96**
Raw Spectra Fusion Modelling	No	SFFVG	32.3	26.8	**0.94**
		iPLS	33.7	19.2	0.93
	Yes	SFFVG	32.1	25.6	**0.94**
		iPLS	**31.8**	**17.7**	**0.94**
Preprocessed Spectra Fusion Modelling	Yes	SFFVG	32.2	25.7	**0.94**
		iPLS	**31.8**	**17.7**	**0.94**
Feature Fusion Modelling	No	SFFVG	27.7	25.1	0.95
		iPLS	29.5	16.8	0.95
	Yes	SFFVG	**26.6**	24.1	**0.96**
		iPLS	29.1	**16.1**	0.95
Decision Fusion Modelling	No	SFFVG	**47.0**	**37.0**	**0.87**
		iPLS	49.6	38.4	0.86
	Yes	SFFVG	49.6	38.3	0.86
		iPLS	49.5	38.5	0.86

Note. iPLS: interval partial least squares; SFFVG: signal fragmentation based feature vector generation; RMSE: root mean square error; IR: improvement ratio (comparing the results of an SFFVG-included model versus the benchmarked non-SFFVG model reported in the raw above.); MAPD: mean absolute percentage deviation; r2: coefficient of determination; NIR: near-infrared; MIR: mid-infrared. The values in bold font indicate the best result for each evaluation metric in each strategy. The grey cells indicate the model(s) with the highest number of best-obtained evaluation metrics amongst models developed using the same modelling strategy.

3.4.2. Reevaluation analysis

To further examine the functionality of SFFVG, after reshuffling the data and performing another 80-20 calibration and validation split, we reconducted the model generation and evaluation analysis. The results of this extra analysis are summarised in Table 5. Values in bold and grey cells in the table denote the same information as explained in subsection 3.2 for Table 3. Overall, according to Table 5, intra-strategy analogies reaffirmed the principal outcomes reported in subsection 3.2. Explicitly, again, models with SFFVG outperformed their counterparts without SFFVG in most scenarios. Also, in all strategies, the model(s) with the highest number of best-obtained evaluation metrics included SFFVG.

Table 5. Results of reevaluation analysis for all investigated scenarios.

Strategy	Preprocessing	SFFVG	RMSE (mg dL^{-1})	MAPD (%)	r^2
NIR Modelling	No	No	101.4	83.3	0.48
		Yes	97.4	53.7	0.52
	Yes	No	103.4	77.6	0.46
		Yes	**95.4**	**52.6**	**0.54**
MIR Modelling	No	No	37.4	31.7	0.93
		Yes	35.6	**21.3**	**0.94**
	Yes	No	37.3	30.9	0.93
		Yes	**35.0**	21.4	**0.94**
Raw Spectra Fusion Modelling	No	No	33.0	25.0	0.94
		Yes	32.8	**17.3**	**0.95**
	Yes	No	33.1	22.1	0.94
		Yes	**31.1**	17.4	**0.95**
Preprocessed Spectra Fusion Modelling	Yes	No	32.8	22.0	0.94
		Yes	**31.0**	**17.3**	**0.95**
Feature Fusion Modelling	No	Yes	**36.7**	36.9	**0.94**
		Yes	37.1	**35.7**	0.93
Decision Fusion Modelling	No	No	57.6	49.3	0.83
		Yes	**54.2**	**30.3**	**0.85**
	Yes	No	101.4	83.3	0.48
		Yes	97.4	53.7	0.52

Note. SFFVG: signal fragmentation based feature vector generation; RMSE: root mean square error; IR: improvement ratio (comparing the results of an SFFVG-included model versus the benchmarked non-SFFVG model reported in the raw above.); MAPD: mean absolute percentage deviation; r2: coefficient of determination; NIR: near-infrared; MIR: mid-infrared. The values in bold font indicate the best result for each evaluation metric in each strategy. The grey cells indicate the model(s) with the highest number of best-obtained evaluation metrics amongst models developed using the same modelling strategy.

4. Summary and conclusion

Feature vector generation based on signal partitioning and framed with model interpretation analysis enhanced in vitro glucose quantification from absorption spectroscopy. First, a given spectrum was sliced into fragments. A base-regressor then analysed these fragments individually, forming preliminary glucose concentration estimations. These estimations were then stacked, generating a feature vector for the original spectrum. Later, leveraging the concept of stack learning, a meta-regressor investigates this feature vector to produce a final estimation of the reference glucose concentration. The versatility of the proposed method was tested under an array of modelling strategies. Moreover, the compatibility of the proposed method with a standard preprocessing technique was investigated. Overall, the results obtained accentuated the efficacy of the proposed method in improving glucose quantifications for all modelling strategies. The method maintained its functionality when a preprocessing step was also incorporated into the modelling process. Finally, SHAP was employed to interpret the outcomes of the quantitative analysis. Such interpretation encourages the adoption of the proposed method by extending the transparency of the analysis. For future work, applying the proposed methodology with ununiformed spectra fragmentation is recommended.

Code availability

We coded in Python (3.6.7); the packages Pandas, NumPy and Sklearn were used for the analysis. The source code of implementations is publicly available in this repository.

Acknowledgement

We thank Dr Osamah Alrezj for his efforts in preparing the experimental data used in this work.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] A.L. Galant, R.C. Kaufman, J.D. Wilson, Glucose: Detection and analysis, Food Chem. 188 (2015) 149–160. https://doi.org/http://dx.doi.org/10.1016/j.foodchem.2015.04.071.

[2] R. Ahmad, M. Khan, N. Tripathy, M.I.R. Khan, A. Khosla, Hydrothermally Synthesized Nickel Oxide Nanosheets for Non-Enzymatic Electrochemical Glucose Detection, J. Electrochem. Soc. 167 (2020) 107504. https://doi.org/10.1149/1945-7111/ab9757.

[3] J. Boudrant, L.P. Fonseca, A.N. Reshetilov, K. Pontius, D. Semenova, Y.E. Silina, K. V Gernaey, H. Junicke, Automated Electrochemical Glucose Biosensor Platform as an Efficient Tool Toward On-Line Fermentation Monitoring: Novel Application Approaches and Insights, Front. Bioeng. Biotechnol. 8 (2020) 1–15. https://doi.org/10.3389/fbioe.2020.00436.

[4] I. Delfino, C. Camerlingo, M. Portaccio, B. Della Ventura, L. Mita, D.G. Mita, M. Lepore, Visible micro-Raman spectroscopy for determining glucose content in beverage industry, Food Chem. 127 (2011) 735–742. https://doi.org/http://dx.doi.org/10.1016/j.foodchem.2011.01.007.

[5] M. Shokrekhodaei, Non-Invasive In-Vitro Glucose Monitoring Using Optical Sensor and Machine Learning Techniques for Diabetes Applications, The University of Texas at El Paso, 2021.

[6] A. Al-Mbaideen, M. Benaissa, Coupling subband decomposition and independent component regression for quantitative NIR spectroscopy, Chemom. Intell. Lab. Syst. 108 (2011) 112–122. https://doi.org/10.1016/j.chemolab.2011.05.012.

[7] J. Haas, B. Mizaikoff, Advances in Mid-Infrared Spectroscopy for Chemical Analysis, Annu. Rev.

Anal. Chem. 9 (2016) 45–68. https://doi.org/10.1146/annurev-anchem-071015-041507.

[8] S.K. Vashist, Non-invasive glucose monitoring technology in diabetes management: A review, Anal. Chim. Acta. 750 (2012) 16–27. https://doi.org/10.1016/j.aca.2012.03.043.

[9] D.A. Burns, E.W. Ciurczak, Handbook of near-infrared analysis, CRC press, 2007.

[10] S. Delbeck, H.M. Heise, Evaluation of opportunities and limitations of mid-infrared skin spectroscopy for noninvasive blood glucose monitoring, J. Diabetes Sci. Technol. 15 (2021) 19–27. https://doi.org/https://doi.org/10.1177/1932296820936224.

[11] B. Rabinovitch, W.F. March, R.L. Adams, Noninvasive glucose monitoring of the aqueous humor of the eye: Part I. Measurement of very small optical rotations, Diabetes Care. 5 (1982) 254–258. https://doi.org/https://doi.org/10.2337/diacare.5.3.254.

[12] J. Yadav, A. Rani, V. Singh, B.M. Murari, Prospects and limitations of non-invasive blood glucose monitoring using near-infrared spectroscopy, Biomed. Signal Process. Control. 18 (2015) 214–227. https://doi.org/10.1016/j.bspc.2015.01.005.

[13] C.-F. So, K.-S. Choi, T.K.S. Wong, J.W.Y. Chung, Recent advances in noninvasive glucose monitoring, Med. Devices Evid. Res. 5 (2012) 45–52. https://doi.org/https://dx.doi.org/10.2147%2FMDER.S28134.

[14] H. von Lilienfeld-Toal, M. Weidenmüller, A. Xhelaj, W. Mäntele, A novel approach to non-invasive glucose measurement by mid-infrared spectroscopy: The combination of quantum cascade lasers (QCL) and photoacoustic detection, Vib. Spectrosc. 38 (2005) 209–215. https://doi.org/https://doi.org/10.1016/j.vibspec.2005.02.025.

[15] B.K. Mekonnen, W. Yang, T.-H. Hsieh, S.-K. Liaw, F.-L. Yang, Accurate prediction of glucose concentration and identification of major contributing features from hardly distinguishable near-infrared spectroscopy, Biomed. Signal Process. Control. 59 (2020) 1–15. https://doi.org/https://doi.org/10.1016/j.bspc.2020.101923.

[16] A. Tura, A. Maran, G. Pacini, Non-invasive glucose monitoring: Assessment of technologies and devices according to quantitative criteria, Diabetes Res. Clin. Pract. 77 (2007) 16–40. https://doi.org/10.1016/j.diabres.2006.10.027.

[17] G. Han, S. Chen, X. Wang, J. Wang, H. Wang, Z. Zhao, Noninvasive blood glucose sensing by near-infrared spectroscopy based on PLSR combines SAE deep neural network approach, Infrared Phys. Technol. 113 (2021) 1–10. https://doi.org/https://doi.org/10.1016/j.infrared.2020.103620.

[18] J. Tenhunen, H. Kopola, R. Myllylä, Non-invasive glucose measurement based on selective near infrared absorption; requirements on instrumentation and spectral range, Meas. J. Int. Meas. Confed. 24 (1998) 173–177. https://doi.org/10.1016/S0263-2241(98)00054-2.

[19] Å. Rinnan, F. van den Berg, S.B. Engelsen, Review of the most common pre-processing techniques for near-infrared spectra, TrAC - Trends Anal. Chem. 28 (2009) 1201–1222. https://doi.org/10.1016/j.trac.2009.07.007.

[20] H. Khadem, M.R. Eissa, H. Nemat, O. Alrezj, M. Benaissa, Classification before regression for improving the accuracy of glucose quantification using absorption spectroscopy, Talanta. 211 (2020) 1–10. https://doi.org/https://doi.org/10.1016/j.talanta.2020.120740.

[21] Z.-H. Zhou, Ensemble methods: foundations and algorithms, Chapman and Hall/CRC, 2019.

[22] E. Mauer, J. Lee, J. Choi, H. Zhang, K.L. Hoffman, I.J. Easthausen, M. Rajan, M.G. Weiner, R. Kaushal, M.M. Safford, others, A predictive model of clinical deterioration among hospitalized COVID-19 patients by harnessing hospital course trajectories, J. Biomed. Inform. 118 (2021) 1–12. https://doi.org/10.1016/j.jbi.2021.103794.

[23] S. Bhatt, A. Cohon, J. Rose, N. Majerczyk, B. Cozzi, D. Crenshaw, G. Myers, Interpretable machine learning models for clinical decision-making in a high-need, value-based primary care setting, NEJM Catal. Innov. Care Deliv. 2 (2021). https://doi.org/https://doi.org/10.1056/CAT.21.0008.

[24] S. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: 31th Conf. Neural Inf. Process. Syst., 2017: pp. 4765–4774.

[25] L.S. Shapley, A value for n-person games, Contrib. to Theory Games. 2 (1953) 307–317.

[26] O. Alrezj, M. Benaissa, S.A. Alshebeili, Digital bandstop filtering in the quantitative analysis of glucose from near-infrared and midinfrared spectra, J. Chemom. 34 (2020) e3206. https://doi.org/https://doi.org/10.1002/cem.3206.

[27] Ian T. Jolliffe, A Note on the Use of Principal Components in Regression, J. R. Stat. Soc. 31 (1982) 300–303.

[28] G.M. Escandar, P.C. Damiani, H.C. Goicoechea, A.C. Olivieri, A review of multivariate calibration methods applied to biomedical analysis, Microchem. J. 82 (2006) 29–42. https://doi.org/10.1016/j.microc.2005.07.001.

[29] Y. Li, Y. Xiong, S. Min, Data fusion strategy in quantitative analysis of spectroscopy relevant to olive oil adulteration, Vib. Spectrosc. 101 (2019) 20–27. https://doi.org/10.1016/j.vibspec.2018.12.009.

[30] Y. Li, J.Y. Zhang, Y.Z. Wang, FT-MIR and NIR spectral data fusion: a synergetic strategy for the geographical traceability of Panax notoginseng, Anal. Bioanal. Chem. 410 (2018) 91–103. https://doi.org/10.1007/s00216-017-0692-0.

[31] W. Sun, X. Zhang, Z. Zhang, R. Zhu, Data fusion of near-infrared and mid-infrared spectra for identification of rhubarb, Spectrochim. Acta - Part A Mol. Biomol. Spectrosc. 171 (2017) 72–79. https://doi.org/10.1016/j.saa.2016.07.039.

[32] L. Tao, B. Via, Y. Wu, W. Xiao, X. Liu, NIR and MIR spectral data fusion for rapid detection of Lonicera japonica and Artemisia annua by liquid extraction process, Vib. Spectrosc. 102 (2019) 31–38. https://doi.org/10.1016/j.vibspec.2019.03.005.

[33] Y. Wang, M. Yang, G. Wei, R. Hu, Z. Luo, G. Li, Improved PLS regression based on SVM classification for rapid analysis of coal properties by near-infrared reflectance spectroscopy, Sensors Actuators, B Chem. 193 (2014) 723–729. https://doi.org/10.1016/j.snb.2013.12.028.

[34] A. Savitzky, M.J.E. Golay, Smoothing and Differentiation of Data by Simplified Least Squares Procedures, Anal. Chem. 36 (1964) 1627–1639. https://doi.org/10.1021/ac60214a047.

[35] M.L.F. Simeone, R.A.C. Parrella, R.E. Schaffert, C.M.B. Damasceno, M.C.B. Leal, C. Pasquini, Near infrared spectroscopy determination of sucrose, glucose and fructose in sweet sorghum juice, Microchem. J. 134 (2017) 125–130. https://doi.org/10.1016/j.microc.2017.05.020.

[36] A. De Myttenaere, B. Golden, B. Le Grand, F. Rossi, Mean absolute percentage error for regression models, Neurocomputing. 192 (2016) 38–48. https://doi.org/https://doi.org/10.1016/j.neucom.2015.12.114.

[37] and W.A.N. Lee Rodgers, Joseph, Thirteen ways to look at the correlation coefficient., Am. Stat. 42 (1988) 59–66. https://doi.org/https://doi.org/10.1080/00031305.1988.10475524.

[38] P. Mlynarski, H. Delingette, A. Criminisi, N. Ayache, 3D convolutional neural networks for tumor segmentation using long-range 2D context, Comput. Med. Imaging Graph. 73 (2019) 60–72. https://doi.org/https://doi.org/10.1016/j.compmedimag.2019.02.001.

[39] H. Hui, X. Zhang, F. Li, X. Mei, Y. Guo, A Partitioning-Stacking Prediction Fusion Network Based on an Improved Attention U-Net for Stroke Lesion Segmentation, IEEE Access. 8 (2020) 47419–47432. https://doi.org/10.1109/ACCESS.2020.2977946.

[40] M. Golic, K. Walsh, P. Lawson, Short-wavelength near-infrared spectra of sucrose, glucose, and fructose with respect to sugar concentration and temperature, Appl. Spectrosc. 57 (2003) 139–145. https://doi.org/10.1366/000370203321535033.

Publication 5.

COVID-19 mortality risk assessments for individuals with and without diabetes mellitus: machine learning models integrated with interpretation framework[5]

Abstract. This research develops machine learning models equipped with interpretation modules for mortality risk prediction and stratification in cohorts of hospitalised coronavirus disease-2019-(COVID-19) patients with and without diabetes mellitus (DM). To this end, routinely collected clinical data from 156 COVID-19 patients with DM and 349 COVID-19 patients without DM were scrutinised. First, a random forest classifier forecasted in-hospital COVID-19 fatality utilising admission data for each cohort. For the DM cohort, the model predicted mortality risk with an accuracy of 82%, area under the curve (AUC) of 80%, sensitivity of 80%, and specificity of 56%. For the non-DM cohort, the achieved accuracy, AUC, sensitivity, and specificity were 80%, 84%, 91%, and 56%, respectively. The models were then interpreted using SHapley Additive exPlanations (SHAP), which explained the predictors' global and local influences on model outputs. Finally, the k-means algorithm was applied to cluster patients on their SHAP values. The algorithm demarcated patients into three clusters. Average mortality rates within the generated clusters were 8%, 20%, and 76% for the DM cohort, 2.7%, 28%, and 41.9% for the non-DM cohort, providing a functional method of risk stratification.

Keywords. Machine Learning; COVID-19; Diabetes Mellitus; Risk Assessment; Model Interpretation

---

[5] This article was published in Computers in Biology and Medicine 144 (2022). Authors: **H. Khadem**, H. Nemat, M.R. Eissa, J. Elliott, M. Benaissa.

1. Introduction

Diabetes mellitus (DM) was identified as a risk factor for coronavirus disease-2019 (COVID-19) shortly after the spread of the new disease [1]–[3]. Later, it was argued that DM comorbidity was a leading cause of death in people hospitalised for COVID-19 [4].

These realisations spurred efforts towards assessing COVID-19 mortality risk in people with DM. For example, Sourij et al. investigated the predictors of in-hospital COVID-19 mortality in patients with DM, followed by the development of a risk score for predicting fatal outcomes [5]. Furthermore, in another study, Ciardullo et al. reported that DM was independently associated with increased in-hospital COVID-19 mortality using multivariable logistic regression to evaluate the effect of DM on COVID-19 mortality [6].

Due to these efforts, the COVID-19 susceptibility of DM patients and the need for more intensive surveillance in hospitalised COVID-19 patients with DM have been well documented. However, additional research is underway to determine the cause of this vulnerability, which has remained a global healthcare challenge [7].

One strategy for elucidating the increased vulnerability of COVID-19 patients with DM is to conduct observational studies on defined populations of COVID-19 patients with and without DM [8]. Such studies aim to identify distinctive characteristics of COVID-19 patients with DM, thereby advancing our understanding of their increased vulnerability. In this respect, several comparative risk assessment studies in COVID-19 patients with and without DM have been conducted [8]–[10]. These studies effectively distinguished risk predictions and risk factors for COVID-19 patients with and without DM, primarily through standard statistical analysis.

Machine learning (ML), as a complementary data analysis tool, possesses significant power in discriminating outcomes due to the capability to discover complex correlated interactions [11]. ML algorithms have demonstrated efficacy in COVID-19 risk assessment research [12]–[14]. For instance, Gao et al. developed an ensemble model to efficiently forecast deterioration and death for COVID-19 patients

up to 20 days ahead of time [15]. This evidence supports further exploration of advanced ML techniques in observational studies of COVID-19 patients with and without DM.

A concern with ML methods in healthcare applications is the black-box nature of these methods, in which the process of generating a specific outcome is unclear [16]. In this context, incorporating interpretation frameworks could further promote the adoption of an ML method designed to combat COVID-19. These frameworks increase analysis transparency and provide results beyond the domain of classical data analysis approaches, e.g., individualised explanations versus generic descriptions [17].

The use of SHapley Additive exPlanations (SHAP) is an elaborate approach to increasing the transparency of ML models. SHAP is a game-theoretic model agnostic technique that can interpret ML models' outputs by integrating optimal credit allocation with local explanations using the classical Shapley values from cooperative game theory [18]. The resulting SHAP values denote the deviation from the average prediction when conditioning on a particular feature, elucidating the influence of individual attributes on the model's outputs [18].

SHAP analysis transforms and scales the features. This conversion enables the formation of meaningful clusters based on explainable similarities. SHAP clustering, as an extension of the original SHAP analysis, partitions data points into groups based on their SHAP values [19].

ML models equipped with SHAP have been considered in previous risk assessment research on DM patients [20] as well as COVID-19 patients [21]–[23]. Specifically, after shortlisting eight out of 100 collated variables, Pan et al. developed SHAP-incorporated ML models for prognosis assessment of COVID-19 patients hospitalised in intensive care units [23].

In this research, first, a model was created for each cohort utilising the random forest (RF) classifier to predict COVID-19 outcomes (death or survival) from admission characteristics. Following that, the outputs of the models were explained globally and locally using SHAP. The most predictive features for each cohort were then identified and rated based on the interpretation results. Finally, patients were clustered according to their SHAP values to form a risk stratification. The main contributions of the work encompass:

- Developing ML models for in-hospital mortality risk assessment of DM and non-DM COVID-19 patients;
- Incorporating an interpretation module into the developed models, explaining significant distinctions between the two cohorts;
- Examining the capability of SHAP clustering for risk stratification of COVID-19 patients with and without DM.

2. Material and methods

Advanced machine learning techniques were employed for mortality risk prediction and stratification of hospitalised COVID-19 patients with and without DM. After cleaning and preprocessing the data, predictive features were determined for each cohort. Then, an RF classifier was assigned to predict admission outcomes for each cohort using the selected features. In the next step, SHAP explained classifiers' outputs at a global and local level. Finally, a k-means algorithm studied generated SHAP values, resulting in the formation of clusters useful in risk assessment practice. For the analysis, we coded in Python (3.6.7); Pandas, NumPy and Sklearn, and shap 0.39.0 packages were also used. The dataset used and the details of how the methodologies were implemented are described in this section. The work was approved by the East-Midlands-Leicester South Research Ethics Committee (20/EM/0145).

2.1. Clinical data

This research developed and evaluated models for mortality risk assessment using demographic, clinical, and laboratory data from 505 participants with confirmed COVID-19. Of the 505 participants, 156 had DM (type 1: 13, type 2: 143). The patients were admitted at Sheffield Teaching Hospitals, Sheffield, UK, between 29 February 2020 and 01 May 2020, coinciding with the first COVID-19 wave in the UK. A comprehensive description of the dataset alongside a detailed explanation of the data collection process can be found in [9]. In line with previous COVID-19 research on individuals with DM [9], in this study, patients

with type 1 and type 2 DM were combined in one cohort (DM cohort) and those without diabetes in another cohort (non-DM cohort). Table 1 summarises admission outcomes for DM and non-DM cohorts.

As this work assessed COVID-19 mortality, 15 individuals, who died due to causes other than COVID-19, were excluded from the remainder of the analysis. Based on the table, the COVID-19 death ratio was higher for the DM cohort (51/156) than in the non-DM cohort (77/349), correlating with existing evidence that people with DM are at an increased risk of COVID-19-related mortality [4].

Table 1. Summary of admission outcomes for DM (diabetes mellitus) and non-DM cohorts.

Outcome of admission	DM cohort	Non-DM cohort
COVID-19 mortality	51	77
Non-COVID-19 mortality	3	12
Survival from COVID-19	102	260

Tables 2 and 3 summarise the attributes collected at the point of hospital admission for both DM and non-DM cohorts. A comprehensive statistical analysis of the data presented in the table can be found in Ref. [9]. The current study leverages ML techniques to determine in-hospital COVID-19 mortality risk.

The two categorical variables *NLRL* (neutrophils-lymphocytes ratio labelled) and *APTTL* (activated partial thromboplastin time labelled), shown in Table 3, were created and added to the feature set by binning corresponding numerical variables. A previous study confirmed the association between these two characteristics and in-hospital COVID-19 mortality in DM patients [9]. For generating the *NLRL* feature, *NLR* values less than eight were labelled as ‘low’, while those greater than eight were labelled as ‘high’. Similarly, for *APTTL*, *APTT* values less than 24s were classified as ‘low’, while those greater than 24s were classified as ‘high’.

Table 2. Numerical baseline clinical characteristics of DM (diabetes mellitus) and non-DM cohorts before hospitalisation for COVID-19.

Feature	Mean standard deviation	
	DM	Non-DM
Frailty score A	5.2±1.8	4.2±2.3
Age (yrs)	71.8 ± 14.9	68.6 ± 18.1
BMI (kg/m2)	29.2 ± 8.5	26.9 ± 7.1
Hb (g/l)	122.1±20.9	130.1±21.5
WCC (g/l)	9.2±5.1	8.1±4.5
Neutrophils (109/l)	7.1±4.6	6.2±4.1
Lymphocytes (109/l)	1.3±1.9	1.2±1.0
Monocytes (109/l)	0.7±0.4	0.6±0.4
Platelets (1/ml)	240.1±106.6	224.8±85.0
Na (mmol/l)	135.2±3.9	136.8±4.6
K (mmol/l)	3.8±1.9	3.7±1.5
Urea (mmol/l)	11.6±7.4	8.2±5.9
Creatinine (µmol/l)	188.4±206.7	114.5±114.5
eGFR(1.73ml.m2/min)	43.2±23.6	27.3±22.7
Bilirubin (µmol/l)	9.7±5.9	12.5±15.9
ALT (u/l)	27.1±27.9	39.5±121.6
Total protein (g/l)	68.3±6.4	68.1±7.3
ALPO4 (g/l)	101.8±62.8	97.7±111.5
Albumin (g/l)	36.8±4.6	38.7±5.1
CRP (mg/dl)	100.3±99.6	82.2±89.2
Procalcitonin (µg/l)	0.7±1.8	1.4±7.7
Ferritin (µg/l)	863.7±1620	834.6±1080
PT(s)	13.5±7.7	11.9±2.3
Fibrinogen (g/l)	5.7±1.2	5.2±1.4
D-dimer (µg/l)	3281±6029	4160±8135
APTT (S)	29.5±14.5	25.9±4.7

F: Frailty measured by Rockwood score; mild (1-3), moderate (4-6), and severe (7-9).
Note. BMI body mass index; Hb haemoglobin; WCC white cell count; Na sodium; K potassium; eGFR estimated glomerular filtration rate; ALT alanine transaminase; ALPO4 alkaline phosphates; CRP c-reactive protein; PT prothrombin time; APTT activated partial thromboplastin time.

Table 3. Categorical baseline clinical characteristics of DM (diabetes mellitus) and non-DM cohorts before hospitalisation for COVID-19.

Feature	Category	Frequency [A]	
		DM	Non-DM
Sex	Male	61	57
	Female	39	43
Ethnicity [B]	White	81	88
	Other	19	12
Smoking status [C]	Non-smoker	45	43
	(ex-)smoker	55	57
IHD	Yes	33	17
	No	67	83
Stroke/TIA	Yes	25	15
	No	75	85
Haemodialysis	Yes	12	2
	No	88	98
Asthma	Yes	10	11
	No	90	89
COPD	Yes	15	15
	No	85	85
Hypertension	Yes	62	60
	No	38	40
HF	Yes	28	12
	No	72	88
CLD	Yes	1	1
	No	99	99
Malignant neoplasm	Yes	14	22
	No	86	78
Dementia	Yes	16	15
	No	84	85
PBC	Yes	14	14
	No	86	86
NLRL [D]	High	36	30
	Low	64	70
APTTL [E]	High	25	35
	Low	75	65

A: percentage population within the category. B: For simplicity, ethnicities other than the dominant white category were united as 'other'. C: Smoker and ex-smoker status were unified as '(ex-)smoker'. D: 'low' for NLR<8, 'high' for NLR > 8. E: 'low' for APTT < 24s, 'high' for APTT > 24s.
Note. IHD ischemic heart disease; TIA transient ischemic attack; COPD chronic obstructive pulmonary disease; HF heart failure; CLD chronic liver disease; PBC positive blood culture; NLRL neutrophils-lymphocytes ratio labelled; APTTL activated partial thromboplastin time labelled.

2.2. Data cleaning

A data cleaning process was considered to exclude entries with a high missingness rate. A 50% inclusion criterion was determined, and thus individuals with a missingness rate of more than 50% in their features and features missing in more than 50% of individuals were excluded from the analysis. As a result, 13 patients, four with and nine without DM, and two features, *ferritin* and *D-dimer*, did not meet the inclusion criteria. Thus, considering the 15 individuals who died from non-COVID-19 causes, a total of 28

individuals were excluded. As a result, 40 features from 477 participants, 149 with and 328 without DM, were used in the subsequent analysis.

2.3. Train test split

After cleaning the dataset, a stratified random sampling approach was employed to perform a 70-30 train test split, considering the unbalanced distribution of classes. For each cohort, 70% of death cases plus 70% of survival cases were selected at random and allocated as the training set, and the remaining 30% of death and survival cases were allocated as the testing set. Table 4 summarises the train test split results for the DM and non-DM cohorts. All model training and hyperparameter tuning operations were carried out on training sets only, with testing sets remaining unseen for evaluation and model interpretation analysis.

Table 4. Summary characteristics of the training set and testing set of DM (diabetes mellitus) and non-DM cohorts.

		DM cohort	Non-DM cohort
Training set	Dead	36	54
	Survived	68	175
	Total	104	229
Testing set	Dead	15	23
	Survived	30	76
	Total	45	99

2.4. Data preprocessing

2.4.1. Outliers treatment

The first preprocessing step considered was dealing with outliers to prevent models from being significantly influenced by extreme values of numerical features. Therefore, the winsorisation technique was employed to limit extreme values of numerical features to the lower and upper boundaries of the 5th and 95th percentiles of the training set, respectively.

2.4.2. Feature values transformation

The following preprocessing step was converting feature values to a format suitable for analysis by ML algorithms. Hence, numerical features were standardised by subtracting the average of the training set from each feature value and then scaling to unit variance by dividing the result by the standard deviation of the

training set. Additionally, categorical variables were transformed into numeric values using the one-hot-encoding technique. One dummy variable was obtained from two categories by dropping the first level. This curtailment may help avert the dummy variable trap by avoiding an unnecessary increase in the feature set size.

2.4.3. Missing values imputation

After converting feature values, missing values were replaced with predictions from k-nearest neighbour imputation, an algorithm compatible with both continuous and categorical features [24] as presented in the data used in this work. With five as the number of neighbours, for a given data point, the algorithm found the five most similar data points in the training set using non-missing values, and each missing value was filled with the average values of the five considered neighbours.

2.4.4. Oversampling

The final stage of preprocessing addressed two imbalance issues in the dataset. One imbalance condition was that, as shown in Table 4, in the training set of both cohorts, the number of survivors (68 for the DM cohort and 175 for the non-DM cohort) was considerably higher than the number of deaths (36 for the DM cohort and 54 for the non-DM cohort). This inequality may cause biased model learning towards the dominant class [25]. The other imbalance condition was that the training set of the non-DM cohort, at 229 entries, was considerably larger than that of the DM cohort, at 104 entries. This difference may result in models with performance commensurate with the size of training sets, making model comparisons less conclusive. Thus, the oversampling technique was deployed to address the concerns regarding imbalanced data. The oversampling increased the number of deaths and survivors in both training sets to 175, the maximum number of deaths and survivors in the original training sets (Table 4). Oversampling was performed using the SMOTE-NC algorithm, a well-suited technique for datasets with continuous and categorical features [26], such as the one used in this study. The testing sets were not oversampled; thereby, evaluation and interpretation analyses were conducted only on actual data.

2.5. Feature selection

A preliminary step in developing models for mortality risk assessment was to perform a feature selection on each cohort to reduce the input data size. Otherwise, the relatively large feature set size may cause the dimensionality curse during the model training process. For feature selection, we considered a voting system that could potentially provide further robustness compared to non-voting systems. To accomplish this, we wrapped the recursive feature elimination (RFE) technique around three different classifiers to create three voter systems. The three classifiers used in each voter system were logistic regression, gradient boosting, and AdaBoost. These algorithms have demonstrated broad capability and have been applied in COVID-19 research [27], [28].

In each voter system, features were ranked using the feature coefficient metric for logistic regression and feature importance metric for gradient boosting and the AdaBoost model, and RFE eliminated the variable that had the least contribution to predictions on the training set. This feature reduction cycle was repeated until RFE dropped half of the variables (commonly used configuration of the RFE function) in each voter system and shortlisted 20 out of the 40 features. The features shortlisted by at least two voters were finally considered for mortality risk assessments.

To fine-tune the hyperparameters of the three classifiers, we used the random search approach. A search space for possible hyperparameter values was defined. Then, after experimenting with 20 different randomly selected combinations of values within the search space, the one that provided the highest five-fold cross-validation accuracy on the training set was chosen. The details of the search spaces considered and the results of hyperparameter tunning are available in Appendix, Table A1.

2.6. Mortality risk assessment

Mortality risk assessments in this work consisted of three main parts; developing a mortality risk prediction model for each cohort, equipping the developed mortality risk prediction models with a model

agnostic framework, and developing a mortality risk stratification model for each cohort based on model interpretation outcomes.

2.6.1. Mortality risk prediction

After selecting predictive features for each cohort, a model was created to predict in-hospital COVID-19 mortality. An RF classifier was used to predict admission outcomes from selected features. This classification technique has been demonstrated to be effective in different fields, including COVID-19 risk assessment [29]. Hyperparameter tuning was performed with a similar approach explained in subsection 2.5 (for classifiers in the voting feature selection systems). The results are presented in Appendix, Table A1.

2.6.2. Model interpretation

Following the development of mortality risk prediction models, an extensive SHAP analysis was performed. Models' predictions on unseen testing data were initially interpreted globally, i.e., by explaining the aggregate effects of selected features on forming predictions across the entire training set. Afterwards, a local interpretation analysis was conducted on a subset of selected individuals, elaborating on the contribution of predictors in forming a specific prediction for each individual. This investigation increases the transparency of the analysis and enables localisation and comparison of the predictors' effects on forecasts for each instance.

2.6.3. Mortality risk stratification

Model interpretation analysis was followed by risk stratification investigations. To this end, first, each patient was represented with a vector containing SHAP values corresponding to the selected features. Then, the k-means algorithm was employed to divide patients of the test data into clusters based on their SHAP value vectors, a demarcation with potential utility in risk stratification practice. The k-means algorithm has been used in previous COVID-19 research [30], [31]. The algorithm partitions samples into groups of equal variance by minimising the inertia criterion. For selecting the number of clusters, values of 1 to 9 were

examined, and the one delivering the elbow point based on the inertia criterion across the entire training set was decided [32].

3. Results

This section presents results related to mortality risk prediction and stratification analysis.

3.1. Feature selection

From feature selection analysis, the predictors selected for the DM cohort were *frailty score*, *age*, *Hb* (haemoglobin), *platelets*, *Na* (sodium), *creatinine*, *eGFR* (estimated glomerular filtration rate), *ALPO4* (alkaline phosphates), *CRP* (c-reactive protein), *fibrinogen*, *sex*, *PT* (prothrombin time), *WCC* (white cell count), *neutrophils*, *lymphocytes*, *monocytes*, *ALT* (alanine transaminase), *smoking status*, *asthma*, *HF* (heart failure), *NLRL*, *APTTL*. On the other hand, the predictors selected for the non-DM cohort consisted of *frailty score*, *age*, *Hb*, *platelets*, *Na*, *creatinine*, *eGFR*, *ALPO4*, *CRP*, *fibrinogen*, *sex*, *PT*, *BMI* (body mass index), *monocytes*, *K* (potassium), *bilirubin*, *total protein*, *albumin*, *procalcitonin*, *PBC* (positive blood culture).

3.2. Mortality risk prediction

The developed RF classifiers to predict COVID-19 mortality were evaluated by measuring the prediction performance on the unseen testing sets. Four metrics were considered for evaluation analysis; accuracy, area under the curve (AUC), sensitivity, and specificity. These metrics have been broadly used in classification tasks. Also, these metrics have evidence supporting their applications in healthcare research [33], [34]. Table 5summarises evaluation results for mortality risk prediction models. As shown in the table, both models resulted in values of at least 80% for three of the four evaluation metrics (accuracy, AUC, and sensitivity).

Table 5. The evaluation result of the mortality prediction models for DM (diabetes mellitus) and non-DM cohort.

Evaluation metric	DM model	Non-DM model
Accuracy (%)	82	80
AUC (%)	80	84
Sensitivity (%)	80	91
Specificity (%)	55	56

Note. AUC area under the receiver operating characteristics curve

3.3. Global interpretation

The variable importance plots in Figure 1 list the most significant features for each model in descending order, according to their collective SHAP values. The length of each bar indicates the mean of absolute SHAP values for the relevant feature(s) across the entire testing set. For legibility and brevity, considering a maximum display of 10, the first nine most influential predictors alongside the aggregated impact of remaining predictors are displayed.

Based on the plots, *frailty score*, *age*, and *CRP* (*c-reactive protein*) were among the nine most predictive variables for both models. *NLRL* was the most predictive variable for the DM model, and *frailty score* and *Na* ranked second and third, respectively. On the other hand, *albumin*, *age*, and *eGFR* (*estimated glomerular filtration rate*) were the first three most predictive variables for the non-DM model.

The Frailty *score* was the second most important variable for the DM model and the fourth for the non-DM model. Therefore, this measure of underlying health status was more influential for the DM model than the non-DM model. Additionally, *albumin* was a critical variable for the non-DM cohort while not a predictive factor for the DM cohort. Further research may elicit this inconsistency also observed in previous work [9].

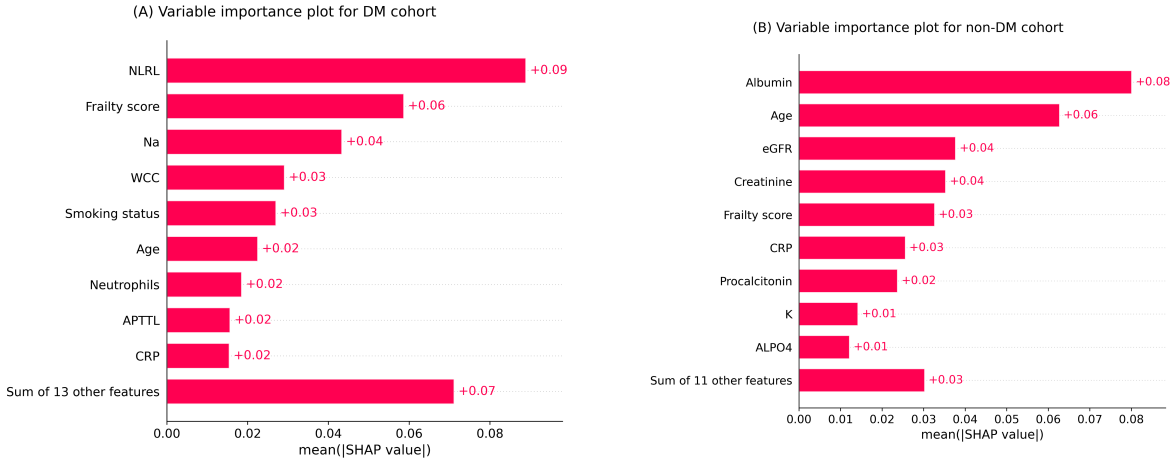

Figure 1. Feature importance plots for (A) DM (diabetes mellitus) cohort (B) non-DM cohort. The plots indicate a rank order for variables upon collective absolute SHAP values of the testing set. Note. NLRL neutrophils-lymphocytes ratio labelled; Na sodium; WCC white cell count; APTTL activated partial thromboplastin time labelled; CRP c-reactive protein; eGFR estimated glomerular filtration rate; K potassium; ALPO4 alkaline phosphates.

Figure 2 provides SHAP value plots as an alternative global interpretation schematic for mortality prediction models. These bee swarm plots express predictors' positive/negative associations with the target variable, in addition to their importance rank. Each point on the graphs corresponds to a sample from the testing set. The position on the x-axis indicates whether a particular feature value is associated with a higher or lower mortality prediction. The colours represent the relative values of variables. For numerical features, blue and red denote low and high values, respectively, while for encoded categorical features, these colours indicate 0 and 1, respectively. With similar explanations given for Figure 2, with a maximum display of 10, the first nine most influential predictors individually along with the remaining features together are shown.

The DM model's nine distinct features were all positively associated with mortality risk prediction, i.e., higher feature values were associated with positive SHAP values, while lower feature values were associated with negative SHAP values. On the other hand, for the non-DM model, age, *frailty score*, and *CRP* were positively associated with mortality risk prediction, whereas *albumin*, *eGFR*, and *K* were negatively associated with mortality risk prediction.

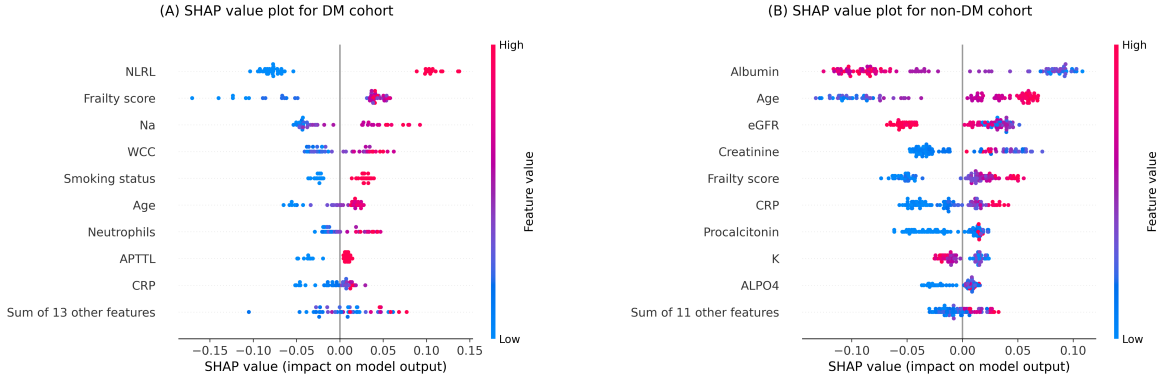

Figure 2. SHAP values plots of the testing set for (A) DM (diabetes mellitus) cohort (B) non-DM cohort. Each point signifies a patient in the testing set. The horizontal locations reflect the effect of features on the model's outputs for a particular individual. Colours indicate whether the variable is high (red) or low (blue) for a particular observation; for encoded categorical variables, blue and red denote 0 and 1, respectively. Note. NLRL neutrophils-lymphocytes ratio labelled; Na sodium; WCC white cell count; APTTL activated partial thromboplastin time labelled; CRP c-reactive protein; eGFR estimated glomerular filtration rate; K potassium; ALPO4 alkaline phosphates

3.4. Local interpretation

After presenting the results of the global interpretation analysis, this subsection presents examples of the outcomes of the local interpretation analysis. To this end, the results concerning a random death and survival case from each cohort are selected to present.

The waterfall plots in Figure 3 display the local interpretation results for a randomly selected individual with a death outcome example in each cohort. These plots show features' contributions to generating a specific prediction for a given instance. The size and direction of each arrow indicate the effect of a particular feature to shift the output from a base prediction (average prediction on the training set) towards a final prediction [18]. According to the figure, the mortality prediction models predicted a probability of death greater than 50% for both cases (DM: 50.2%, non-DM: 62.5%) and thus classified them in the death category.

Based on Figure 3A, *NLRL* was the most adverse feature for the DM instance, with *frailty score*, *age*, and *PT* being second to fourth, respectively. In contrast, in terms of protective impact, variable *Na* was ranked first, *smoking status* second, *fibrinogen* third, *neutrophils* fourth, and *WCC* (*white cell count*) fifth.

In comparison, the leading five predictors of death in the non-DM case were *age*, low *albumin*, *creatine*, *eGFR*, and *frailty score*, whereas *CRP*, *procalcitonin*, *bilirubin*, and *K* were the main features decreasing the prediction of death in this case.

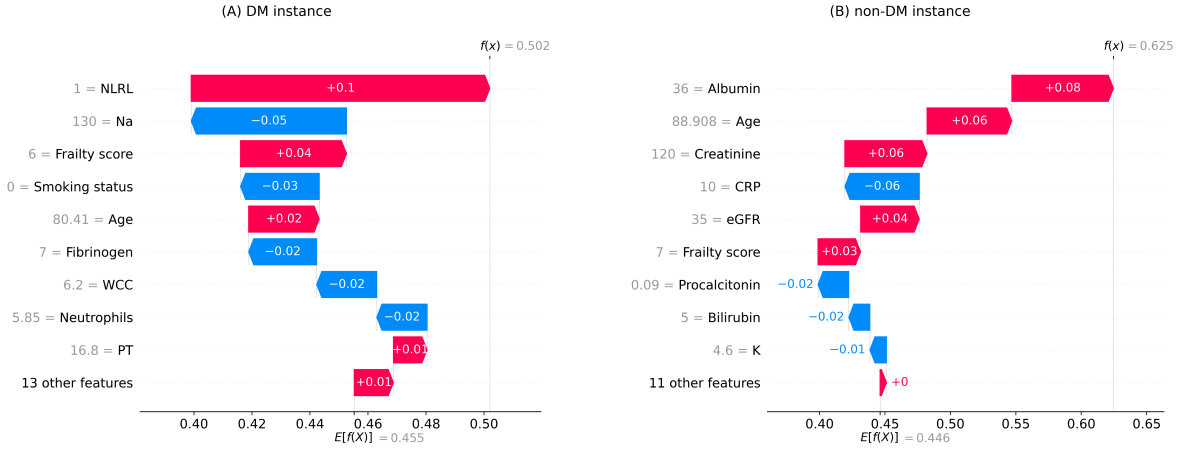

Figure 3. Local interpretation waterfall plots for an individual who died due to COVID-19 in the testing set of (A) DM (diabetes mellitus) cohort (B) non-DM cohort. The bottom of the plots starts at a base expectation under training data (E[f(x)]). Then, each row shows the contribution of its relevant feature to increase (red) or decrease (blue) the expectation value. The final model prediction value is indicated by f(x) in the end. Note. NLRL neutrophils-lymphocytes ratio labelled; Na sodium; WCC white cell count; PT prothrombin time; CRP c-reactive protein; eGFR estimated glomerular filtration rate; K potassium

Figure 4 illustrates the local interpretation results for two randomly selected instances with a survival outcome (one from each cohort). According to the plots, the mortality prediction models classified both cases in the survival category, predicting a mortality chance of less than 50% for both cases (DM: 27.6%, non-DM: 13.8%). The most protective features for the DM case were *NLRL*, *frailty score*, *age*, *APTTL*, and *Na*, whereas *WCC* and *smoking status* were the most adverse features for this instance. On the other hand, the primary protective variables for the non-DM case were *age*, *albumin, eGFR*, *CRP, creatine*, *ALPO4*, and K, whereas the primary adverse variables for this case were *frailty score* and *bilirubin*.

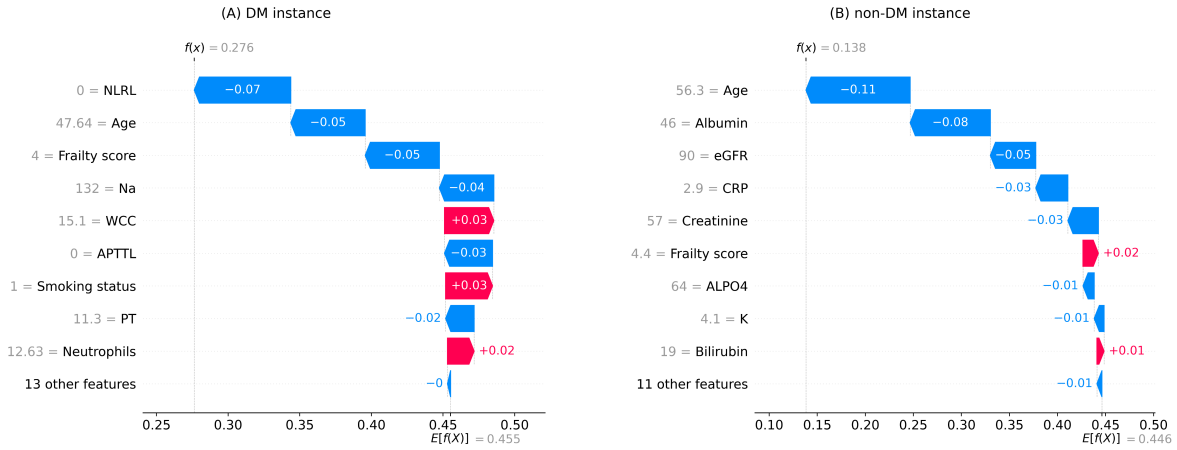

Figure 4. Local interpretation waterfall plot for an individual who survived COVID-19 in the testing set of (A) DM (diabetes mellitus) cohort (B) non-DM cohort. The bottom of the plots starts at a base expectation under training data (E[f(x)]). Then, each row shows the contribution of its relevant feature to increase (red) or decrease (blue) the expectation value. The final model prediction value is indicated by f(x) in the end. Note. NLRL neutrophils-lymphocytes ratio labelled; Na sodium; WCC white cell count; APTTL activated partial thromboplastin time labelled; PT prothrombin time; eGFR estimated glomerular filtration rate; CRP c-reactive protein; ALPO4 alkaline phosphates; K potassium

3.5. Mortality risk stratification

In this subsection, the results of the mortality risk stratification analysis are presented and discussed. Figure 5 presents the results of the elbow method analysis. According to the figure, three clusters were decided for both cohorts as it was an elbow point in both cases.

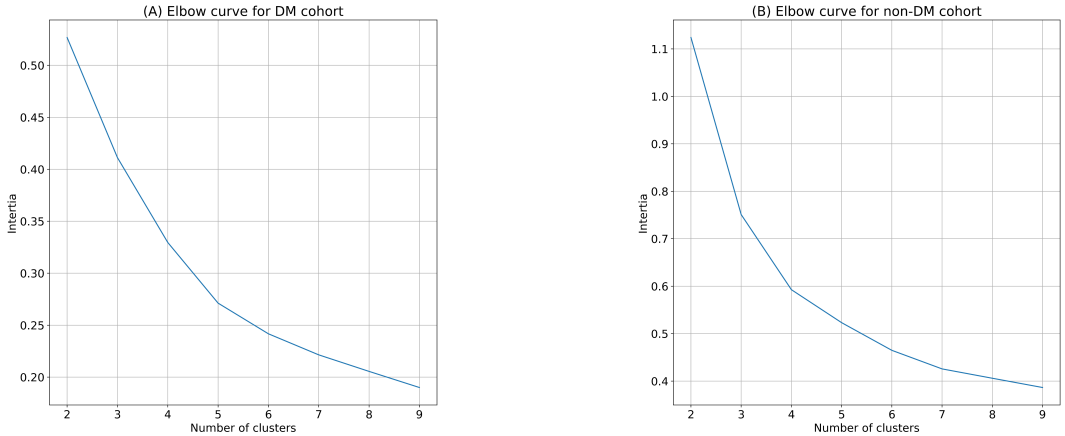

Figure 5. Elbow method graph to determine the optimal number of clusters for SHAP clustering on (A) DM (diabetes mellitus) cohort (B) non-DM cohort.

Table 6 shows the results of SHAP clustering, including the ratio of the cohorts population distributed among the clusters and the death rate within each cluster. The clustering technique has allocated patients of both cohorts into three categories with relatively low, moderate, and high mortality rates (clusters 1, 2, and 3, respectively).

Table 6. The results of clustering patients on their SHAP values for DM (diabetes mellitus) and non-DM cohorts.

	DM cohort		**Non-DM cohort**	
	Frequency [A] (%)	Mortality rate (%)	Frequency [A] (%)	Mortality rate (%)
Cluster 1	26.6	8	36.4	2.7
Cluster 2	44.4	20	32.3	28.1
Cluster 3	28.8	76	31.3	41.9

A: The percentage of the cohort placed in each cluster

4. Discussion

In general, clinical studies have demonstrated an association between most selected features in this research (presented in subsection 3.1) and COVID-19 complications [35]. More specifically, *smoking status*, *asthma*, and *HF* were among the features selected for the DM cohort, underlining the established increased risk of these preexisting factors for COVID-19 patients with DM [36]. Moreover, it is noteworthy that the selection of *NLRL* and *APTTL* for the DM cohort was consistent with the findings of previous work [9]. Such congruence with the literature implies the effectiveness of the feature selection analysis in laying a reliable foundation for the ensuing ML-based mortality risk assessments.

The evaluation results of the mortality risk prediction models (presented in subsection 3.2) emphasise the overall effectiveness of the analysis in predicting mortality risk for both cohorts. Moreover, the models' performance was comparable, enabling fair intercohort analogies. This comparable performance may imply that the oversampling process has effectively addressed the concerns regarding data group imbalances.

As illustrated in Figure 2A, overall, contributions of high *NLRLs* to increased mortality risk predictions were more than contributions of low *NLRLs* to decreased mortality risk predictions in the DM cohort. Conversely, low *frailty scores* contributed more to lower mortality risk predictions than high *frailty scores* did to higher mortality risk predictions. Similarly, among other high-impact variables for the DM cohort, *Na*, *WCC*, *smoking status*, and *neutrophils* contributed more to increased mortality risk predictions overall,

whereas age, *APTTL* and *CRP* contributed more to decreased mortality risk predictions. Likewise, it can be implied from Figure 2B that creatine had a greater adverse than protective impact in the DM cohort, while *age* and *procalcitonin* had a greater protective than adverse impact overall. However, the differences between the protective and adverse contributions were inconclusive for other high-impact predictors of the non-DM cohort. Such analysis could help compare the protective versus adverse impact of features. For instance, since high *NLRL* showed a stronger adverse impact compared to the protective impact from low *NLRL* in the DM cohort, it could be inferred that this feature was overall a stronger adverse risk factor rather than a protective factor in this cohort.

Of note, older *ages* were associated with increased mortality risk predictions in both cohorts (positive SHAP values in Figure 2), but this effect was more marked in the DM cohort. One possible explanation could be that the chance of having co-existing features that increased mortality predictions occurred more often in older DM cases than non-DM cases.

Overall, local interpretation results (presented in subsection 3.4) show how explaining the model's output for an individual can differ from explaining the model's output globally across the cohort. This evidence stresses the advantages of individualised risk explanations over generic risk descriptions.

The SHAP clustering outcomes (presented in subsection 3.5) are in line with real-world risk stratification requirements, namely for applications in triage systems, where the aim is to allocate patients into predefined categories with different risk grades [37]. This evidence supports the potential capability of SHAP clustering in practical COVID-19 mortality risk stratification.

Table 7 summarises some statistical characteristics of features within the three formed clusters for each cohort to explore patterns apart from the frequency and mortality rate presented in Table 6. For conciseness, only the three most predictive variables, according to Figure 1, are investigated for each cohort. Based on the table, one noteworthy intercluster pattern for the DM cohort was that all patients in Cluster 3 had a high *NLRL*. Another marked pattern was that patients in Cluster 1 had a considerably lower average *frailty score* than patients in Clusters 2 and 3. On the other hand, for the non-DM cohort, a significant pattern was that

the average *albumin* for patients in Cluster 3 was considerably higher than that in Clusters 2 and 1. Also, the average *age* in Cluster 1 was considerably lower than that in Clusters 2 and 3. Finally, there was a decrease in the average *eGFR* from Clusters 1 towards 3.

Table 7. Characteristics of the three most predictive features of DM (diabetes mellitus) and non-DM cohort sin the three clusters created on SHAP values.

		Cluster 1	Cluster 2	Cluster 3
DM cohort	High NLRL ratio	16%	0%	100%
	Average frailty score	3.4	6.1	5.8
	Average Na	134.2	134.9	135.9
Non-DM cohort	Average albumin	40.7	41.7	34.9
	Age (year)	51.7	79.1	82.4
	eGFR	79.5	60.9	51.9

Note. Note. NLRL neutrophils-lymphocytes ratio labelled; Na sodium; eGFR estimated glomerular filtration rate

5. Summary and conclusion

Fatality risk assessments were conducted in parallel for cohorts of COVID-19 patients with and without DM. First, using the RF algorithm, a model was developed for each cohort to predict in-hospital death due to COVID-19 from admission data. The evaluation results showed that the generated mortality prediction models provided comparable performances. The models were then interpreted globally and locally through SHAP. The global interpretations delineated distinct characteristics of each cohort, such as their features' relative importance and positive/negative association with the predicted probability of death. Finally, the k-means algorithm was implemented on the SHAP values to generate clusters pertaining to risk stratification practice. Clustering on SHAP values formed three clusters with relatively low, moderate, and high mortality rates, highlighting the potential functionality of SHAP clustering for COVID-19 risk stratification.

Overall, these ML algorithms offered additional results beyond that provided by standard statistical approaches, such as the rate and order of the most important predictors, global and local interpretation of outcomes, and risk stratification based on interpretation analysis. In conclusion, this article contributes to bridging the gap between advanced ML techniques and routinely collected clinical data in a critical field of medicine. The research findings encourage further exploitation of ML models framed with interpretation analysis in observational studies of COVID-19 patients with and without DM. These advanced data analysis

tools, underused previously in this field, have been shown to facilitate knowledge discovery and inferences. Consequently, implementing similar methodologies on recent COVID-19 datasets is recommended for future work.

Although this study successfully highlights the effectiveness of interpretable machine learning models in assessing COVID-19 mortality risk among patients with and without diabetes mellitus, several avenues for future research remain. Expanding the research to larger and more diverse cohorts would strengthen the generalisability of the findings and ensure that the models perform effectively across different populations. Additionally, investigating the practical aspects of implementing these models in clinical settings is crucial. This includes examining data integration with electronic health records, clinician training, and real-time decision support systems to maximise the practical utility of the models in everyday medical practice. Future work should focus on validating the models in real-world environments to ensure that they deliver meaningful insights and enhance patient care.

Acknowledgement

We would like to thank Ahmed Iqbal, Marni Greig, Muhammad Fahad Arshad, Thomas H Julian, and Sher Ee Tan for their efforts in collecting the clinical data used in this paper.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix

Table A1. Summary results of the randomised hyperparameter tunning for the voter and final mortality prediction models

Model	Hyperparameter	Search space	Selected hyperparameter	
			DM cohort	Non-DM cohort
LR	regularisation strength	{0, .01, 0.02, …,1}	0.02	0.04
	class weight	{0, 1, …,10}	3	5
	maximum number of iterationsfelbo	{1000, 2000, …,10000}	40000	6000
GB	learning rates	{0.01, 0.02, …,1}	0.03	0.06
	number of boosting stages	{20, 40, …,200}	50	160
	minimum number of samples required to split an internal node	{2, 4, …,8}	6	4
	minimum number of samples required to be at a leaf node	{2, 4, …,8}	4	4
	maximum depth of the individual estimators	{1, 2, …,10}	5	4
AB	maximum number of estimators at which boosting is terminated	{10, 20, …,100}	50	70
	learning rates	{0.01, 0.02, …,1}	0.04	0.06
RF	number of trees	{50, 100, …,500}	200	400
	maximum depth of the tree	{1, 2, …,10}	4	6
	minimum number of samples required to split an internal node,	{2, 4, …,8}	6	6
	minimum number of samples required to be at a leaf node	{2, 4, …,8}	4	4
	maximum number of leaf nodes	{2, 4, …,8}	6	8
	minimum impurity decrease	{0, 0.01}	0	0
	cost complexity pruning factor	{0.01, 0.02, …,0.10}	0.06	0.04
	minimum weighted fraction of the sum total of weights	{0.01, 0.02, …0.05}	0.03	0.04

References

[1] M. Wargny *et al.*, "Predictors of hospital discharge and mortality in patients with diabetes and COVID-19: updated results from the nationwide CORONADO study," *Diabetologia*, vol. 64, no. 4, pp. 778–794, Apr. 2021, doi: 10.1007/s00125-020-05351-w.

[2] C. Wu *et al.*, "Risk Factors Associated with Acute Respiratory Distress Syndrome and Death in Patients with Coronavirus Disease 2019 Pneumonia in Wuhan, China," *JAMA Intern. Med.*, vol. 180, no. 7, pp. 934–943, 2020, doi: 10.1001/jamainternmed.2020.0994.

[3] G. Onder, G. Rezza, and S. Brusaferro, "Case-Fatality Rate and Characteristics of Patients Dying in Relation to COVID-19 in Italy," *JAMA*, vol. 323, no. 18, pp. 1775–1776, 2020, doi: 10.1001/jama.2020.4683.

[4] G. Corona *et al.*, "Diabetes is most important cause for mortality in COVID-19 hospitalized patients: Systematic review and meta-analysis.," *Rev. Endocr. Metab. Disord.*, vol. 22, no. 2, pp. 275–296, 2021, doi: 10.1007/s11154-021-09630-8.

[5] H. Sourij *et al.*, "COVID-19 fatality prediction in people with diabetes and prediabetes using a simple score upon hospital admission," *Diabetes, Obes. Metab.*, vol. 23, no. 2, pp. 589–598, 2021, doi: 10.1111/dom.14256.

[6] S. Ciardullo *et al.*, "Impact of diabetes on COVID-19-related in-hospital mortality: a retrospective study from Northern Italy," *J. Endocrinol. Invest.*, vol. 44, no. 4, pp. 843–850, 2021, doi: 10.1007/s40618-020-01382-7.

[7] H. Shah, M. S. H. Khan, N. V Dhurandhar, and V. Hegde, "The triumvirate: why hypertension, obesity, and diabetes are risk factors for adverse effects in patients with COVID-19," *Acta Diabetol.*, vol. 58, no. 11, pp. 831--843, 2021, doi: https://doi.org/10.1007/s00592-020-01636-z.

[8] N. Holman *et al.*, "Risk factors for COVID-19-related mortality in people with type 1 and type 2 diabetes in England: a population-based cohort study," *Lancet Diabetes Endocrinol.*, vol. 8, no. 10, pp. 823–833, 2020, doi: 10.1016/S2213-8587(20)30271-0.

[9] A. Iqbal, M. Arshad, T. Julian, S. Tan, M. Greig, and J. Elliott, "Higher admission activated partial thromboplastin time, neutrophil-lymphocyte ratio, serum sodium, and anticoagulant use predict in-hospital covid-19 mortality in people with diabetes: Findings from two university hospitals in the UK," *Diabet. Med.*, vol. 178, no. 108955, pp. 1–12, 2021, doi: 10.1016/j.diabres.2021.108955.

[10] S. J. McGurnaghan *et al.*, "Risks of and risk factors for COVID-19 disease in people with diabetes: a cohort study of the total population of Scotland," *Lancet Diabetes Endocrinol.*, vol. 9, no. 2, pp. 82–93, 2021, doi: 10.1016/S2213-8587(20)30405-8.

[11] D. Bzdok, N. Altman, and M. Krzywinski, "Statistics versus machine learning," *Nat. Methods*, vol. 15, no. 4, pp. 233–234, 2018, doi: 10.1038/nmeth.4642.

[12] S. Lalmuanawma, J. Hussain, and L. Chhakchhuak, "Applications of machine learning and artificial intelligence for Covid-19 (SARS-CoV-2) pandemic: A review," *Chaos, Solitons and Fractals*, vol. 139, no. 110059, pp. 1–6, 2020, doi: 10.1016/j.chaos.2020.110059.

[13] O. Shahid *et al.*, "Machine learning research towards combating COVID-19: Virus detection, spread prevention, and medical assistance," *J. Biomed. Inform.*, vol. 117, no. 103751, pp. 1–16, 2021, doi: 10.1016/j.jbi.2021.103751.

[14] A. Alimadadi, S. Aryal, I. Manandhar, P. B. Munroe, B. Joe, and X. Cheng, "Artificial intelligence and machine learning to fight covid-19," *Physiol. Genomics*, vol. 52, no. 4, pp. 200–202, 2020, doi: 10.1152/physiolgenomics.00029.2020.

[15] Y. Gao *et al.*, "Machine learning based early warning system enables accurate mortality risk prediction for COVID-19," *Nat. Commun.*, vol. 11, no. 5033, pp. 1–10, 2020, doi: 10.1038/s41467-020-18684-2.

[16] E. Mauer *et al.*, "A predictive model of clinical deterioration among hospitalized COVID-19 patients by harnessing hospital course trajectories," *J. Biomed. Inform.*, vol. 118, no. 103794, pp. 1–12, 2021, doi: 10.1016/j.jbi.2021.103794.

[17] A. McGovern *et al.*, "Making the black box more transparent: Understanding the physical implications

of machine learning," *Bull. Am. Meteorol. Soc.*, vol. 100, no. 11, pp. 2175–2199, 2019, doi: 10.1175/BAMS-D-18-0195.1.

[18] S. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *31th Conference on Neural Information Processing Systems*, 2017, pp. 4765–4774.

[19] J. C. Forte *et al.*, "Identifying and characterizing high-risk clusters in a heterogeneous ICU population with deep embedded clustering," *Sci. Rep.*, vol. 11, no. 12109, pp. 1–12, 2021, doi: 10.1038/s41598-021-91297-x.

[20] Q. A. Hathaway *et al.*, "Machine-learning to stratify diabetic patients using novel cardiac biomarkers and integrative genomics," *Cardiovasc. Diabetol.*, vol. 18, no. 78, pp. 1–16, 2019, doi: 10.1186/s12933-019-0879-0.

[21] A. D. Haimovich *et al.*, "Development and Validation of the Quick COVID-19 Severity Index: A Prognostic Tool for Early Clinical Decompensation," *Ann. Emerg. Med.*, vol. 76, no. 4, pp. 442–453, 2020, doi: 10.1016/j.annemergmed.2020.07.022.

[22] B. Zheng *et al.*, "An Interpretable Model-Based Prediction of Severity and Crucial Factors in Patients with COVID-19.," *Biomed Res. Int.*, vol. 2021, no. 8840835, pp. 1–9, 2021, doi: 10.1155/2021/8840835.

[23] P. Pan *et al.*, "Prognostic assessment of COVID-19 in the intensive care unit by machine learning methods: Model development and validation," *J. Med. Internet Res.*, vol. 22, no. 11, pp. 1–16, 2020, doi: 10.2196/23128.

[24] P. Jonsson and C. Wohlin, "An evaluation of k-nearest neighbour imputation using likert data," in *10th International Symposium on Software Metrics, 2004. Proceedings.*, 2004, pp. 108–118, doi: 10.1109/METRIC.2004.1357895.

[25] H. Ali, M. N. M. Salleh, R. Saedudin, K. Hussain, and M. F. Mushtaq, "Imbalance class problems in data mining: A review," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 14, no. 3, pp. 1560–1571, 2019, doi: 10.11591/ijeecs.v14.i3.pp1560-1571.

[26] N. V Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002, doi: 10.1613/jair.953.

[27] A. K. Das, S. Mishra, and S. S. Gopalan, "Predicting CoVID-19 community mortality risk using machine learning and development of an online prognostic tool," *PeerJ*, vol. 8, no. e10083, pp. 1–12, 2020, doi: 10.7717/peerj.10083.

[28] A. M. U. D. Khanday, S. T. Rabani, Q. R. Khan, N. Rouf, and M. M. U. Din, "Machine learning based approaches for detecting COVID-19 using clinical text data," *Int. J. Inf. Technol.*, vol. 12, no. 3, pp. 731–739, 2020, doi: 10.1007/s41870-020-00495-9.

[29] J. Wang *et al.*, "A descriptive study of random forest algorithm for predicting COVID-19 patients outcome," *PeerJ*, vol. 8, no. e9945, pp. 1–19, 2020, doi: 10.7717/peerj.9945.

[30] D. Abdullah, S. Susilo, A. S. Ahmar, R. Rusli, and R. Hidayat, "The application of K-means clustering for province clustering in Indonesia of the risk of the COVID-19 pandemic based on COVID-19 data," *Qual. Quant.*, pp. 1–9, 2021, doi: 10.1007/s11135-021-01176-.

[31] J. Hutagalung, N. L. W. S. R. Ginantra, G. W. Bhawika, W. G. S. Parwita, A. Wanto, and P. D. Panjaitan, "COVID-19 Cases and Deaths in Southeast Asia Clustering using K-Means Algorithm," *J. Phys. Conf. Ser.*, vol. 1783, no. 12027, 2021, doi: 10.1088/1742-6596/1783/1/012027.

[32] M. A. Syakur, B. K. Khotimah, E. M. S. Rochman, and B. D. Satoto, "Integration k-means clustering method and elbow method for identification of the best customer profile cluster," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 336, no. 1, p. 012017, 2018, doi: 10.1088/1757-899X/336/1/012017.

[33] S. Simons, D. Abasolo, and J. Escudero, "Classification of Alzheimer's disease from quadratic sample entropy of electroencephalogram," *Healthc. Technol. lLtters*, vol. 2, no. 3, pp. 70–73, 2015, doi: 0.1049/htl.2014.0106.

[34] J. M. Ahn, S. Kim, K.-S. Ahn, S.-H. Cho, K. B. Lee, and U. S. Kim, "A deep learning model for the detection of both advanced and early glaucoma using fundus photography," *PLoS One*, vol. 13, no. 11, p. e0207982, 2018, doi: 10.1371/journal.pone.0207982.

[35] D. Wolff, S. Nee, N. S. Hickey, and M. Marschollek, "Risk factors for Covid-19 severity and fatality: a structured literature review," *Infection*, vol. 49, no. 1, pp. 15–28, 2021, doi: 10.1007/s15010-020-01509-1.

[36] G. Targher *et al.*, "Patients with diabetes are at higher risk for severe illness from COVID-19," *Diabetes Metab.*, vol. 46, no. 4, pp. 335–337, 2020, doi: 10.1016/j.diabet.2020.05.001.

[37] S. R. Knight *et al.*, "Risk stratification of patients admitted to hospital with covid-19 using the ISARIC WHO Clinical Characterisation Protocol: Development and validation of the 4C Mortality Score," *BMJ*, vol. 370, pp. 1–13, 2020, doi: 10.1136/bmj.m3339.

Publication 6.

Interpretable Machine Learning for Inpatient COVID-19 Mortality Risk Assessments: Diabetes Mellitus Exclusive Interplay[6]

Abstract. People with diabetes mellitus (DM) are at elevated risk of in-hospital mortality from coronavirus disease-2019 (COVID-19). This vulnerability has spurred efforts to pinpoint distinctive characteristics of COVID-19 patients with DM. In this context, the present article develops ML models equipped with interpretation modules for inpatient mortality risk assessments of COVID-19 patients with DM. To this end, a cohort of 156 hospitalised COVID-19 patients with pre-existing DM is studied. For creating risk assessment platforms, this work explores a pool of historical, on-admission, and during-admission data that are DM-related or, according to preliminary investigations, are exclusively attributed to the COVID-19 susceptibility of DM patients. First, a set of careful pre-modelling steps are executed on the clinical data, including cleaning, pre-processing, subdivision, and feature elimination. Subsequently, standard machine learning (ML) modelling analysis is performed on the cured data. Initially, a classifier is tasked with forecasting COVID-19 fatality from selected features. The model undergoes thorough evaluation analysis. The results achieved substantiate the efficacy of the undertaken data curation and modelling steps. Afterwards, SHapley Additive exPlanations (SHAP) technique is assigned to interpret the generated mortality risk prediction model by rating the predictors' global and local influence on the model's outputs. These interpretations advance the comprehensibility of the analysis by explaining the formation of outcomes and, in this way, foster the adoption of the proposed methodologies. Next, a clustering algorithm

---

[6] This articles was published in Sensors 22 (2022). Authors: **H. Khadem**, H. Nemat, J. Elliott, M. Benaissa.

demarcates patients into four separate groups based on their SHAP values, providing a practical risk stratification method. Finally, a re-evaluation analysis is performed to verify the robustness of the proposed framework.

Keywords. COVID-19; Diabetes Mellitus; Machine Learning; SHAP

1. Introduction

Shortly after the outbreak of coronavirus disease-2019 (COVID-19), pre-existing diabetes mellitus (DM) was recognised as a risk factor for the new disease [1,2]. Subsequently, extensive research has been underway to study this vulnerability [3]. For instance, adopting logistic regression (LR) analysis, Sourij et al. investigated the prognostic prediction in hospitalised COVID-19 patients with DM. They also offered a simple yet effective score for forecasting the risk of fatal outcomes from age and on-admission values of arterial occlusive disease, c-reactive protein (CRP), estimated glomerular filtration rate, and aspartate aminotransferase [4].

DM comorbidity was later declared a leading cause of in-hospital COVID-19 mortality in some studies [5]. As an example, in a uni-centre retrospective study, Ciardullo et al. deployed LR to perform death prediction analysis for 373 hospitalised COVID-19 patients with DM from diabetes status, comorbid conditions, and laboratory data. Based on the results achieved, the authors affirmed DM as an independent culprit for in-hospital COVID-19 mortality [6].

Although the COVID-19 susceptibility of DM patients has been well documented, explaining this vulnerability remains a challenge [3,7]. The use of explainable machine learning (ML) is one strategy to contribute to addressing this challenge [8–11].

In general, ML algorithms carry excellent potency in discovering intricate correlated interactions [8,12]. These tools have found practical implementation in COVID-19 research. As a representative, using machine learning techniques, Kar et ll. studied 63 clinical and laboratory factors in relation to 1393 subjects

hospitalised for COVID-19 to forecast the probability of mortality at 7 and 28 days. As a result, they generated an effective bespoke death risk score [13].

Integrating underlying ML pipelines with model interpretation frameworks promotes the transparency of the analysis, offsetting the block-box nature of plain ML algorithms [14–16]. SHapley Additive exPlanations (SHAP) is an exemplar of elaborate ML explainability techniques [17,18]. SHAP employs the classical notion of Shapley values from cooperative game theory to measure the contribution of input data in forming a given output by the model [19]. The measured SHAP values for a particular input feature indicate the deviation from the average prediction when conditioned on that feature [18].

SHAP analysis has seen promising applications in COVID-19 risk assessment research [19–21]. For example, Pan et al. designed ML models dressed with SHAP analysis for COVID-19 prognosis assessment in individuals hospitalised in intensive care units [20].

In a recent publication, we developed machine learning pipelines incorporated with interpretation components for mortality risk prediction and stratification in hospitalised COVID-19 patients with and without DM in parallel [22]. For this purpose, a set of features collected at the point of hospital admission for both groups were investigated. Consequently, the generated risk assessment models possessed potential application to triage systems for both groups and enabled inter-cohort comparative analysis. In the original dataset, though, there existed a pool of historical, on-admission, and inpatient variables collated only for the DM group. These features were either DM-relevant or the primary investigations did not persuade the clinical data acquisition team to collate for the non-DM cohort. Due to the significance of the topic and the value of further knowledge discovery in this field, the current sequel study is conducted to create risk assessment platforms comparable to those in the earlier study for the same DM cohort by scrutinising only the abovementioned DM-exclusive data opted out of the former investigation.

First, the clinical data are cleansed and prepared for formal ML modelling analysis. After that, an ML model is constructed for inpatient fatality risk assessments. In-depth evaluation and interpretation analyses are performed on the model, and the results obtained are discussed in detail. This follow-on work initially

recruits similar main methods contrived in the prior paper, focusing on new findings, discussions, and applications. This homogeneity facilitates the analogical study of the two relevant articles. Next, some compartments in the primary skeleton of the pipelines are replaced with new units, and the investigations are re-conducted. This complementary analysis further inspects the robustness of the infrastructure utilised in the two studies by a side-by-side comparison of the new and old outcomes.

The remainder of the paper is organised as follows. The clinical data utilised in this work are outlined in Section 2. In Section 3, data pre-treatment steps undertaken before the conventional ML modelling analysis are explained. Section 4 describes the primary methodologies implemented for mortality risk assessments. The results achieved and the associated discussion is represented in Section 5. Section 6 reports further stability investigations on the proposed work frames. Finally, Section 7 summarises and concludes the work.

2. Material

The source of the clinical data used in this paper is the dataset primarily described in [23]. The present research explores demographic, clinical, and laboratory data from 156 individuals in the main dataset with confirmed COVID-19 and DM comorbidity. All these participants were admitted to Sheffield Teaching Hospitals (Sheffield, United Kingdom) between 29 February 2020 and 01 May 2020. Of the 156 patients, 103 survived, and 51 died due to COVID-19; the other three died due to causes other than COVID-19, according to their death certificates. Tables 1 and 2 summarise the statistical characteristics of the data used in this work. Table 1 includes categorical variables' information encompassing the name of categories and the number of recorded data in each category. In Table 2, the mean and standard deviation (SD) of numerical variables are presented alongside the frequency of records for each feature.

Table 1. A summary of properties of the categorical clinical data used in this article. For each feature, the categories' names and the number of patients with recorded data in each category are given.

Feature	Category	Count	Feature	Category	Count
CKD	Yes	131	PC-Dizziness	Yes	7
	No	21		No	145
Radiology-RPLD	Yes	12	PC-Headache	Yes	5
	No	132		No	147
Radiology-Consolidation on report	Yes	99	PC-Hyperglycaemia	Yes	12
	No	46		No	140
Radiology-Worsening consolidation	Yes	34	DM type	Type 1	12
	No	25		Type 2	140
Radiology-CT chest or CTPA	Yes	5	DM complications	Yes	127
	No	143		No	26
Radiology-PE	Yes	1	PVD	Yes	47
	No	5		No	105
Diabetes autoantibodies	Yes	1	Peripheral neuropathy	Yes	67
	No	152		No	84
PC-Fever	Yes	64	Background retinopathy	Yes	73
	No	88		No	64
PC-Cough	Yes	81	Preproliferative retinopathy	Yes	15
	No	72		No	122
PC-SOB	Yes	66	Proliferative retinopathy	Yes	17
	No	86		No	120
PC-Chest pain	Yes	11	Previous foot ulcer	Yes	28
	No	141		No	124
PC-Abdominal pain	Yes	8	Active foot ulceration	Yes	8
	No	144		No	144
PC-Diarrhoea	Yes	21	VRII treatment during admission	Yes	23
	No	131		No	129
PC-Myalgia	Yes	16	DNAR	Yes	109
	No	136		No	41

Note. CKD: chronic kidney disease; CT: computed tomography; CTPA: computed tomography pulmonary angiogram; DNAR: do not attempt resuscitation; PC: presenting complaint; PE: pulmonary embolism; PVD: peripheral vascular disease; RPLD: reported pre-existing lung disease; SOB: shortness of breath; VRII: variable rate intravenous insulin infusion.

Table 2. A summary of properties of the numerical clinical data used in this article. For each feature, mean and standard deviation, together with the number of patients that a value is recorded for the feature, are given.

Feature	Mean ± SD	Count	Feature	Mean ± SD	Count
BGV-pH (mmol/l)	7.39 ± 0.09	80	LV-ALPO4 (g/L)	87.27 ± 49.08	100
BGV-HCO3 (mmol/l)	22.43 ± 4.33	79	HV-ALPO4 (g/L)	134.35 ± 87.84	100
BGV-Lactate (mmol/l)	2.09 ± 1.71	80	LAYBA-Albumin (g/L)	39.82 ± 5.17	149
BGV-Na (mmol/l)	134.83 ± 5.45	80	LV-Albumin (g/L)	30.15 ± 4.91	104
BGV-K (mmol/l)	4.13 ± 0.86	80	HV-Albumin (g/L)	37.4 ± 4.2	104
BGV-Cl (mmol/l)	99.65 ± 10.91	80	LV-CRP (mg/dL)	49.37 ± 57.93	133
BGV-Anion Gap (mmol/l)	16.86 ± 10.44	79	HV-CRP (mg/dL)	165.89 ± 124.04	133
LAYBA-HbA1c (mmol/mol)	61.34 ± 17	144	LV-Procalcitonin (µg/L)	0.58 ± 1.04	68
FI-HbA1c (mmol/mol)	68.01 ± 19.52	74	HV-Procalcitonin (µg/L)	3.04 ± 11.98	68
LAYBA-Hb (g/l)	123.9 ± 20.89	146	LV-Ferritin (µg/L)	1019.87 ± 1363.47	45
LV-Hb (g/l)	106.5 ± 20.22	140	HV-Ferritin (µg/L)	1596.16 ± 2861.80	45
HV-Hb (g/l)	125.91 ± 18.29	140	OA-Troponin (ng/L)	49.07 ± 70.64	25
LAYBA-WCC (g/l)	8.09 ± 2.95	147	LV-Troponin (ng/L)	90.75 ± 224.18	32
LV-WCC (g/l)	6.03 ± 2.84	140	HV-Troponin (ng/L)	108.94 ± 237.01	32
HV-WCC (g/l)	11.01 ± 5.69	140	LAYBA-UACR	35.07 ± 76.42	62
LAYBA- NEUT (109/l)	5.35 ± 2.47	147	LAYBA-Vitamin D (ng/mL)	44.25 ± 25	44
LV- NEUT (109/l)	4.24 ± 2.43	140	LAYBA-PT (s)	12.26 ± 4.12	92
HV- NEUT (109/l)	8.72 ± 5.05	140	LV-PT (s)	12.08 ± 2.46	65
LAYBA-LYM (109/l)	1.76 ± 1.01	147	HV-PT (s)	16.54 ± 14.43	65
LV- LYM (109/l)	0.93 ± 1.41	140	LAYBA-APTT (s)	27.18 ± 5.48	93
HV- LYM (109/l)	1.66 ± 1.93	140	LV-APTT (s)	26.41 ± 4.77	66
LaV- LYM (109/l)	1.34 ± 1.45	151	HV-APTT (s)	33.27 ± 11.81	66
LAYBA-MN (109/l)	0.69 ± 0.32	147	LAYBA-Fibrinogen (g/L)	5.14 ± 1.42	94
OA-MN (109/l)	0.68 ± 0.41	148	LV-Fibrinogen (g/L)	5.43 ± 1.24	66
LV-MN (109/l)	0.38 ± 0.24	140	HV-Fibrinogen (g/L)	6.44 ± 1.1	66
HV-MN (109/l)	0.87 ± 0.46	140	LV-D-dimer (µg/L)	5313.47 ± 10366.62	15
LAYBA-Platelets (1/mL)	253.84 ± 102.75	147	HV-D-dimer (µg/L)	5397.80 ± 10339.12	15
LV-Platelets (1/mL)	196.06 ± 82.68	140	FI-BGL (mmol/L)	10.63 ± 5.84	152
HV-Platelets (1/mL)	332.41 ± 158.11	140	FI-Ketones (mmol/L)	1.05 ± 1.71	54
LAYBA-Na (mmol/l)	138.21 ± 3.45	150	Preadmission-RR (1/min)	23.89 ± 8.04	150
LV-Na (mmol/l)	133.49 ± 4.56	142	Preadmission-Saturations (%)	90.5 ± 9.96	142
HV-Na (mmol/l)	140.15 ± 5.77	142	Preadmission-Temperature (°C)	37.47 ± 1.21	149
LAYBA-K (mmol/l)	4.56 ± 0.54	150	Preadmission-SBP (mmHg)	139.27 ± 24.98	150
LV-K (mmol/l)	3.86 ± 0.55	140	Preadmission-DBP (mmHg)	75.56 ± 14.14	150
HV-K (mmol/l)	4.94 ± 0.86	140	Preadmission-Pulse (1/min)	90.65 ± 24.15	150
LAYBA-Urea (mmol/l)	8.85 ± 5.45	150	DM duration (years)	14.15 ± 10.8	145
LV-Urea (mmol/l)	7.75 ± 5.5	142	FI-RR (1/min)	26.26 ± 7.88	152
HV-Urea (mmol/l)	14.51 ± 9.28	142	FI-Saturations (%)	91.94 ± 6.24	152
LAYBA-Creatinine (µmol/l)	140.88 ± 125.46	150	FI-FiO2 (%)	47.16 ± 22.41	103
LV-Creatinine (µmol/l)	142.98 ± 156.87	142	FI-Temperature (°C)	37.1 ± 1.4	152
HV-Creatinine (µmol/l)	223.4 ± 244.63	142	FI-SBP (mmHg)	127.74 ± 31.82	152
eGFR	2.62 ± 1.1	152	FI-Pulse (1/min)	94.47 ± 22.33	152
LAYBA-Bilirubin (µmol/l)	7.55 ± 4.08	149	HR-O2 (%)	55.46 ± 22.85	124
LV-Bilirubin (µmol/l)	6.73 ± 3.93	101	LV-BGL (mmol/L)	4.92 ± 2.34	152
HV-Bilirubin (µmol/l)	12.86 ± 11.02	101	HV-BGL (mmol/L)	16.51 ± 6.65	152
LAYBA-ALT (u/l)	20.37 ± 13.93	142	Average BGL (mmol/L)	9.55 ± 3.02	152
LV-ALT (u/l)	29.38 ± 41.64	68	RBGLR below 3 mmol/L	0.01 ± 0.03	151
HV-ALT (u/l)	76.91 ± 160.6	68	RBGLR 4—10 mmol/L	0.61 ± 0.3	151
LAYBA-TP(g/l)	68.75 ± 7.36	149	RBGLR 10.1—14 mmol/L	0.22 ± 0.19	151
LV-TP (g/l)	59.62 ± 7.51	102	RBGLR 14—21 mmol/L	0.11 ± 0.15	151
HV-TP (g/l)	70.02 ± 6.6	102	RBGLR 21.0—27.8 mmol/L	0.03 ± 0.07	151
LAYBA-ALPO4 (g/l)	99.43 ± 45	149	RBGLR above 27.8 mmol/L	0.01 ± 0.03	151

Note. ALT: alanine transaminase; ALPO4: alkaline phosphatase; APTT: activated partial thromboplastin time; BGL: blood glucose level; BGV: blood gas value; Cl: chloride; CRP: c-reactive protein; DBP: diastolic blood pressure; D-dimer: disseminated intravascular coagulation; DM: diabetes mellitus; eGFR: estimated glomerular filtration rate; FI: first inpatient; FiO2: fraction of inspired oxygen; HbA1c: glycated haemoglobin; HCO3: bicarbonate; HR: highest requirement; HV: highest value; K: potassium; LaV: last value; LAYBA: latest available within one year before admission; LV: lowest value; LYM: lymphocytes; MN: monocytes; Na: sodium; NEUT: neutrophils; O2: oxygen; OA: on admission; pH: potential of hydrogen; PT: prothrombin time; RBGLR: ratio of blood glucose level readings; RR: respiratory rate; SBP: systolic blood pressure; SD: standard deviation; TP: total protein; WCC: white cell count. Note. eGFR > 90= Stage 1, 60–89 = Stage 2, 30–59= Stage 3, 15–29= Stage 4, <15= Stage 5-Stage CKD.

3. Data Curation

The following four pre-treatment stages are undertaken to prepare the data for the ensuing ML modelling analysis.

3.1. Cleaning

In the first data-cleaning step, tainted entities and features are excluded from the rest of the analysis. First, the three individuals with reported non-COVID-19 mortality reasons (as discussed in Section 2) are omitted from the rest of the analysis. Next, features and participants with a high missingness rate are discarded. For this purpose, an inclusion criteria of having a missingness rate of no more than 50% is determined for both features and individuals [24]. Initially, attributes with missing rates larger than the 50% threshold are discarded. Next, the same criterion is applied to data contributors. As a result, the following features are omitted from the rest of the analysis: FI-HbA1c, LV-ALT, HV-ALT, LV-Procalcitonin, HV-Procalcitonin, LV-Ferritin, HV-Ferritin, OA-Troponin, LV-Troponin, HV-Troponin, LAYBA-UACR, LAYBA-Vitamin D, LV-PT, HV-PT, LV-APTT, HV-APTT, LV-Fibrinogen, HV-Fibrinogen, LV-D-dimer, HV-D-dimer, and FI-Ketones. However, no further individual is obviated from the rest of the analysis, as no one holds more than 50% missingness after discarding the abovementioned high-missing rate features.

3.2. Subsetting

After the cleaning phase, data that have met the inclusion criteria and qualified for the subsequent analysis are subdivided into training and testing sets as per the requirements of upcoming supervised ML analysis. For data subsetting, 70% of the cases are allocated as the training set and 30% as the testing set. Stratified random sampling carries out the train-test split process to take into account the distribution of classes. All model training and hyperparameter tuning operations are undertaken on training sets only, with testing sets remaining unseen for evaluation and model interpretation analysis.

3.3. Pre-processing

Three pre-processing steps are conducted to render the data more suitable for ML analysis: outlier treatment, missing value imputation, and feature transformation.

Initially, leveraging the winsorisation technique, we shift the numerical variables placed outside the 5th to 95th percentile to the corresponding boundary. This confinement pre-empts extreme values of skewing the results.

Next, the missing values for numerical features are treated using the k-nearest neighbour data imputation technique [25], configuring the number of neighbours as five. Exploring all non-missing features, the algorithm selects five data entities from the training set with the most congruency with a given data contributor. Then the average of these akin points is used to interpolate the missing values of the given data instance. For categorical variables, the most repeated value is used to fill in the missing values.

Lastly, features are transformed into a more digestible form for ML algorithms. Categorical attributes are converted to numerical form using the binary encoding technique. The numerical features are standardised. The mean of the training set is subtracted from each feature, and then the results are scaled to unit variance by dividing them by the standard deviation of the training set.

3.4. Feature Elimination

After the data curation steps, a voting feature selection is performed on the pre-processed data to reduce the input size and help preclude the occurrence of a dimensionality curse. First, regular LR, gradient boosting (GB), and AdaBoost (AB) models, which all have already succeeded in applications to COVID-19 research, are fine-tuned. To this end, the random search approach is used to select the hyperparameter values delivering the highest five-fold cross-validation accuracy on the training set. The outcomes of hyperparameter tuning are given in Table A1, Appendix. Next, the recursive feature elimination technique is enfolded around each model, forming a voting system. Each voting system then shortlists 15 features (approximately one-tenth the number of data points, a common practice in ML modelling) by investigating

training data only. The features picked by at least two voting systems are then used as predictors to generate the final mortality risk prediction model. The shortlisted features encompass LAYBA-NEUT, HV-NEUT, LaV-LYM, LAYBA-MN, OA-MN, LV-Platelets, HV-Platelets, LV-CRP, LAYBA-PT, FI-BGL, FI-RR, FI-FiO2, and HR-O2.

4. Modelling

This section develops explainable ML models for mortality risk assessment analysis from the selected features. Prior to representing model implementations, providing a brief description of SHAP theory and calculations is of use.

4.1. Preliminary

As a game-theoric model agnostic method, SHAP simulates the formation of outputs by an ML model as a game. In this gamification process, the input features have the role of involved players. Subsequently, the payoff for each player in the game is calculated as Equation (1) [18], based on the principles of Shapley value [19]. On elucidating the formula, the SHAP value of a particular feature for a given individual is calculated by integrating the payoff shares for the feature in all possible coalitions with other variables. The payoff share of the feature in each coalition is determined by calculating the difference between the whole payoff of the coalition with and without the given feature included and then dividing the outcome between the members of the coalition equally.

$$SHAP_x(f) = \sum_{F:f\in F} \left(|F| \times \binom{N}{|F|} \right)^{-1} \times \left(\hat{x}_F - \hat{x}_{F\setminus f} \right) \quad (1)$$

f: a given feature; *x*: a given data point; $SHAP_x(f)$: SHAP value of variable *f* for *x* (the payoff of feature *f* in the designed game); *F*: all permutations of feature with *f* included; $|F|$: the size of *F* (number of features in *F*); *N*: the whole number of features in the models; $\hat{x}_F$: the model's output for x using the feature subset *F*; $\hat{x}_{F\setminus f}$: the model's output for x from the feature subset *F* excluding *f*.

4.2. Mortality Risk Prediction

The first risk assessment analysis is to forecast in-hospital COVID-19 mortality from selected features. For this purpose, a random forest (RF) classifier, which has proven its capability in COVID-19 risk assessment research [26], is fine-tuned using the same approach explained in subsection 3.4. The results of this optimisation analysis are presented in Table A1, Appendix. The fine-tuned RF classifier is then trained on the entire training set to predict inpatient death due to COVID-19. Following that, the generated model undergoes careful evaluation analysis employing four widely used metrics: accuracy, the area under the curve (AUC), sensitivity, and specificity. After evaluating the mortality risk prediction model, SHAP is leveraged to interpret the model globally and locally.

4.3. Mortality Risk Stratification

The second risk assessment analysis is to stratify the in-hospital mortality risk of patients. In order to do so, SHAP clustering [27], an extension of SHAP analysis, is deployed. The k-means [28], an algorithm used in previous COVID-19 research [29,30], is employed on SHAP values to search for meaningful clusters of individuals. The algorithm clusters the subjects into an optimised number of groups [22] with identical variance by optimising a criterion known as inertia [28]. For deciding the number of clusters, the heuristic elbow method is employed. Values of 1 to 9 are explored, and the one resulting in an elbow point, based on inertia values achieved, is chosen [31]. According to the outcome of elbow analysis shown in Figure 1, four is determined as the number of clusters, as the diagram has the sharpest break point for this value.

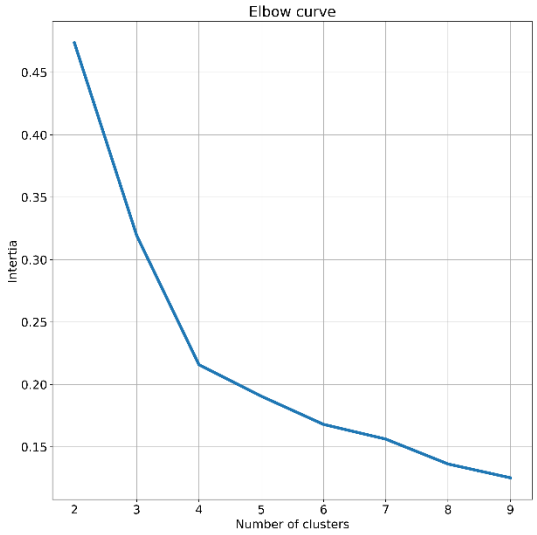

Figure 1. A schematic of elbow analysis operated to decide the number of clusters.

5. Results and Discussion

This section presents the results of model evaluation and interpretation analysis alongside the corresponding discussion. First, the outcomes of mortality risk prediction analysis are given and then those of mortality risk stratification analysis.

5.1. Mortality Risk Prediction

5.1.1. Evaluation

The generated mortality risk prediction model provides these evaluation results across the testing set: 97% accuracy, 78% AUC, 78% sensitivity, and 80% specificity. Such practical evaluation results support the overall effectiveness of the implemented methodologies, including the feature selection, and hyperparameter tuning processes coupled with the final RF classifier. These dependable outcomes also backend the following SHAP-based analysis.

5.1.2. Global Interpretations

The next results to present for the mortality risk prediction model are the outcome of the interpretation analysis. In this regard, first, the results of global interpretation analysis are reported, followed by those of local interpretation analysis.

Figure 2 illustrates the results of global interpretations for the generated mortality risk prediction model in two plots. Both plots represent the features in descending order as per their overall influence on the model's outcomes.

The bee swarm plot in Figure 2A shows SHAP values and their relative association with predictions of death. Each point on the scheme represents a feature value from the testing set. The values of features are colour-coded from blue to red, encoding low to high values. A positive SHAP value for each point denotes the adverse effect of the feature, viz, its contribution level to a higher risk of death. In contrast, a negative SHAP value indicates the protective effect of the relevant feature, i.e., decreasing the risk of death.

The bar chart in Figure 2B is the variable importance plot for the developed mortality risk prediction model. The plot summarises the features' overall impacts on the model outputs according to their mean absolute SHAP values, represented by the length of the bars.

According to Figure 2A, the predictors positively associated with mortality risk predictions are HR-O2, FI-FiO2, HV-NEUT, FI-RR, LV-CRP, LAYBA-PT, OA-MN, LAYBA-NEUT, and LAYBA-MN. By comparing the positive and negative SHAP values of these variables, it is noticeable that HR-O2, FI-FiO2, HV-NEUT, and LAYBA-PT show greater adverse than protective effects. In other words, the sinister roles of higher values for these features are relatively more significant than the protective roles of lower values. On the other hand, with similar explanations, it can be inferred that FI-RR carries a stronger protective than adverse power. The rest of the aforementioned variables have comparable protective and adversarial influences. On the other side, based on the plot, it can also be seen that the modalities negatively associated with the prediction of death comprise LaV-LYM, HV-Platelets, and LV-Platelets. Overall, the first two variables possess more substantial sinister impacts (in lower values) than protective impacts (in higher

values), whereas the last one holds stronger protective effects (in higher values) than sinister effects (in lower values).

Furthermore, based on Figure 2A, one noteworthy inference is that a more influential feature does not necessarily have stronger adverse and protective power at once. To exemplify, notwithstanding the greater overall importance of HR-O2 over FI-FiO2, the latter possesses a more substantial adverse impact than the former on average. This deduction is formed based on predominantly bigger positive SHAP values for FI-FiO2 compared to HR-O2.

Additionally, according to the plots in Figure 2, HR-O2, FI-FiO2, and HV-NEUT form the top three influential features with considerably higher impacts than others. Therefore, undesired measures for these features may be an indicator of high death risk. These findings underscore the importance of careful inpatient surveillance and the monitoring of peak values of oxygen requirement and NEUT, along with the imperative role of immediate inspection of FI-FiO2 after admission for COVID-19 patients with DM.

Another notable point is that two features from the patients' historical profiles, LAYBA-PT and LAYBA-NEUT, have shown considerable mortality predictivity power even in the presence of many on-admission and during-admission data. This observation stresses the potential utility of accessing and considering the history profile of COVID-19 patients with DM and specifies two features as candidates with high priority for consideration in this respect.

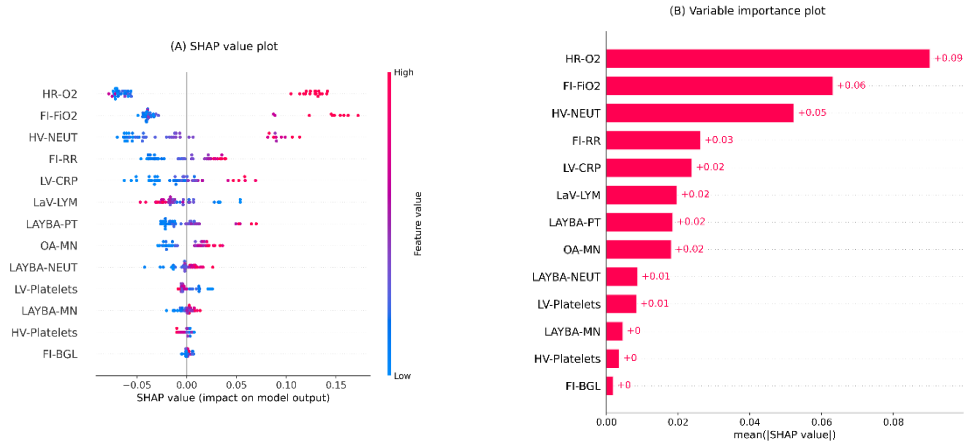

Figure 2. Global interpretation plots for the developed inpatient COVID-19 mortality prediction model. (A) Bee swarm SHAP values plot, (B) SHAP summary importance plot. The bee swarm plot shows all SHAP values in accord with predictors values. The

summary plot presents predictors in descending order based on their overall importance on the model's outcomes derived from mean absolute SHAP values. Note. BGL: blood glucose level; CRP: c-reactive protein; FI: first inpatient; FiO2: fraction of inspired oxygen; HR: highest requirement; HV: highest value; LaV: last value; LAYBA: latest available within one year before admission; LV: lowest value; LYM: lymphocytes; MN: monocytes; NEUT: neutrophils; OA: on admission; O2: oxygen; PT: prothrombin time; RR: respiratory rate; SHAP: SHapley Additive exPlanations.

5.1.3. Local Interpretations

The next outcome to be presented entails the results of local interpretation analysis for the mortality risk prediction model. In this respect, Figure 3 shows the waterfall plots for two randomly selected examples of data entities, one from individuals with death and the other from those with survival as the outcomes of their admissions. These plots start at the base from E[f(x)], representing the average risk of death according to the training set. Next, each arrow illustrates the influence of a feature, i.e., the feature's SHAP value, towards forming the specific prediction for the given entry. The positive associations of the given feature with increased mortality risk prediction are exhibited by red rightward arrows and the negative associations with blue leftward arrows. Finally, at the top of the plot, the model's output for the given sample is represented by f(x). It merits mentioning that each arrow's length denotes the level of impact from its relevant feature, i.e., absolute SHAP value. Moreover, the arrows are displayed in ascending order from the bottom to the top of the plots according to their size.

One immediate recognition from both plots in Figure 3 is that the grade and order for the features' impacts on local interpretations are different from global interpretations. This evidence shows how local interpretations can evolve the transparency of the analysis by explaining the formation of each specific outcome through localising and contrasting the effect of the components, as opposed to giving a generic explanation based on all outcomes.

For the death instance represented in Figure 3A, relatively high values for features FI-FiO2, HR-O2, HV-NEUT, and LAYBA-PT have been the most effective predictors of a fatal outcome. In this regard, it is worth remarking that, in line with the aforementioned discussion, FI-FiO2 had a more adverse impact than HR-O2, the most influential feature overall.

For the survival case reported in Figure 3B, features with protective impacts are HR-O2, FI-FiO2, LaV-LYM, and HV-Platelets. It is worth highlighting that this case has received a non-fatal outcome prediction,

whilst its influential feature, HV-NEUT, shows an adverse impact. Moreover, for this data instance, HR-O2 and Fi-FO2 both deliver a protective effect, with the former's being stronger. This perception is also in line with the overall higher protective influence of HR-O2, as discussed before.

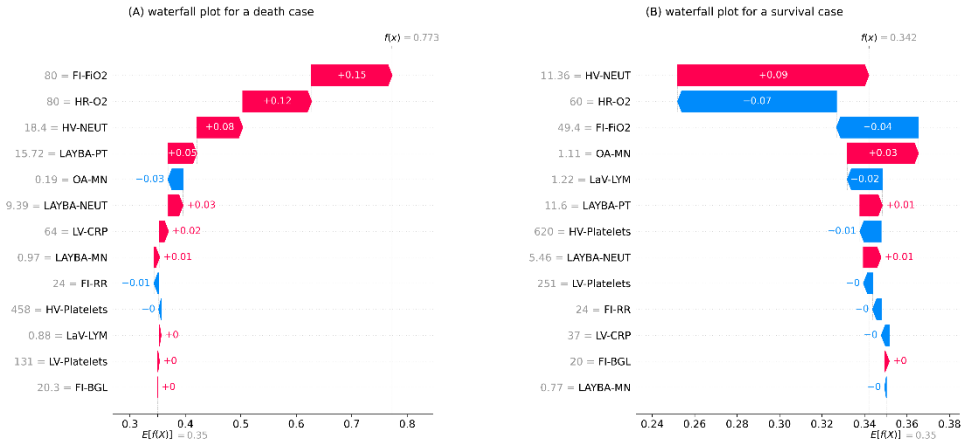

Figure 3. The local interpretation plots for the developed inpatient COVID-19 mortality prediction model. (A) an example of patients with survival as the outcome of admission, (B) an example of patients with death as the outcome of admission. The plots start from the bottom with a predefined prediction for the risk of death equal to the average death rate in the training set. Next, the arrows with an ascending order show how each feature has contributed to the formation of a final prediction specified for the given data instance shown at the top of the plot. Note. BGL: blood glucose level; CRP: c-reactive protein; FI: first inpatient; FiO2: fraction of inspired oxygen; HR: highest requirement; HV: highest value; LaV: last value; LAYBA: latest available within one year before admission; LV: lowest value; LYM: lymphocytes; MN: monocytes; NEUT: neutrophils; OA: on admission; O2: oxygen; PT: prothrombin time; RR: respiratory rate; SHAP: SHapley Additive exPlanations.

5.2. Mortality Risk Stratification

Table 3 outlines the results of the SHAP clustering analysis, including the distribution of patients in the generated clusters, the rate of mortality outcome in each cluster, and a summary of the statistical characteristics of predictors within clusters. Based on the table, it can be apprehended that the clustering approach has made an appropriate risk stratification system by forming four categories with disparate characteristics.

In terms of mortality rates, cluster 1 poses a zero mortality rate, cluster 2 has a moderate mortality rate, and clusters 3 and 4 have relatively high mortality rates. Further distinctive patterns can be found based on the feature distributions within clusters, specifically for the more critical variables. For example, a pronounced discriminator between clusters 1 and 2 compared to clusters 3 and 4 is that the first two have an average HR-O2 considerably lower than the other two. Additionally, comparing clusters 1 and 2, a

prominent discrepancy is that the former, in general, includes patients with more desired values for HV-NEUT and LV-CRP. One more pattern to mention is that a significant distinguisher between clusters 3 and 4 is the relatively higher average values for FI-FiO2 and LV-CRP in cluster 3.

Table 3. The results of the performed SHAP clustering to generate a mortality risk stratification system.

Characteristics	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Count	21	7	10	8
Mortality rate (%)	0.00 ± 0.00	0.43 ± 0.53	0.70 ± 0.48	0.75 ± 0.46
HR-O2	37.31 ± 9.85	37.57 ± 10.83	80.00 ± 0.00	80.00 ± 0.00
FI-FiO2	36.49 ± 10.06	34.03 ± 10.55	78.00 ± 6.32	32.08 ± 6.87
HV-NEUT	6.08 ± 1.97	14.95 ± 3.46	8.99 ± 4.36	7.47 ± 3.32
FI-RR	22.74 ± 4.95	25.57 ± 7.35	31.80 ± 6.36	28.62 ± 6.7
LV-CRP	29.97 ± 21.68	42.67 ± 41.62	75.09 ± 62.95	37.35 ± 33.85
LaV-LYM	1.25 ± 0.5	1.18 ± 0.49	1.10 ± 0.4	0.99 ± 0.62
LAYBA-PT	11.57 ± 0.89	11.36 ± 0.58	12.32 ± 1.87	11.58 ± 0.51
OA-MN	0.72 ± 0.32	0.68 ± 0.3	0.48 ± 0.24	0.63 ± 0.3
LAYBA-NEUT	4.57 ± 1.56	5.19 ± 0.97	5.61 ± 1.73	5.32 ± 1.48
LV-Platelets	189.2 ± 69.29	175 ± 69.63	210.26 ± 70.85	151.43 ± 42.15
LAYBA-MN	0.71 ± 0.25	0.54 ± 0.15	0.64 ± 0.23	0.64 ± 0.13
HV-Platelets	272.1 ± 113.01	420.00 ± 157.79	366.96 ± 152.35	269.68 ± 103.64
FI-BGL	9.67 ± 4.13	13.09 ± 6.42	10.59 ± 4.37	9.85 ± 4.95

Note. BGL: blood glucose level; CRP: c-reactive protein; FI: first inpatient; FiO2: fraction of inspired oxygen; HR: highest requirement; HV: highest value; LaV: last value; LAYBA: latest available within one year before admission; LV: lowest value; LYM: lymphocytes; MN: monocytes; NEUT: neutrophils; OA: on admission; O2: oxygen; PT: prothrombin time; RR: respiratory rate; SD: standard deviation.

6. Complementary Analysis

This section presents some extra analysis embarked on for robustness assessments. The following set of amendments is applied to the compartments of the proposed learning environment. The data are reshuffled in their entirety, and a new round of 30–70 stratified random samplings is performed to reallocate training and testing sets. In addition, missing values are interpolated using a different technique by applying the iterative imputation. In order to do so, a Bayesian ridge regressor is set as the estimator for the numerical variables and an RF classifier for the categorical variables. In addition, the mortality risk prediction modelling is performed again using a support vector classifier (SVC). The SVC is fine-tuned using the random search approach described in Subsection 3.4, and the results are presented in Table A1, Appendix. Following these updates, the evaluation, interpretation, and clustering analyses are re-conducted. The new results are presented concisely below, and the consistency of the proposed core workflow in producing practical outcomes in line with previously discussed findings is inspected.

The features selected in the updated analysis are BGV-Na, BGV-Cl, LV-NEUT, HV-NEUT, LaV-LYM, LV-Platelets, HV-Platelets, LV-Albumin, LV-CRP, Preadmission-SBP, DM duration, FI-RR, FI-FiO2, and HR-O2. In comparison, the top six most important features, according to the previous analysis (HR-O2, FI-FiO2, HV-NEUT, FI-RR, LV-CRP, and LaV-LYM), are all shortlisted in the renewed analysis as well.

Furthermore, the updated fatality risk prediction model yields these new evaluation outcomes over the testing set: 87% accuracy, 92% AUC, 72% sensitivity, and 74% specificity. Similar to the primary analysis, these results are practical, with an outcome of more than 70% for every metric.

Moreover, Figure 4 illustrates the global interpretation plots for the renewed models. According to the plots, HR-O2 and FI-FiO2 are the first and second most important features, similar to the primary analysis. In addition, the top five ranks are occupied by the same features in the original and renewed analysis.

Furthermore, Table 4 shows the results of the new SHAP clustering analysis. For brevity, only the top four important features are shown in the table. As can be seen, similar to the primary analysis, the new SHAP clustering analysis successfully groups patients into four distinguishable categories.

All in all, based on the discussion above, there is a significant agreement between the results of both the updated and the original analysis. This alignment promises the robustness of the core interpretable ML workflow proposed for mortality risk assessment in COVID-19 patients.

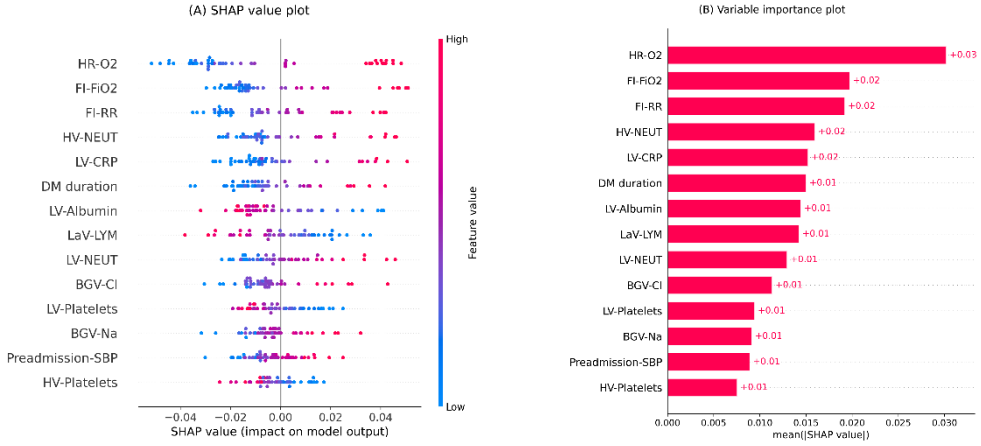

Figure 4. Global interpretation plots for the updated inpatient COVID-19 mortality prediction model. (A) Bee swarm SHAP values plot, (B) SHAP summary importance plot. The bee swarm plot shows all SHAP values in accord with predictors values. The

summary plot presents predictors in descending order based on their overall importance on the model's outcomes derived from mean absolute SHAP values. Note. BGV: blood gas value; Cl: chloride; CRP: c-reactive protein; DM: diabetes mellitus; FI: first inpatient; FiO2: fraction of inspired oxygen; HR: highest requirement; HV: highest value; LaV: last value; LV: lowest value; LYM: lymphocytes; Na: sodium; NEUT: neutrophils; O2: oxygen; RR: respiratory rate; SBP: systolic blood pressure; SHAP: SHapley Additive exPlanations.

Table 4. The results of the performed updated SHAP clustering to generate a mortality risk stratification system.

Characteristics	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Count	23	10	7	6
Mortality rate (%)	0.04 ± 0.021	0.50 ± 0.53	0.71 ± 0.49	0.83 ± 0.41
HR-O2	40.07 ± 10.69	41.31 ± 10.70	80 ± 0.00	80 ± 0.00
FI-FiO2	34.07 ± 7.69	33.67 ± 7.16	80.00 ± 0.00	51.60 ± 11.36
HV-NEUT	5.68 ± 1.91	14.59 ± 3.60	9.04 ± 4.09	9.07 ± 4.76
FI-RR	22.78 ± 4.17	26.10 ± 7.29	28.71 ± 6.82	30.17 ± 6.46
LV-CRP	28.83 ± 14.59	38.44 ± 33.25	89.35 ± 62.95	106.84 ± 58.60

Note. FI: first inpatient; FiO2: fraction of inspired oxygen; HR: highest requirement; HV: highest value; LV: lowest value; NEUT: neutrophils; O2: oxygen; RR: respiratory rate; SD: standard deviation.

7. Conclusions

Inpatient COVID-19 mortality risk assessment specifically designed for patients with pre-existing DM was performed in this work. This goal was achieved by investigating a set of clinical features exclusively pertaining to DM and COVID-19 interplay for 156 individuals. Initially, the clinical data of the studied subjects were carefully pre-treated for the subsequent standard ML modelling analysis. After that, a mortality risk prediction model was created, exercising established ML pipelines. Evaluation analysis was then performed on the generated model. The results underpinned the effectiveness of the data treatment and modelling analysis. Afterwards, the generated model was interpreted globally and locally using SHAP. These interpretations help extend the transparency of the analysis. Next, a mortality risk stratification system was developed upon the outcomes of the SHAP analysis. Finally, an extra analysis was performed to further examine the stability of the core pipelines, where the outcomes corroborated this. The analysis reported in this work can be applied to online surveillance of hospitalised patients. The findings suggest some critical features to be reviewed more carefully in this monitoring process. To further expand upon this area of knowledge, future work could include more rigorous scrutiny of SHAP clustering results by devising a nested model interpretation mechanism.

While this study offers valuable insights into the COVID-19 mortality risks for diabetic patients using interpretable machine learning models, several areas for improvement remain. Expanding the cohort size in future studies will enhance the generalisability of the findings and allow for more robust model validation. Moreover, incorporating other comorbidities, alongside diabetes, will broaden the relevance of the results and provide a more comprehensive understanding of the complex interplay between various health conditions and COVID-19. Additionally, comparative analyses with existing risk prediction models would offer a clearer perspective on the relative performance of the proposed methodology. Addressing these challenges, alongside considerations for real-world implementation, such as integration with healthcare systems and clinician training, will be critical for maximising the practical impact of the model.

Institutional Review Board Statement

For analysis, an anonymised dataset is used, and the study is conducted under National Health Service (NHS) ethics as approved by the East-Midlands-Leicester South Research Ethics Committee (20/EM/0145).

Code and Data Availability

For the analysis, we coded in Python (3.6.7) [32]. The libraries used include; Pandas [33], NumPy [34], and Sklearn [35]. The implementation source code is publicly available on this Gitlab repository: https://gitlab.com/Heydar-Khadem/DM-interplay-COVID19.git

Acknowledgements

We want to thank Ahmed Iqbal, Marni Greig, Muhammad Fahad Arshad, Thomas H Julian, and Sher Ee Tan for their efforts in collecting the clinical data used in this paper.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A

In this section, the results of hyperparameter tuning are given. Table A1 represents the outcomes of randomised fine-tuning analysis on hyperparameters for all models in the paper. For each model, a search space is studied for the associated hyperparameters then the random search approach is conducted to select hyperparameter values according to the highest performance on the training set.

Table A1. The results of the conducted randomised hyperparameter tuning processes for the classifiers in the article.

Model	Hyperparameter	Search Space	Selected
LR	Regularisation strength	{0, 0.10, 0.20, …,1}	0.40
	class weight	{0, 1, …,10}	7
	Maximum number of iterations	{1000, 2000, …,10000}	7000
GB	Learning rates	{0.01, 0.02, …,1}	0.10
	Number of boosting stages	{20, 40, …,200}	160
	Minimum number of samples required to split an internal node	{1, 2, …,10}	2
	Minimum number of samples required to be at a leaf node	{{1, 2, …,10}	6
	Maximum depth of the individual estimators	{1, 2, …,10}	9
AB	Maximum number of estimators at which boosting is terminated	{10, 20, …,100}	90
	Learning rates	{0.01, 0.02, …,1}	1.58
RF	number of trees	{50, 100, …,500}	20
	Maximum depth of the tree	{1, 2, …,10}	3
	Minimum number of samples required to split an internal node,	{1, 2, …,10}	4
	Minimum number of samples required to be at a leaf node	{1, 2, …,10}	6
	Maximum number of leaf nodes	{1, 2, …,10}	3
	minimum impurity decrease	{0, 0.001, 0.002, …, 0.010}	0.004
	Cost complexity pruning factor	{0.01, 0.02, …,0.10}	0.01
	Minimum weighted fraction of the sum total of weights	{0.01, 0.02, …0.10}	0.01
SVC	Class weight	{0, 1, …,10}	6
	Maximum integration	{100, 200, …,10000}	2400

Note. AB: AdaBoost; LR: logistic regression; GB: gradient boosting; RF: random forest; SVC: support vector classifier.

References

[1] Zhou, K.; Sun, Y.; Li, L.; Zang, Z.; Wang, J.; Li, J.; Liang, J.; Zhang, F.; Zhang, Q.; Ge, W.; et al. Eleven Routine Clinical Features Predict COVID-19 Severity Uncovered by Machine Learning of

Longitudinal Measurements. Comput. Struct. Biotechnol. J. 2021, 19, 3640–3649, doi:10.1016/j.csbj.2021.06.022.

[2] Onder, G.; Rezza, G.; Brusaferro, S. Case-Fatality Rate and Characteristics of Patients Dying in Relation to COVID-19 in Italy. JAMA 2020, 323, 1775–1776. https://doi.org/10.1001/jama.2020.4683.

[3] Wargny, M.; Potier, L.; Gourdy, P.; Pichelin, M.; Amadou, C.; Benhamou, P.-Y.; Bonnet, J.-B.; Bordier, L.; Bourron, O.; Chaumeil, C.; et al. Predictors of Hospital Discharge and Mortality in Patients with Diabetes and COVID-19: Updated Results from the Nationwide CORONADO Study. Diabetologia 2021, 64, 778–794. https://doi.org/10.1007/s00125-020-05351-w.

[4] Sourij, H.; Aziz, F.; Bräuer, A.; Ciardi, C.; Clodi, M.; Fasching, P.; Karolyi, M.; Kautzky-Willer, A.; Klammer, C.; Malle, O.; et al. COVID-19 Fatality Prediction in People with Diabetes and Prediabetes Using a Simple Score upon Hospital Admission. Diabetes Obes. Metab. 2021, 23, 589–598. https://doi.org/10.1111/dom.14256.

[5] Corona, G.; Pizzocaro, A.; Vena, W.; Rastrelli, G.; Semeraro, F.; Isidori, A.M.; Pivonello, R.; Salonia, A.; Sforza, A.; Maggi, M. Diabetes Is Most Important Cause for Mortality in COVID-19 Hospitalized Patients: Systematic Review and Meta-Analysis. Rev. Endocr. Metab. Disord. 2021, 22, 275–296. https://doi.org/10.1007/s11154-021-09630-8.

[6] Ciardullo, S.; Zerbini, F.; Perra, S.; Muraca, E.; Cannistraci, R.; Lauriola, M.; Grosso, P.; Lattuada, G.; Ippoliti, G.; Mortara, A.; et al. Impact of Diabetes on COVID-19-Related in-Hospital Mortality: A Retrospective Study from Northern Italy. J. Endocrinol. Investig. 2021, 44, 843–850. https://doi.org/10.1007/s40618-020-01382-7.

[7] Shah, H.; Khan, M.S.H.; Dhurandhar, N.V.; Hegde, V. The Triumvirate: Why Hypertension, Obesity, and Diabetes Are Risk Factors for Adverse Effects in Patients with COVID-19. Acta Diabetol. 2021, 58, 831–843.

[8] Campbell, T.W.; Wilson, M.P.; Roder, H.; MaWhinney, S.; Georgantas, R.W.; Maguire, L.K.; Roder, J.; Erlandson, K.M. Predicting Prognosis in COVID-19 Patients Using Machine Learning and Readily Available Clinical Data. Int. J. Med. Inform. 2021, 155, 104594, doi:10.1016/J.IJMEDINF.2021.104594.

[9] Dennis, J.M.; Mateen, B.A.; Sonabend, R.; Thomas, N.J.; Patel, K.A.; Hattersley, A.T.; Denaxas, S.; McGovern, A.P.; Vollmer, S.J. Diabetes and COVID-19 Related Mortality in the Critical Care Setting: A Real-Time National Cohort Study in England. 2020. Available online: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3615999.

[10] Haimovich, A.D.; Ravindra, N.G.; Stoytchev, S.; Young, H.P.; Wilson, F.P.; van Dijk, D.; Schulz,

W.L.; Taylor, R.A. Development and Validation of the Quick COVID-19 Severity Index: A Prognostic Tool for Early Clinical Decompensation. Ann. Emerg. Med. 2020, 76, 442–453. https://doi.org/10.1016/j.annemergmed.2020.07.022.

[11] Zheng, B.; Cai, Y.; Zeng, F.; Lin, M.; Zheng, J.; Chen, W.; Qin, G.; Guo, Y. An Interpretable Model-Based Prediction of Severity and Crucial Factors in Patients with COVID-19. Biomed Res. Int. 2021, 2021, 8840835. https://doi.org/10.1155/2021/8840835.

[12] Lalmuanawma, S.; Hussain, J.; Chhakchhuak, L. Applications of Machine Learning and Artificial Intelligence for Covid-19 (SARS-CoV-2) Pandemic: A Review. Chaos Solitons Fractals 2020, 139, 110059.

[13] Kar, S.; Chawla, R.; Haranath, S.P.; Ramasubban, S.; Ramakrishnan, N.; Vaishya, R.; Sibal, A.; Reddy, S. Multivariable Mortality Risk Prediction Using Machine Learning for COVID-19 Patients at Admission (AICOVID). Sci. Rep. 2021, 11, 12801. https://doi.org/10.1038/s41598-021-92146-7.

[14] Khadem, H.; Nemat, H.; Elliott, J.; Benaissa, M. Signal Fragmentation Based Feature Vector Generation in a Model Agnostic Framework with Application to Glucose Quantification Using Absorption Spectroscopy. Talanta 2022, 243, 123379. https://doi.org/10.1016/j.talanta.2022.123379.

[15] Mauer, E.; Lee, J.; Choi, J.; Zhang, H.; Hoffman, K.L.; Easthausen, I.J.; Rajan, M.; Weiner, M.G.; Kaushal, R.; Safford, M.M.; et al. A Predictive Model of Clinical Deterioration among Hospitalized COVID-19 Patients by Harnessing Hospital Course Trajectories. J. Biomed. Inform. 2021, 118, 103794. https://doi.org/10.1016/j.jbi.2021.103794.

[16] Bhatt, S.; Cohon, A.; Rose, J.; Majerczyk, N.; Cozzi, B.; Crenshaw, D.; Myers, G. Interpretable Machine Learning Models for Clinical Decision-Making in a High-Need, Value-Based Primary Care Setting. NEJM Catal. Innov. Care Deliv. 2021, 2. https://doi.org/10.1056/CAT.21.0008.

[17] Lundberg, S.M.; Erion, G.G.; Lee, S.-I. Consistent Individualized Feature Attribution for Tree Ensembles. arXiv 2018, arXiv1802.03888.

[18] Lundberg, S.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions. In Proceedings of the 31th Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 4765–4774.

[19] Shapley, L.S. A Value for N-Person Games. Contrib. Theory Games 1953, 2, 307–317.

[20] Pan, P.; Li, Y.; Xiao, Y.; Han, B.; Su, L.; Su, M.; Li, Y.; Zhang, S.; Jiang, D.; Chen, X.; et al. Prognostic Assessment of COVID-19 in the Intensive Care Unit by Machine Learning Methods: Model Development and Validation. J. Med. Internet Res. 2020, 22, e23128. https://doi.org/10.2196/23128.

[21] Hathaway, Q.A.; Roth, S.M.; Pinti, M.V.; Sprando, D.C.; Kunovac, A.; Durr, A.J.; Cook, C.C.; Fink, G.K.; Cheuvront, T.B.; Grossman, J.H.; et al. Machine-Learning to Stratify Diabetic Patients Using

Novel Cardiac Biomarkers and Integrative Genomics. Cardiovasc. Diabetol. 2019, 18, 78. https://doi.org/10.1186/s12933-019-0879-0.

[22] Khadem, H.; Nemat, H.; Eissa, M.R.; Elliott, J.; Benaissa, M. COVID-19 Mortality Risk Assessments for Individuals with and without Diabetes Mellitus: Machine Learning Models Integrated with Interpretation Framework. Comput. Biol. Med. 2022, 144, 105361. https://doi.org/10.1016/j.compbiomed.2022.105361.

[23] Iqbal, A.; Arshad, M.; Julian, T.; Tan, S.; Greig, M.; Elliott, J. Higher Admission Activated Partial Thromboplastin Time, Neutrophil-Lymphocyte Ratio, Serum Sodium, and Anticoagulant Use Predict in-Hospital Covid-19 Mortality in People with Diabetes: Findings from Two University Hospitals in the UK. Diabet. Med. 2021, 178, 108955. https://doi.org/10.1016/j.diabres.2021.108955.

[24] Zwart, D.L.; Langelaan, M.; van de Vooren, R.C.; Kuyvenhoven, M.M.; Kalkman, C.J.; Verheij, T.J.; Wagner, C. Patient Safety Culture Measurement in General Practice. Clinimetric Properties of "SCOPE." BMC Fam. Pract. 2011, 12, 117. https://doi.org/10.1186/1471-2296-12-117.

[25] Jonsson, P.; Wohlin, C. An Evaluation of K-Nearest Neighbour Imputation Using Likert Data. In Proceedings of the 10th International Symposium on Software Metrics, Chicago, IL, USA, 11–17 September 2004; pp. 108–118.

[26] Wang, J.; Yu, H.; Hua, Q.; Jing, S.; Liu, Z.; Peng, X.; Luo, Y. A Descriptive Study of Random Forest Algorithm for Predicting COVID-19 Patients Outcome. PeerJ 2020, 8, e9945. https://doi.org/10.7717/peerj.9945.

[27] Forte, J.C.; Yeshmagambetova, G.; van der Grinten, M.L.; Hiemstra, B.; Kaufmann, T.; Eck, R.J.; Keus, F.; Epema, A.H.; Wiering, M.A.; van der Horst, I.C.C. Identifying and Characterizing High-Risk Clusters in a Heterogeneous ICU Population with Deep Embedded Clustering. Sci. Rep. 2021, 11, 12109. https://doi.org/10.1038/s41598-021-91297-x.

[28] MacQueen, J. Some Methods for Classification and Analysis of Multivariate Observations. In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, University of California, Berkeley, 7 January 1966; Volume 1, pp. 281–297.

[29] Abdullah, D.; Susilo, S.; Ahmar, A.S.; Rusli, R.; Hidayat, R. The Application of K-Means Clustering for Province Clustering in Indonesia of the Risk of the COVID-19 Pandemic Based on COVID-19 Data. Qual. Quant. 2021, 56, 1283–1291. https://doi.org/10.1007/s11135-021-01176-w.

[30] Hutagalung, J.; Ginantra, N.L.W.S.R.; Bhawika, G.W.; Parwita, W.G.S.; Wanto, A.; Panjaitan, P.D. COVID-19 Cases and Deaths in Southeast Asia Clustering Using K-Means Algorithm. J. Phys. Conf. Ser. 2021, 1783, 012027. https://doi.org/10.1088/1742-6596/1783/1/012027.

[31] Syakur, M.A.; Khotimah, B.K.; Rochman, E.M.S.; Satoto, B.D. Integration K-Means Clustering

Method and Elbow Method for Identification of the Best Customer Profile Cluster. IOP Conf. Ser. Mater. Sci. Eng. 2018, 336, 012017. https://doi.org/10.1088/1757-899X/336/1/012017.

[32] Van Rossum, G.; Drake, F.L. Python 3 Reference Manual; CreateSpace: Scotts Valley, CA, USA, 2009; ISBN 1441412697.

[33] McKinney, W. Data Structures for Statistical Computing in Python. In Proceedings of the 9th Python in Science Conference, Austin, TX, USA, 28 June–3 July 2010; Volume 445, pp. 51–56.

[34] Harris, C.R.; Millman, K.J.; van der Walt, S.J.; Gommers, R.; Virtanen, P.; Cournapeau, D.; Wieser, E.; Taylor, J.; Berg, S.; Smith, N.J.; et al. Array Programming with NumPy. Nature 2020, 585, 357–362. https://doi.org/10.1038/s41586-020-2649-2.

[35] Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-Learn: Machine Learning in Python. J. Mach. Learn. Res. 2011, 12, 2825–2830.