

Designing and Evaluating Heterogeneous Ensembles for Blood Glucose Level Forecasting

Mamoune Benaida, Ibtissam Abnane, and Ali Idri(✉)

Software Project Management Research Team, ENSIAS, Mohammed V University, Rabat, Morocco
{mamoune_benaida, ali.idri}@um5.ac.ma, ibtissam.abnane19@gmail.com

Abstract. Accurate forecasting of blood glucose levels (BGL) in individuals with type 1 diabetes mellitus (T1DM) is crucial for effective disease management. This study focuses on advancing BGL prediction by performing a comparative evaluation of single machine learning models and heterogeneous ensembles. The main objective of this study is to determine the optimal model configuration for accurate and reliable blood glucose predictions by examining various optimization strategies. For this purpose, several machine learning techniques were used, including Long Short-Term Memory (LSTM), Convolutional Neural Networks (CNN), Gated Recurrent Units (GRU), Support Vector Regression (SVR) and Deep Belief Networks (DBN). All these models were incorporated into various ensembles and fine tuned using Particle Swarm Optimization (PSO), Random Search (RS) and Bayesian Optimization (BO). We evaluated these models using metrics such as Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Mean Magnitude Relative Error (MMRE) and Predictive Level at 25% (PRED). Our findings demonstrated that PSO fine-tuned models, particularly those using GRU, performed exceptionally well. Most significantly, the PGRU model achieved a RMSE of 7.87 mg/dl. Ensembles optimized using RS and PSO, in particularly ERME and EPAV, outperformed all others, with RMSE values of 12.25 mg/dl and 12.26 mg/dl, respectively. Statistical validation through the Scott-Knott test (SK) and Borda Count (BC) analysis corroborated that PSO significantly enhanced accuracy in both individual and ensemble models. In conclusion, this study underscores the high efficacy of PSO in optimizing BGL forecasting models. Additionally, heterogeneous ensembles offer more consistent and reliable predictions than single models, thereby providing a promising solution for real-time diabetes management.

Keywords: Diabetes Mellitus · Blood Glucose Level · Time Series Forecasting · Ensemble Learning · Heterogeneous Ensembles · Optimization Techniques

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
O. Gervasi et al. (Eds.): ICCSA 2025, LNCS 15649, pp. 116–133, 2025.
https://doi.org/10.1007/978-3-031-96997-3_8

1 Introduction

T1DM is a lifelong autoimmune condition that results in the destruction of insulin-producing beta cells in the pancreas, causing complete deficiency of insulin [1]. This disease predominantly affects children and young adults, with an increasing global incidence. According to the International Diabetes (IDF) Federation, the incidence of T1DM increases by approximately 3% annually, particularly among children and adolescents [2]. This upward trend emphasizes the importance of developing effective management strategies to reduce the associated health risks and improve the quality of life of individuals with T1DM.

Complications associated with T1DM can manifest acutely and chronically. Acute complications, such as diabetic ketoacidosis, present a significant life-threatening risk due to severe insulin deficiency [3]. The development of chronic complications has been associated with long-term uncontrolled BGL levels, among them cardiovascular disease, neuropathy, retinopathy, and nephropathy [3]. These complications can lead to significant impairment in patients' quality of life and impose substantial financial burdens on healthcare systems [4]. Effective management of T1DM necessitates continuous glucose monitoring to prevent such complications [5].

By providing real-time blood glucose data, Continuous glucose monitoring (CGM) systems have played a significant role in the management of type 1 diabetes, enabling patients to accurately adjust their insulin therapy and lifestyle. The integration of artificial intelligence and predictive algorithms into these systems enables the prediction of blood glucose trends, aiding patients in avoiding the potentially harmful conditions of both hyperglycemia and hypoglycemia. These predictive capabilities are essential components in the development of artificial pancreas systems. These systems automate insulin delivery, enhance glycemic control, and reduce the risk of long-term complications.

In order to advance the management of T1DM, it is necessary to address the issue of time-series forecasting. This process involves predicting future BGLs based on historical data. The development of accurate predictive models for blood glucose values holds significant potential to empower both patients and healthcare professionals by furnishing them with actionable information regarding insulin dosages, dietary intake, and physical activity [5]. A comprehensive approach to the management of T1DM is imperative to decrease the risk of serious glycemic complications. Traditional statistical models, including the Autoregressive Integrated Moving Average (ARIMA) model, have been used to predict BGL. However, these models are limited by the complexity of the non-linear, dynamic nature of glucose metabolism. Indeed, linear models, which form the basis of ARIMA models, are unsuited to the complexity of high-frequency data generated by CGM systems.

In recent years, machine learning (ML) and deep learning (DL) techniques have offered more robust solutions for BGL forecasting [6,7]. Models such as SVR, DBN, CNN, LSTM, and GRU have demonstrated significant potential for effectively capturing both short- and long-term dependencies in glucose data [8]. Each model demonstrates distinct strengths: SVR is effective in high-dimensional

spaces [9], DBNs excel at learning hierarchical data representations [10], CNNs are proficient at modeling local patterns in time-series data [11], and LSTMs and GRUs are well-suited for sequential data, addressing the vanishing gradient problem that affects traditional RNNs [7,10]. These advances have significantly improved the accuracy of BGL forecasting.

Although single models have been successful in predicting BGLs, ensemble methods, which combine multiple models, have proven to be a powerful strategy for improving accuracy. Heterogeneous ensembles are particularly promising because they leverage the strengths of various models [12]. For example, a GRU may excel in capturing long-term trends, whereas a CNN is better at identifying short-term fluctuations. Despite their potential, the use of heterogeneous ensembles in BGL forecasting remains underexplored, and further research is required to determine the optimal combination of models for enhanced performance.

To enhance the performance of both the single models and ensembles, optimization techniques such as RS, BO, and PSO are employed to fine-tune the hyperparameters. RS explores the hyperparameter space by randomly sampling configurations, whereas BO uses a probabilistic model to make more informed choices about which hyperparameters are likely to yield the best results. Inspired by the social behaviors of birds and fish, PSO iteratively refines candidate solutions to find optimal configurations. Integrating these techniques into heterogeneous ensembles for BGL forecasting introduces a novel and powerful approach for improving prediction accuracy.

Numerous studies have already explored the use of traditional models, such as ARIMA and advanced ML/DL models, such as SVR, DBN, CNN, LSTM, and GRU for BGL prediction. These models were also explored in previous studies [6,10]. However, a gap remains in evaluating heterogeneous ensemble models systematically, particularly when optimized with techniques such as PSO, RS, and BO in the context of One-Step Ahead Forecasting (OSF). This study aimed to bridge this gap by developing and assessing a comprehensive framework for building and optimizing heterogeneous ensembles for BGL forecasting.

This study addresses the following research questions:

- RQ1) Is there any single predictive model that outperforms others for BGL forecasting?
- RQ2) Which heterogeneous ensemble provides the most accurate and reliable BGL predictions?
- RQ3) How do heterogeneous ensembles compare to their constituent singles in BGL forecasting?
- RQ4) How do different optimization techniques improve the performance of single and ensemble models for BGL forecasting?

The rest of the paper is structured as follows: Sect. 2 presents the ML/DL and optimization techniques used, Sect. 3 provides a review of related work; Sect. 4 outlines the methodology for developing and evaluating models and ensembles; Sect. 5 presents the results and answers to the research questions. Threats to validity are discussed in Sect. 6, and Sect. 7 concludes the paper with key findings and future research directions.

2 Background

This section provides a comprehensive background on the ML/DL single techniques and the optimization techniques used.

2.1 Single Models

In the context of BGL forecasting, several ML/DL models have been employed to improve the prediction accuracy. Each of these models has unique strengths, leveraging different aspects of the data to enhance forecasting performance. This section provides an overview of the single models utilized in the study: SVR, DBN, CNN, LSTM, and GRU.

Support Vector Regression. SVR is a concept introduced by Vapnik in 1992 [13], is an extension of support vector machines. It was developed for regression tasks that utilize an insensitivity loss function, designated as ϵ-insensitivity, with the objective of mapping input data into a higher-dimensional feature space. The aim of this approach is to identify a function that, although remaining within a specified range of ϵ values relative to the target values, maintains the simplicity of the initial model. This approach is designed to enhance the generalizability of the results obtained.

Deep Belief Networks. Introduced by Geoffrey Hinton in 2006 [14], DBNs are sophisticated neural networks consisting of multiple layers of probabilistic models with hidden stochastic variables. These networks use a generative model, typically a Restricted Boltzmann Machine (RBM), to pre-train each layer individually. In this method, the output of one layer serves as the input to the next, making it easier for DBNs to learn generative representations of data. After pre-training, the network was fine-tuned using supervised techniques such as back-propagation to improve its discrimination capabilities.

Convolutional Neural Networks. The conceptualisation of CNNs can be traced back to the 1980 s [15], with Yann LeCun. The inspiration for CNNs derived from the study of the animal visual cortex. These networks have acquired major importance in the field of image recognition and sequential data analysis. The application of CNNs is based on the use of convolutional filters to identify the proprities of input data. Such networks are typically characterized by their standard architectures, which include convolutional layers, pooling layers, and fully connected layers. When applied to time series data, one-dimensional CNNs effectively capture temporal dependencies, making them particularly useful for predicting BGL by using sequential data to improve prediction accuracy [7].

Long Short-Term Memory. LSTMs as introduced by Hochreiter and Schmidhuber [16], solve the problem of gradient vanishing in traditional recurrent neural networks (RNNs). These networks are designed with the intention to capture long-term dependencies in sequential data, making them particularly well suited to video, speech, and time-series tasks. Their architecture integrates memory cells with a cell state and three key gates: input, forget, and output. The function of these gates is to regulate the flow of information, thereby enabling selective retention and updating [7,16].

Gated Recurrent Units. GRUs, as proposed in 2014, by Kyunghyun Cho et al. [17], represent a variant of recurrent neural networks (RNNs) that aims to reduce their computational complexity by combining the input and forget gates of the LSTM model into a unified update gate. This approach has been shown to reduce computational complexity and accelerate the network training process. The efficacy of GRUs in handling long-term dependencies, particularly in scenarios involving shorter sequences, has been demonstrated. Moreover, GRUs have exhibited noteworthy efficacy in domains such as image and speech recognition, attaining results that are comparable to those of LSTMs despite their more streamlined design.

2.2 Optimization Techniques

Optimization techniques play a significative role in adjustment of model parameters, thereby ensuring optimal performance. Indeed, the accuracy and predictive capability of a model are closely related to its parameter tuning, which varies according to the considered data sets. Consequently, optimization methods are indispensable for identifying the optimal values that will maximize the efficiency of a model. This study will focus on three such approaches: RS, BO, and PSO methods.

Random Search. RS is a robust optimization methodology for tuning the hyperparameters of ML and DL models. As distinct from grid search, RS employs a random sampling method, whereby values are selected from predefined ranges. This process is repeated iteratively, with performance evaluated at each iteration until a predetermined criterion is fulfilled [18]. This approach has the advantage to explore extensive hyperparameter spaces without requiring an exhaustive search, making it well-suited for high-dimensional spaces where optimal values are often not known.

Bayesian Optimization. BO has become a widely used technique in ML/DL research for hyperparameter optimization due to its ability to identify the optimal parameters. This technique employs a probabilistic model, most often a Gaussian process (GP), in order to estimate the performance of different parameter combinations. In addition, to select the next set of hyperparameters to be

assessed, BO uses a learning function, generally the Expected Improvement (EI) or Upper Confidence Bound (UCB) [19]. This approach reduces the number of evaluations required, making BO particularly useful for optimizing complex DL models.

Particle Swarm Optimization. Developed by Kennedy and Eberhart in 1995 [20], is a meta-heuristic inspired by bird flocking and fish schooling. It optimizes problems by iteratively adjusting each particle's position based on its own experience and that of its neighbors, aiming to improve solutions relative to a fitness function. PSO is particularly effective for high-dimensional, complex problems, as it often finds optimal solutions in a single run. Key parameters include the number of iterations, swarm size, velocity limits (v_{min}, v_{max}), learning factors (c_1, c_2), and inertial weight (w).

3 Related Work

Recent advances in BGL forecasting increasingly involve the use of ML and DL techniques, in contrast to traditional statistical models such as AR methods. These techniques excel because of their ability to automatically extract and learn data representations, making them well-suited for time-series predictions in T1DM patients.

Early studies introduced neural networks for BGL predictions. Zecchin et al. [21] implemented an artificial neural network with polynomial extrapolation for 30-minute BGL forecasting, laying the foundation for ML applications in healthcare. Mirshekarian et al. [22] later showed that LSTM models outperformed SVR for 30- and 60-minute BGL predictions, demonstrating DL's advantage in managing temporal dependencies. Li et al. [23] further improved the accuracy by combining CNN and RNN in a hybrid model, effectively handling CGM data by leveraging both spatial and temporal feature extractions.

In addition to LSTM and CNN architectures, ensemble learning has emerged as a promising approach for improving BGL prediction accuracy. Alfian et al. [24] demonstrated that XGBoost outperformed models such as SVR and decision trees, yielding lower RMSE values for 30- and 60-minute forecasts. Liu et al. [25] further validated ensemble methods, using a tree-based ensemble with majority voting-based feature selection, which enhanced the accuracy by identifying key glucose indicators. Optimizations such as BO, as applied by Wang et al. [26] to a LightGBM model, showed significant performance gains. RNNs also remain central in glucose forecasting, with El Idrissi et al. [11] and Sun et al. [27] confirming superiority of LSTM over AR models and baseline methods for sequential data prediction in healthcare.

Literature shows that deep learning techniques, particularly LSTM and CNN models, outperform traditional methods in BGL prediction. Integrating ensemble learning and optimization techniques, such as BO further boosts predictive accuracy, indicating a promising future for these models in real-time diabetes management systems.

4 Framework Design

This section outlines our empirical framework for BGL forecasting in T1DM, covering data preparation, performance measures, framework architecture, single-model configurations, and ensemble construction. We also describe the significance tests used to evaluate the model performance. Experimental evaluations were conducted on Google Colaboratory to leverage GPU capabilities, using Keras 2.3.1 [28] with TensorFlow API 2.2.0 [29].

4.1 Data Description

The Diabetes Research in Children Network (DirecNet) dataset [30] includes historical BGL data from 110 T1DM patients aged 3–17, collected from 2000 to 2002. With BGL values recorded every 5 min via CGM sensors, the dataset spans 73,085 records, ranging from 40 to 400 mg/dl with an average of 177.90 mg/dl. To ensure data quality, only patients with at least 200 records were included, resulting in 108 patients with an average of 620 BGL records per patient.

4.2 Performance Measures

To assess BGL predictions, we used MAE and RMSE metrics, which provide insights into model accuracy by measuring the closeness between predicted values $\hat{x_i}$ and actual values x_i [31]. These metrics quantitatively evaluate prediction accuracy, with the formulas for MAE and RMSE given as follows:

– **Mean Absolute Error:**

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{x_i} - x_i| \tag{1}$$

– **Root Mean Square Error:**

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{x_i} - x_i)^2} \tag{2}$$

To improve our BGL prediction accuracy, we employed the MMRE and PRED metrics, which provided insights into the comparison between predicted and actual values [10].

– **Magnitude Relative Error:**

$$MRE = \left|\frac{x - \hat{x}}{x}\right| \times 100 \tag{3}$$

– **Mean Magnitude Relative Error:**

$$MMRE = \frac{\sum_{i=1}^{n}\left|\frac{x_i - \hat{x}_i}{x_i}\right| \times 100}{n} \tag{4}$$

– **Prediction Level:** The PRED metric is determined by the proportion of predictions that fall within a specified threshold p, which is conventionally set to 25%. The formula for PRED is given in Equation (5).

$$Pred(p) = \frac{K}{n} \times 100 \tag{5}$$

In this context, x and $\hat{x}$ denote the actual and predicted values, respectively, and n represents the total number of predictions. In addition, K signifies the number of predictions with an MMRE that falls within the specified error threshold p.

4.3 Significance Tests

To validate the results and facilitate model comparison, we employed statistical tests, beginning with the SK test. This non-parametric method ranks models based on performance by utilizing hierarchical clustering within an ANOVA to identify significant differences [32]. Furthermore, we implemented the BC method to aggregate the SK rankings. The BC method allocates points according to rank, thereby ensuring fairness and minimizing bias [33]. In our study, the BC method was applied across four metrics (RMSE, MAE, MMRE, and PRED), with the technique achieving the highest score considered the most favorable.

4.4 Abbreviations Used

To shorten the names of singles and ensembles, the following abbreviation rule was adopted:

– **For single models,** the format used is:
Optimization technique-Single model name
where optimization techniques are abbreviated as follows: R for RS, P for PSO, and B for BO.
For example, PGRU refers to a GRU model optimized using PSO.
– **For heterogeneous ensembles,** the format is:
E-Optimization technique-Combination rule
where combination rules are abbreviated as follows: AV for Mean, ME for Median, IR for Inverse Ranked Weighted Mean.
For example, EPAV refers to a heterogeneous ensemble composed of five ML/DL single models optimized using PSO, with the mean used as the combination rule.

4.5 Framework Architecture

The framework architecture used in this study is outlined in seven distinct steps, as illustrated in Fig. 1. These steps comprise the entire process from data preparation to model optimization, training, and evaluation. A comprehensive description of each step of the framework is provided below.

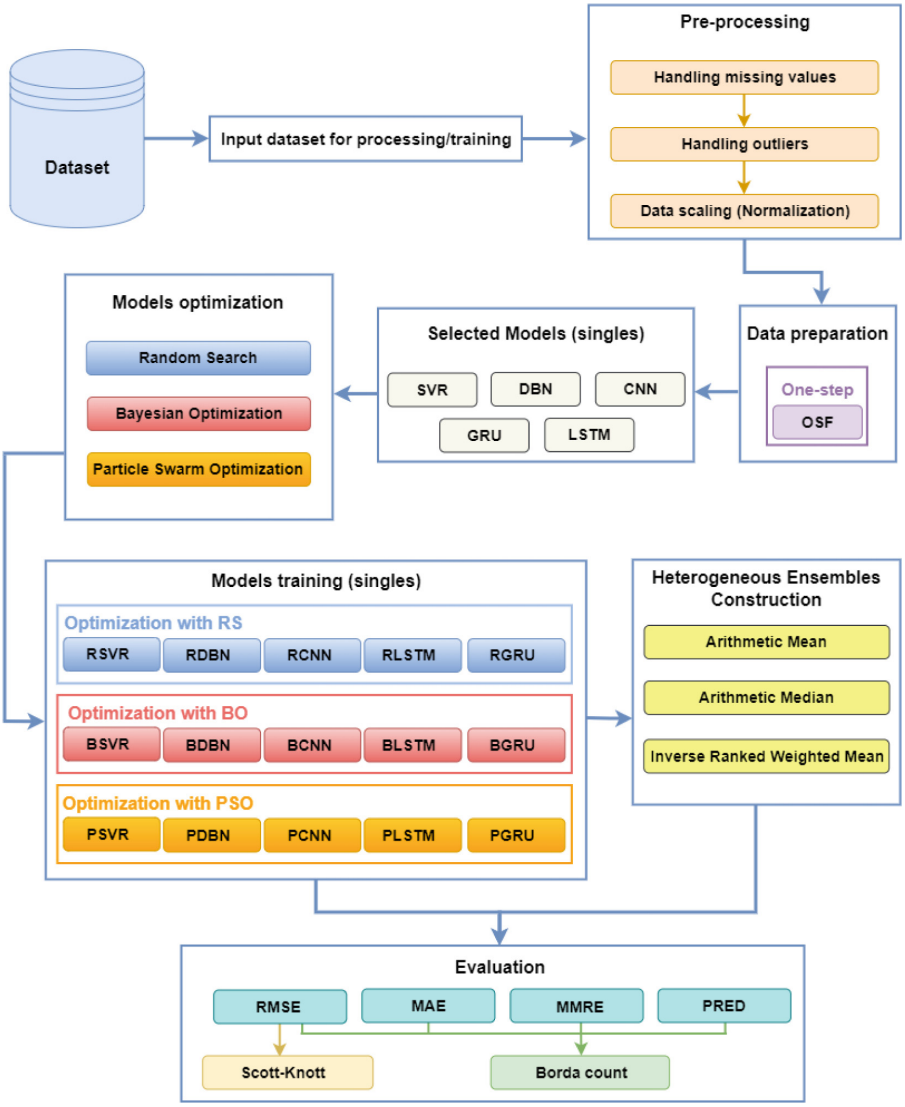

Fig. 1. Overview of the framework architecture.

Step 1) Pre-processing: The initial phase involved importing the dataset for further processing and training. Pre-processing is a crucial part of this step, where missing values are handled, outliers are addressed, and data scaling is performed. This ensures that the data are clean and standardized, which is essential for accurate modeling.

Step 2) Data Preparation for Models: In this step, the prepared data are organized for modeling using OSF forecasting strategies.

Step 3) Model Selection and Optimization: Models were optimized using BO, RS, and PSO to explore the key parameter ranges. For GRU and LSTM, the hidden layer units were tuned, with GRU exploring a wider range (64–128 for the first layer, 32–64 for the second) than LSTM (32–64 for the first, 16–32 for the second). CNN optimization included filter counts in convolution layers and hidden units (32–128), while SVR focused on continuous hyperparameters like regularization C (0.1–2) and kernel parameter γ (0.1–1). DBNs optimized hidden layer units (32-256), learning rates (0.01–1), and training parameters. BO used discrete unit/filter ranges, RS uniformly sampled parameters, and PSO employed lower and upper bounds for a global search.

Step 4) Model Training: This phase involves training each selected and optimized model on the prepared data, ensuring that they are well-fitted and capable of accurate predictions.

Step 5) Heterogeneous Ensembles Construction: After training individual models, we constructed heterogeneous ensembles using methods such as Arithmetic Mean (AV), Arithmetic Median (ME), and Inverse Ranked Weighted Mean (IR). These ensembles combine the model strengths to enhance the overall prediction accuracy through model diversity.

Step 6) Model Evaluation: The models and ensembles are evaluated using the RMSE, MAE, MMRE, and PRED metrics, which quantitatively assess prediction accuracy by measuring the alignment with actual values.

Step 7) Model Comparison: Significance tests, including the SK and BC test, were used to statistically compare the models. The SK test ranks models by performance, whereas the BC method aggregates these rankings to select top-performing models, ensuring a fair and unbiased comparison.

5 Results and Discussion

This section addresses key research questions by analyzing the model performance for BGL forecasting. Performance metrics and SK test results provide detailed model comparisons, whereas BC rankings supplement the analysis with comparative model rankings. Each research question was answered through an in-depth discussion based on these results (Tables 1, 2, 3).

Table 1. Performance comparison of RS ensembles and their singles for OSF.

Techniques	RLSTM	RGRU	RCNN	RSVR	RDBN	ERAV	ERME	ERIR
RMSE(mg/dl)	13.27	**8.03**	24.69	12.17	18.52	13.63	12.25	12.92
MAE(mg/dl)	10.19	**6.13**	19.98	9.66	14.29	10.47	9.31	10.05
MMRE(mg/dl)	6.59	**3.89**	11.71	6.22	8.98	6.3	5.76	6.31
PRED(25%)	98.36	**99.92**	95.44	99.35	95.74	99.05	99.09	98.92

Table 2. Performance comparison of PSO ensembles and their singles for OSF

Techniques	RLSTM	RGRU	RCNN	RSVR	RDBN	ERAV	ERME	ERIR
RMSE(mg/dl)	12.98	**7.87**	22.6	12.17	15.02	12.84	12.26	12.58
MAE(mg/dl)	9.88	**6.09**	17.89	9.62	11.58	9.79	9.33	9.71
MMRE(mg/dl)	6.35	**3.84**	11.57	6.11	7.46	6.12	5.83	6.31
PRED(25%)	98.79	**99.98**	94.42	99.41	97.68	99.22	99.14	99.07

Table 3. Performance comparison of BO ensembles and their singles for OSF

Techniques	RLSTM	RGRU	RCNN	RSVR	RDBN	ERAV	ERME	ERIR
RMSE(mg/dl)	13.25	**7.96**	23.98	15.62	22.64	14.61	13.21	14.01
MAE(mg/dl)	10.18	**6.15**	19.37	12.43	17.62	11.31	9.92	10.67
MMRE(mg/dl)	6.51	**3.8**	12.22	7.89	11.19	6.95	6.14	6.47
PRED(25%)	98.76	**99.97**	93.31	97.1	92.06	98.52	98.64	98.83

5.1 Evaluation of Singles for BGL Forecasting. (RQ1)

The PGRU model was identified as the best-performing single model with an RMSE of 7.87 mg/dl, making it the most accurate model for BGL forecasting.

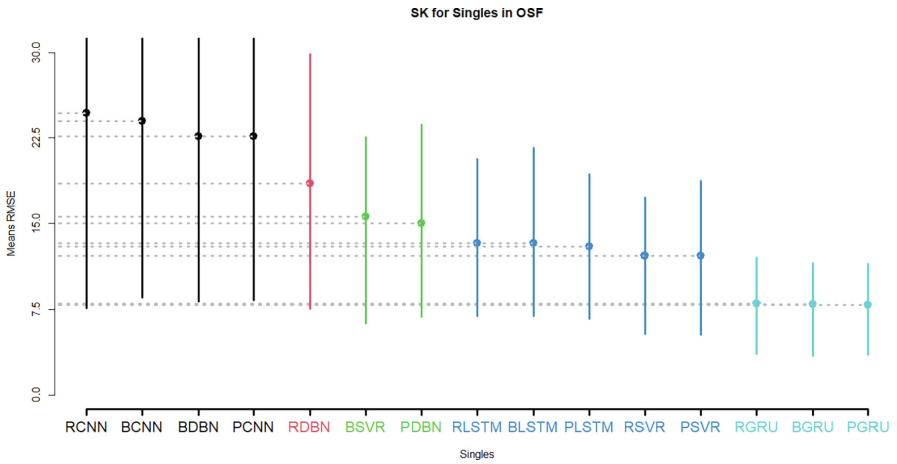

Fig. 2. Plot of SK test for singles in OSF.

Figure 2 presents the SK test results, identifying PGRU as statistically superior, with other GRU-based models like RGRU and BGRU closely following (RMSEs of 8.03 mg/dl and 7.96 mg/dl, respectively) in the top performance group. This grouping underscores the effectiveness of GRU models for BGL

forecasting, given their ability to capture long-term dependencies. In contrast, CNN-based models, such as RCNN, with higher RMSEs (e.g., 24.69 mg/dl), were placed in a lower performance group, highlighting GRU architectures as better suited for time-series predictions in this domain.

Table 4. BC ranking for singles, ensembles and their combination for OSF of the top 10 best models based on SK test.

Techniques	Singles		Ensembles		Ensembles and their Singles	
BC Rank	Model	Score	Model	Score	Model	Score
1st	PGRU	59	ERME	34	PGRU	95
2nd	BGRU	56	EPME	32	BGRU	92
3rd	RGRU	53	EPAV	28	RGRU	89
4th	PSVR	48	EPIR	24	PSVR	80
5th	RSVR	44	EBME	17	ERME	78
6th	PLSTM	40	ERAV	16	EPME	76
7th	BLSTM	36	ERIR	16	RSVR	74
8th	RLSTM	32	EBIR	9	EPAV	69
9th	PDBN	28	EBAV	4	EPIR	64
10th	BSVR	24	-	-	ERIR	55

In addition to the SK results, the BC rankings presented in Table 4 reinforce the performance of PGRU, placing it first among single models with a score of 59 and first overall, including ensembles, with a combined score of 95. This highlights the consistency of PGRU across various metrics and optimization techniques.

5.2 Evaluation of Heterogeneous Ensembles for BGL Forecasting. (RQ2)

ERME was the most accurate and reliable heterogeneous ensemble, with an RMSE of 12.25 mg/dl. The SK test presented in Fig. 3, grouped this ensemble into the highest performance cluster. The EPME ensemble also performed well, with an RMSE of 12.26 mg/dl, and was placed in the same performance group, demonstrating that PSO can still provide competitive results in ensemble configurations.

The SK results indicated that the EBAV ensemble, with an RMSE of 14.61 mg/dl, performed lower, suggesting BO is less effective in tuning ensembles than RS and PSO for BGL forecasting. The BC rankings in Table 4 further contextualize ensemble performance: ERME ranked first with a score of 34, followed by EPME with 32, and EPAV ranked third with 28. This underscores the advantages of RS and PSO in enhancing the ensemble accuracy and reliability.

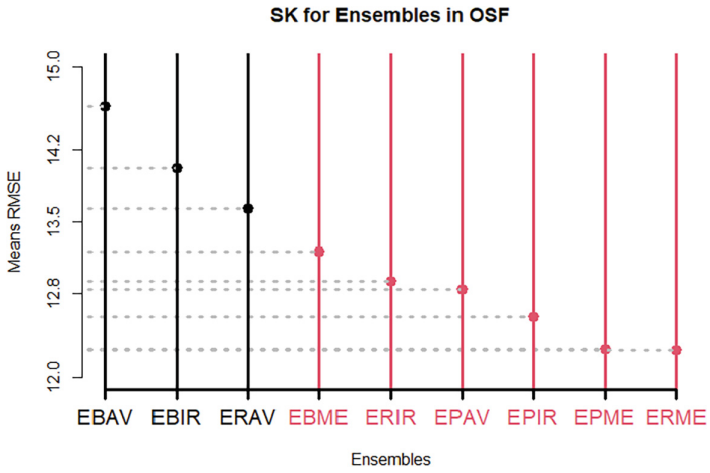

Fig. 3. Plot of SK for heterogeneous ensembles in OSF.

Table 5. Comparison of ensembles and their singles for 5 min forecasting.

	Techniques	Min	Mean	Median	Max
RS	Singles	8.03	15.34	13.27	24.69
	Ensembles	12.25	12.93	12.92	13.63
PSO	Singles	7.87	14.13	12.98	22.6
	Ensembles	12.26	12.56	12.58	12.84
BO	Singles	7.96	16.69	15.62	23.98
	Ensembles	13.21	13.14	14.01	14.61
All	Singles	7.87	15.38	13.27	24.69
	Ensembles	12.25	13.14	12.92	14.61

5.3 Evaluation of Heterogeneous Ensembles Vs Their Constituent Singles in BGL Forecasting. (RQ3)

Figure 4 illustrates the SK test, which identifies single models as belonging to the top performance cluster. However, heterogeneous ensembles consistently outperformed single models across key metrics, such as the mean, median, and maximum RMSE, as shown in Table 5. For instance, while PGRU achieved a low RMSE of 7.87 mg/dl, ensembles showed more stable and consistent performance. Notably, single models reached a maximum RMSE of 24.69 mg/dl, whereas ensembles had a lower maximum RMSE of 14.61 mg/dl, demonstrating their ability to reduce extreme errors and enhance the reliability of BGL forecasting. Across all optimization techniques (PSO, RS, and BO), ensemble models achieved consistently lower mean and median RMSE, proving their effectiveness in improving both accuracy and stability compared to single models.

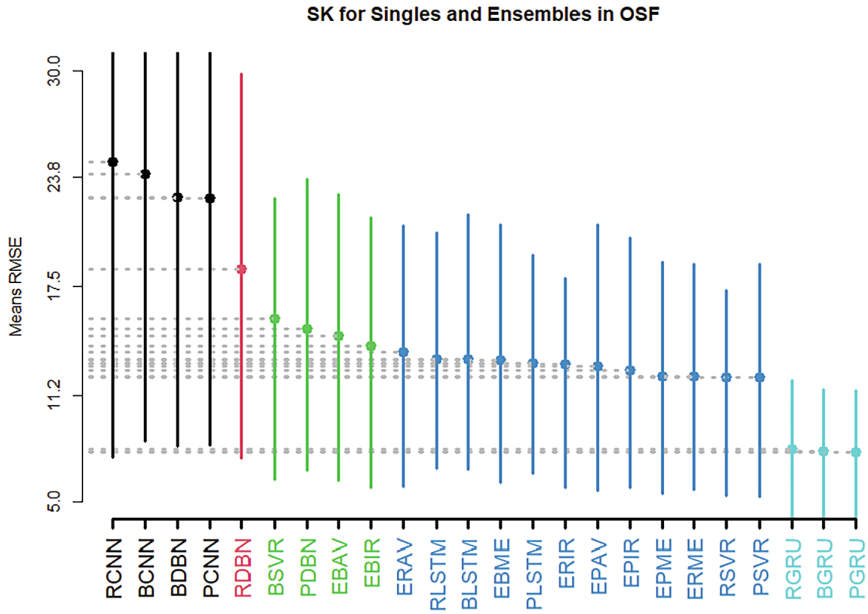

Fig. 4. Plot of SK for singles and ensembles in OSF.

5.4 Evaluation of Optimization Techniques of Singles and Ensembles for BGL Forecasting. (RQ4)

PSO has emerged as the most effective optimization technique for both individual models and ensembles. The SK test revealed that the PSO-optimized models and ensembles consistently ranked in the highest performance groups, with PGRU and EPAV achieving the highest RMSE scores. In contrast, RS also provided competitive results, particularly for ensembles like ERME (RMSE: 12.25 mg/dl), but showed greater variability when applied to individual models, as evidenced by RCNN's poor performance.

BO, while effective in certain cases, was placed in a lower performance group according to the SK test. For instance, BGRU performed well with an RMSE of 7.96 mg/dl, but EBME had higher variability, with an RMSE of 13.21 mg/dl, leading it to be ranked in a lower performance cluster.

Overall, PSO's ability to balance global and local optimization searches makes it the most reliable method, especially for optimizing complex models such as GRU and LSTM, as well as ensemble configurations. The SK test confirmed the superiority of PSO, statistically placing its optimized models ahead of the others across both individual and ensemble configurations.

Table 6 prensents the number of occurences of each optimization technique in the BC analysis reinforcing the effectiveness of PSO. PSO-optimized models appeared 12 times in the top 10 rankings, accumulating a total score of 643, while RS-optimized models appeared 10 times with a score of 491, and BO-

Table 6. Number of occurrences of each optimization technique for top 10 best models based on SK test for singles, ensembles and both.

	Singles		Ensembles		Ensembles and Singles		Total	
	Occurrences	Score	Occurrences	Score	Occurrences	Score	Occurrences	Score
RS	3	129	3	66	4	296	10	491
BO	3	116	3	30	1	92	7	238
PSO	4	175	3	84	5	384	12	643

optimized models appeared 7 times with a score of 238. These results highlight the consistency and effectiveness of PSO in optimizing both single models and heterogeneous ensembles for BGL forecasting.

The findings of this study, supported by the SK test and BC rankings, indicate that PSO-optimized ensembles are the most effective approach for BGL forecasting. The GRU-based models optimized with PSO consistently performed better than the other models, confirming that recurrent architectures are well-suited for handling temporal dependencies in glucose data. Additionally, the superiority of PSO for ensemble configurations suggests that evolutionary optimization techniques should be considered in the future developments of BGL forecasting systems.

6 Threats to Validity

This section considers potential threats to the validity of our study, classified into three categories (internal, external and construct validity). We describe the actions undertaken to counter these threats, reinforcing the robustness and reliability of our findings.

Internal Validity. Internal validity refers to ensuring reliable results and minimizing bias. We used 4-fold cross-validation, which is more robust than a random train-test split and reduces overfitting. In addition, an early stopping mechanism was applied to prevent overfitting during training. To avoid bias from poor hyperparameter choices, optimization techniques, such as RS, BO, and PSO were employed to fine-tune the models effectively.

External Validity. External validity concerns the generalizability of research findings. To ensure broader applicability, we used a dataset of 110 patients with diabetes that offered diverse representation. This contrasts with studies that used smaller datasets and improves the relevance of our findings. We focused on 5-minute prediction horizons, although factors such as physical activity, diet, and stress were not fully considered. Testing these external factors in real-world settings is crucial for validating their practical applicability.

Construct Validity. Construct validity refers to the extent which our performance measures reflect the underlying constructs of model performance. To strengthen the construct validity, we used four metrics: RMSE, MAE, MMRE, and PRED, to ensure a comprehensive evaluation. We applied the SK and BC methods for the balanced model assessment. The SK test clusters models based on RMSE, minimizing variance within clusters and maximizing variance between them, supported by statistical tests such as ANOVA. This method provides a clear ranking of models based on their performance. The BC method aggregates rankings from multiple metrics (MAE, RMSE, MMRE, PRED), assigns points to models and determines the overall rank. This approach ensures fair comparison by simultaneously considering multiple criteria.

7 Conclusion and Future Works

In this study, we evaluated various ML/DL models, including LSTM, CNN, GRU, SVR, and DBN, as well as heterogeneous ensembles, for BGL forecasting in individuals with type 1 diabetes. We aimed to identify the most accurate and reliable models for OSF using optimization techniques such as PSO, RS, and BO.

RQ1: Is there any single predictive model that outperforms others for BGL forecasting?

The results showed that GRU optimized using PSO outperformed the other models. The PGRU model achieved the best performance, with an RMSE of 7.87 mg/dl, indicating its effectiveness in capturing short-term dependencies in BGL data. This highlights the suitability Of GRU for time-series forecasting in diabetes management.

RQ2: Which heterogeneous ensemble provides the most accurate and reliable BGL predictions?

Among the ensemble configurations, the ERME model provided the most reliable forecasts, achieving an RMSE of 12.25 mg/dl, followed by EPME with a quite similar RMSE of 12.26 mg/dl. The combination of different models in the ensemble mitigated the weaknesses of the single models, leading to more stable and accurate predictions.

RQ3: How do heterogeneous ensembles compare to their constituent singles in BGL forecasting?

Heterogeneous ensembles consistently outperformed their constituent single models, showing more robust performance. For example, while the PGRU model had a lower RMSE, the ensemble models provided more balanced and reliable predictions across different metrics, particularly in mitigating single model variability.

RQ4: How do different optimization techniques improve the performance of single and ensemble models for BGL forecasting?

PSO has emerged as the most effective optimization technique, significantly enhancing the performance of both single models and heterogeneous ensembles. The PSO-optimized models consistently ranked among the top performance groups, as validated by the SK test and BC rankings.

Future research should explore long-term multi-step forecasting, incorporate additional physiological data, and develop adaptive models. Additionally, further investigation into other evolutionary optimization techniques, such as genetic algorithms, could improve the model performance. These approaches will enhance real-time diabetes management systems.

References

1. Burrack, A., Martinov, T., Fife, B.: T cell-mediated beta cell destruction: autoimmunity and alloimmunity in the context of type 1 diabetes. Front. Endocrinol. **8** (2017)
2. Chiang, J., et al.: Type 1 diabetes in children and adolescents: a position statement by the american diabetes association. Diabetes Care **41**, 2026–2044 (2018)
3. Mattila, T., Boer, A.: Influence of intensive versus conventional glucose control on microvascular and macrovascular complications in type 1 and 2 diabetes mellitus. Drugs **70**, 2229–2245 (2010)
4. Colberg, S., Kannane, J., Diawara, N.: Physical activity, dietary patterns, and glycemic management in active individuals with type 1 diabetes: an online survey. Int. J. Environ. Res. Public Health. **18** (2021)
5. Kovatchev, B., Gonder-Frederick, L., Cox, D., Clarke, W.: Evaluating the accuracy of continuous glucose-monitoring sensors: Continuous glucose-error grid analysis illustrated by TheraSense Freestyle Navigator data. Diabetes Care. **27** (2004)
6. Benaida, M., Abnane, I., Idri, A., El Idrissi, T.: Machine and deep learning predictive techniques for blood glucose level. Inform. Syst. Technol, pp. 476–485 (2022)
7. El Idrissi, T., Idri, A.: Evaluating a comparing deep learning architectures for blood glucose prediction. In: Biomedical Engineering Systems And Technologies: 13th International Joint Conference, BIOSTEC 2020, Valletta, Malta, February 24–26, 2020, Revised Selected Papers 13, pp. 347–365 (2021)
8. Shahid, S., Hussain, S., Khan, W.: Predicting continuous blood glucose level using deep learning. In: Proceedings of The 14th IEEE/ACM International Conference on Utility and Cloud Computing Companion, pp. 1–5 (2021)
9. Kitchainukoon, N., Sae-Bae, N.: Predicting blood glucose levels with machine learning techniques. In: 2023 IEEE 6th International Conference On Knowledge Innovation And Invention (ICKII), pp. 515–519 (2023)
10. Benaida, M., Abnane, I., Idri, A.: Deep learning based one step and multi-steps ahead forecasting blood glucose level. Expert Systems, pp. e13393 (2023)
11. El Idrissi, T., Idri, A.: Deep learning for blood glucose prediction: CNN vs LSTM. In: Gervasi, O., Murgante, B., Misra, S., Garau, C., Blečić, I., Taniar, D., Apduhan, B.O., Rocha, A., Tarantino, E., Torre, C.M., Karaca, Y. (eds.) ICCSA 2020. LNCS, vol. 12250, pp. 379–393. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58802-1_28
12. Wadghiri, M., Idri, A., Idrissi, T., Hakkoum, H.: Ensemble blood glucose prediction in diabetes mellitus: a review. Comput. Biol. Med. **147** (2022)
13. Vapnik, V., Golowich, S., Smola, A., et al.: Support vector method for function approximation, regression estimation, and signal processing. Adv. Neural Inform. Process. Syst. 281–287 (1997)
14. Ren, Y., Mao, J., Liu, Y., Li, Y.: A novel DBN model for time series forecasting. IAENG Int. J. Comput. Sci. **44**, 79–86 (2017)

15. LeCun, Y., Bengio, Y.: Convolutional networks for images, speech, and time series. In: The Handbook of Brain Theory And Neural Networks, pp. 255–258 (1998)
16. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**, 1735–1780 (1997)
17. Cho, K., Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using rnn encoder–decoder for statistical machine translation. In: Proceedings of The 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1724–1734 (2014,10). https://aclanthology.org/D14-1179
18. Bergstra, J., Bengio, Y.: Random search for hyper-parameter optimization. J. Mach. Learn. Res. **13** (2012)
19. Frazier, P.: A tutorial on Bayesian optimization. ArXiv Preprint ArXiv:1807.02811 (2018)
20. Wang, D., Tan, D., Liu, L.: Particle swarm optimization algorithm: an overview. Soft. Comput. **22**, 387–408 (2018)
21. Zecchin, C., Facchinetti, A., Sparacino, G., De Nicolao, G., Cobelli, C.: Neural network incorporating meal information improves accuracy of short-time prediction of glucose concentration. IEEE Trans. Biomed. Eng. **59**, 1550–1560 (2012)
22. Mirshekarian, S., Shen, H., Bunescu, R., Marling, C.: LSTMs and neural attention models for blood glucose prediction: comparative experiments on real and synthetic data. In: 2019 41st Annual International Conference Of The IEEE Engineering In Medicine And Biology Society (EMBC), pp. 706–712 (2019)
23. Li, K., Daniels, J., Liu, C., Herrero-Vinas, P., Georgiou, P.: Convolutional recurrent neural networks for glucose prediction. IEEE J. Biomed. Health Inform. **24**, 603-613 (2020) https://nottingham-repository.worktribe.com/output/2302666
24. Alfian, G., Syafrudin, M., Rhee, J., Anshari, M., Mustakim, M., Fahrurrozi, I.: Blood glucose prediction model for type 1 diabetes based on extreme gradient boosting. IOP Conf. Series: Materials Sci. Eng. **803** (2020)
25. Liu, J., Wang, L., Zhang, L., Zhang, Z., Zhang, S.: Predictive analytics for blood glucose concentration: an empirical study using the tree-based ensemble approach. Library Hi Tech. **38**, 835–858 (2020)
26. Wang, Y., Wang, T.: Application of improved LightGBM model in blood glucose prediction. Appl. Sci. (Switzerland). **10** (2020)
27. Sun, Q., Jankovic, M., Bally, L., Mougiakakou, S.: Predicting blood glucose with an LSTM and Bi-LSTM based deep neural network. In: 2018 14th Symposium on Neural Networks and Applications (NEUREL), pp. 1–5 (2018)
28. Chollet, F., et al.: Keras. Seattle, WA, USA (2015)
29. Abadi, M., et al. TensorFlow: a system for Large-Scale machine learning. In: 12th USENIX Symposium on Operating Systems Design And Implementation (OSDI 16), pp. 265–283 (2016)
30. DirecNet Diabetes Research in Children Network (DirecNet). (2020 3). https://public.jaeb.org/direcnet/
31. El Idrissi, T., Idri, A., Kadi, I., Bakkoury, Z.: Strategies of multi-step-ahead forecasting for blood glucose level using LSTM neural networks: A comparative study. In: HEALTHINF 2020 - 13th International Conference on Health Informatics, Proceedings; Part of 13th International Joint Conference on Biomedical Engineering Systems And Technologies, BIOSTEC 2020, pp. 337–344 (2020)
32. Scott, A., Knott, M.A.: Cluster analysis method for grouping means in the analysis of variance. Biometrics. **30**, 507–512 (1974,9)
33. Borda, J.: Memoire sur les elections au scrutin. Mémoires De L'académie Royale, pp. 657–664 (1781)