Enhancing Bagging Ensemble Regression with Data Integration for Time Series-Based Diabetes Prediction

Vuong M. Ngo[1,2] ✉, Tran Quang Vinh[3] ✉, Patricia Kearney[4], and Mark Roantree[1]

[1] Insight Centre, School of Computing, Dublin City University, Dublin, Ireland
[2] Ho Chi Minh City Open University, Ho Chi Minh City, Vietnam
[3] Ho Chi Minh City University of Transport, Ho Chi Minh City, Vietnam
[4] School of Public Health, University College Cork, Cork, Ireland
vuong.ngo@dcu.ie or vuong.nm@ou.edu.vn, vinh.tran@ut.edu.vn, patricia.kearney@ucc.ie, mark.roantree@dcu.ie

Abstract. Diabetes is a chronic metabolic disease characterized by elevated blood glucose levels, leading to complications like heart disease, kidney failure, and nerve damage. Accurate state-level predictions are vital for effective healthcare planning and targeted interventions, but in many cases, data for necessary analyses are incomplete. This study begins with a data engineering process to integrate diabetes-related datasets from 2011 to 2021 to create a comprehensive feature set. We then introduce an enhanced bagging ensemble regression model (EBMBag+) for time series forecasting to predict diabetes prevalence across U.S. cities. Several baseline models, including SVMReg, BDTree, LSBoost, NN, LSTM, and ERMBag, were evaluated for comparison with our EBMBag+ algorithm. The experimental results demonstrate that EBMBag+ achieved the best performance, with an MAE of 0.41, RMSE of 0.53, MAPE of 4.01, and an R^2 of 0.91.

1 Introduction

Diabetes is a chronic metabolic disease characterized by elevated blood glucose levels, which can cause severe damage to the heart, blood vessels, eyes, kidneys, and nerves over time. Globally, approximately 830 million people live with diabetes, primarily in low- and middle-income countries, with 1.5 million deaths attributed to the disease annually [29]. In the U.S., diabetes presents a major public health challenge, affecting 38.4 million people, or 11.6% of the population, with substantial state-level variations influenced by demographic and economic factors [5]. Accurate prediction of diabetes prevalence by state is essential for effective healthcare planning, resource allocation, and targeted interventions, especially considering the high rates among seniors (29.2% or 16.5 million) and youth (0.35% or 352 thousands under 20). Diabetes also disproportionately affects racial and ethnic minorities, highlighting the need for culturally tailored approaches.

In the U.S., the disease's economic toll is significant, costing $412.9 billion in 2022, with individuals with diabetes incurring healthcare costs 2.6 times higher than those without [1]. Given that diabetes was the eighth leading cause of death in 2021, accurate state-level predictions are crucial to reducing the health and economic burden of diabetes nationwide and improving population health outcomes.

In recent years, data mining and ML have become essential, reliable tools in the medical field. Data mining is used to preprocess healthcare data and select relevant features, while ML automates the prediction of conditions like diabetes. Many studies require careful integration of data that has been sourced from separate, often heterogenous, databases and repositories [23]. Then, by uncovering hidden patterns in the data, complex ML methods may enable accurate and reliable decision-making [15], [11]. However, recent studies have rarely incorporated time-series features in their models and have primarily focused on predicting diabetes at the individual level rather than at a broader, city-wide scale. Specifically, to improve performance, ML models should be enhanced to better adapt to healthcare applications and disease detection.

Our contribution can be articulated as follows:

- Engineering a novel diabetes dataset containing a comprehensive feature set suitable for machine learning algorithms by integrating various U.S. diabetes-related data sources.
- The application of time series techniques to several popular ML models to establish baseline models for diabetes prediction.
- The development of an enhanced bagging ensemble regression model (EBM-Bag+) that incorporates time series techniques for predicting diabetes prevalence.
- Delivering a robust evaluation framework incorporating the analysis of EBM-Bag+ against baseline models to demonstrate its superior performance.

2 Related Work

Several studies have applied ML algorithms to predict diabetes using patients' medical records, including [25], [14], [9], [17], [8] and [19]. In [25], the authors used the Pima Indian diabetes dataset and collected additional samples from 203 individuals working at a local textile factory in Bangladesh, focusing on six features: pregnancy, glucose level, blood pressure, skin thickness, BMI, age, and diabetes outcome. To address class imbalance in the dataset and enhance ML performance, they applied Synthetic Minority Over-sampling Technique and Adaptive Synthetic Sampling.

In [14], five ML algorithms were employed to classify binary outcomes, focusing specifically on type 2 diabetes. Type 2 diabetes is characterized by the body's reduced ability to use insulin effectively, often linked to lifestyle factors. The authors used a Kaggle dataset containing 768 patient records with nine features: number of pregnancies (for female patients), glucose level, diastolic blood

pressure, skinfold thickness, insulin level, body mass index, family history of diabetes, age, and diabetes outcome (yes/no).

In [9], the authors proposed using the XGB method in their mobile app to predict diabetes. Their dataset includes 300 data samples from volunteers, collected from specialty hospitals in Saudi Arabia and Egypt during the 2022-2023 academic year. A key contribution of this study is the development of a mobile app that allows users to input relevant features and instantly receive a diabetes prediction.

In [17], the SHapley Additive Explanation technique was used to identify the most influential factors for predicting the 10-year risk of developing type 2 diabetes, followed by the use of the XGBoost model for diabetes classification. Using a dataset of 12,148 participants with type 2 diabetes, they concentrated on the top ten features associated with diabetes risk. HbA1c emerged as the strongest predictor, followed by BMI, waist circumference, and blood glucose levels.

In [8], the authors presented an intuitive, self-explanatory interface for diabetes prediction, utilizing four ML algorithms: Decision Tree (DT), K-nearest Neighbor (KNN), Support Vector Classification (SVC), and Extreme Gradient Boosting (XGB). The authors applied these algorithms to open-source clinical data, focusing on features such as pregnancies, glucose, blood pressure, and insulin levels.

In [19], the authors introduced a diabetes prediction model that utilized multiple ML techniques. Specifically, the study employed algorithms such as Logistic Regression, SVM, Naïve Bayes, and Random Forest, alongside various ensemble learning methods, including XGBoost, LightGBM, CatBoost, AdaBoost, and Bagging. These ensemble methods integrate predictions from multiple base learners to enhance the model's accuracy and robustness. The models were evaluated on a Kaggle dataset containing 5,000 patient records.

Summary. In the studies mentioned above, none of ([25], [14], [9], [17], [8], and [19]) incorporated time-series features into their models. Furthermore, these studies focused on predicting diabetes at the individual patient level rather than across broader populations and thus, may not be suited to policy level decisions which require greater numbers to better inform decision making. Similar to our approach, the research in [16] proposed a novel progressive self-transfer framework for time-series disease prediction. They used data from the Korean National Institute of Health, which included biannual medical check-up and survey information from participants aged 40 to 69 years between 2001 and 2018. The authors employed the least absolute shrinkage and selection operator (LASSO) for feature selection. However, they did not integrate datasets to obtain more comprehensive information. Notably, their work also focused on individual-level data rather than addressing predictions at the greater population level. Additionally, they introduced new features to existing models rather than improving the model itself.

3 Our Datasets

In addition to primary factors that increase an individual's risk of diabetes, such as glucose level, insulin level, and body mass index, we identified other potential state-level influences. These include: (1) chronic diseases such as alcohol consumption, smoking, asthma, and high cholesterol; (2) demographic factors like race and gender; (3) housing conditions, including home ownership or rental status; and (4) economic factors such as employment status, income, and poverty levels. To explore these relationships, we utilized four freely available datasets covering chronic diseases, population demographics, housing, and economic conditions across U.S. states from 2011 to 2021.

3.1 Separate Datasets

U.S. CDI: The chronic disease indicator (CDI) dataset was sourced from the U.S. Centers for Disease Control and Prevention (CDC) [5]. In collaboration with the Council of State and Territorial Epidemiologists and the National Association of Chronic Disease Directors, the CDC has developed a comprehensive suite of 115 chronic disease indicators. These indicators enable consistent definitions, data collection, and reporting across states and territories, supporting public health initiatives and allowing for uniform tracking and analysis of chronic diseases across regions.

The dataset, US-CDI Version 13 [6], contains 34 columns, with 24 populated by health-related data from multiple sources covering the years 2001 to 2021. It provides information on 17 distinct chronic diseases for each U.S. state and includes demographic details such as race and gender, organized by topic and specific health-related questions.

U.S. population: The second dataset, obtained from the U.S. Census Bureau[6] and the U.S. National Cancer Institute[7], provides detailed demographic information on the U.S. population. It includes race data classified into five categories: Non-Hispanic White, Non-Hispanic Black, Non-Hispanic American Indian/Alaska Native, Non-Hispanic Asian or Pacific Islander, and Hispanic. Additionally, the dataset includes data on sex (male and female) and age, organized into 19 distinct groups, covering all U.S. states.

For this paper, data covering the period from 2011 to 2021 was used. Specifically, data from 2011 to 2020 was directly extracted from the dataset provided by the U.S. National Cancer Institute [21], while data for 2021 was estimated through linear interpolation based on population trends from the preceding decade and the 2021 U.S. Census [28]. This approach ensures a consistent and comprehensive demographic perspective aligned with the study period.

U.S. Housing: The third data source includes housing data for each state from the years 2010 and 2020, provided by the U.S. Census Bureau via the DEC

[5] https://www.cdc.gov/
[6] https://www.census.gov
[7] https://cancer.gov/

Redistricting Data (PL 94-171) product [26]. This dataset contains information on the total number of housing units, as well as occupied and vacant units across all 51 states. Linear interpolation was applied to estimate housing data for the years 2011 to 2019 and 2021 based on this dataset alongside the U.S. Census dataset.

Table 1: Categories and their feature counts in our dataset, presented as percentages for each U.S. state by year

Category	**Description**	**#Fea.**
Diabetes	The percentage of diagnosed diabetes among adults aged 18 and older	1
Age groups	The percentage of the population, categorized by age groups. The age groups are 0-19, 20-39, 40-59 and 60+	4
Races	The percentage of the population categorized by races. The racial groups include Hispanic, Non-Hispanic White, Non-Hispanic Asian or Pacific Islander, Non-Hispanic American Indian or Alaska Native, and Non-Hispanic Black	5
Gender	The percentages of males and females, as well as the percentage of the overall population	3
House	The percentage of total houses, as well as The percentages of vacant and occupied houses	3
Economy	The percentage of the employed population, per capita income, and the poverty rate	3
Chronic Disease Indicators	The percentage of adults diagnosed with at least one chronic disease (excluding diabetes) or a chronic disease indicator, such as asthma, arthritis, kidney disease, high cholesterol, smoking, or having had a foot examination, dilated eye examination, or glycosylated hemoglobin measurement	71
	Total Feature Count:	**90**

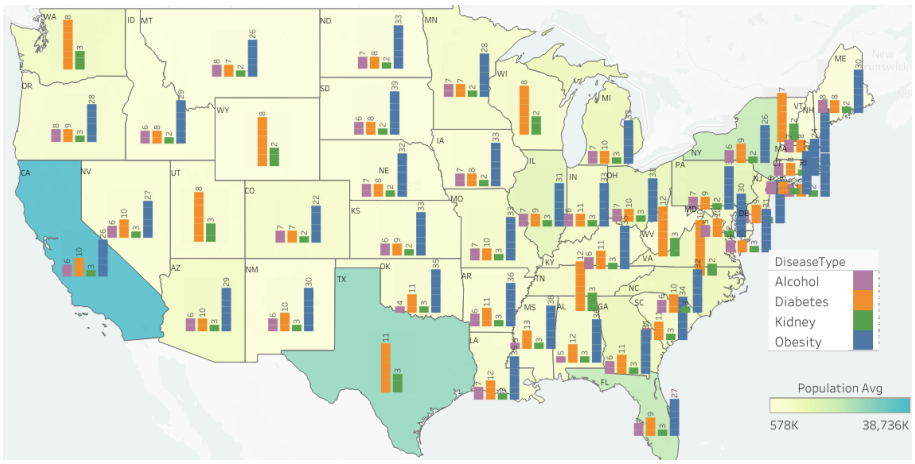

Fig. 1: Average Percentage for Diabetes, Alcohol Consumption, Kidney Disease, and Obesity across U.S. states

U.S. Economy: The fourth data source includes economic data for each U.S. state from 2011 to 2021, combining information from the U.S. Bureau of Labor Statistics [4], the U.S. Census Bureau [27], and the Bureau of Economic Analysis [3]. This dataset provides details on the employment rate, per capita income, and poverty rate across states over an 11-year period.

3.2 Our Integrated Diabetes Dataset

We have combined the four datasets into a unified dataset to predict the prevalence of diagnosed diabetes among adults aged 18 and older across all 51 U.S. states for the following year. This work is not covered here as the methodology is presented in earlier work in [22] where integration from semi-structured sources was described and in [23] where the deployment of the HL7 common model was used to integrate Covid-19 datasets.

Table 1 presents seven categories together with their descriptions and feature counts in this newly formed dataset. Each feature is standardized to display the percentage of the relevant feature for each state by year. The dataset contains 90 features, including one representing the percentage of adults diagnosed with diabetes. Figure 1 illustrates the average percentages over eleven years of populations in all states affected by diabetes, heavy alcohol consumption, kidney disease, and obesity.

4 Methodology

4.1 System Architecture

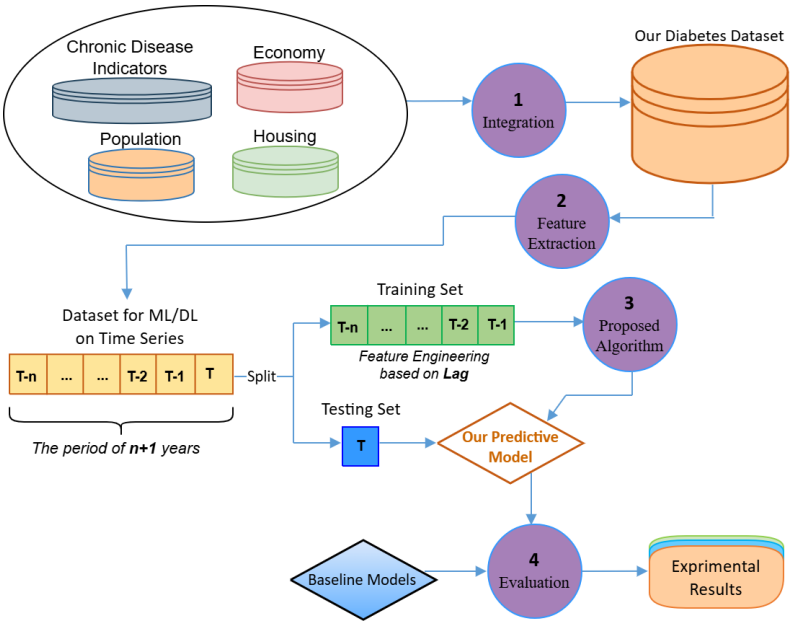

Fig. 2: System Architecture: Data Engineering & Predictive Algorithms

The system architecture, illustrated in Figure 2, outlines a structured workflow for implementing our algorithm to predict diabetes trends across different states

using time-series data. This architecture ensures a systematic approach to data integration, feature engineering, model development, and performance evaluation, facilitating the creation of an accurate and efficient predictive model.

The system consists of four primary processing modules. The first module, `Integration`, combines various datasets into our diabetes dataset, as discussed in Section 3. The second module, `Feature Extraction`, explains how features are used to create time-series training and testing sets, detailed in Section 4.2. The third module, `Proposed Algorithm`, uses the training set with engineered features based on *Lag* to build our diabetes prediction model, described in Section 4.4. Finally, the `Evaluation` module (discussed in Section 5) assesses our proposed model's predictive performance and computational efficiency on test dataset, comparing our approach with baseline models (outlined in Section 4.3).

4.2 Structuring Time Series Training and Testing Datasets

The objective of this study is to predict the diabetes class for a given year t. The input data includes 89 predictor features for each year: 4 features related to age, 5 related to race, 3 related to gender, 3 related to housing, 3 related to economic factors, and 71 related to chronic disease indicators, as shown in Table 1.

To capture the temporal dynamics of the data, the model incorporates lagged observations from previous years. For a given *Lag* l, where $l = 1$ to 9, the input dataset consists of $89 \times (l+1)+l$ variables. This includes 89 predictor variables for the current year t and additional sets of 89 predictors for each of the preceding years up to $t-l$, along with l corresponding target output values for years $t-1$ through $t-l$.

For example, with $Lag = 1$, there are 179 variables in total: 89 input variables for year t, 89 for year $t-1$, and 1 output variable from year $t-1$. For $Lag = 2$, the model includes 269 variables, comprising 89 input variables each for years t, $t-1$, and $t-2$, plus 2 output values for years $t-1$ and $t-2$. With $Lag = 9$, the model aggregates 899 variables, including 9 past output values and 890 input variables spanning years $t-9$ through t.

The training set includes observations from 2011 to 2020, while the test set comprises data for 2021. The test set consistently contains 51 observations, corresponding to the 51 states to be forecast for 2021. In contrast, the number of observations in the training set varies depending on the lag years considered. For each $Lag = l$, the number of training observations is calculated as $51 \times (10-l)$, covering the period from $2011 + l$ through 2020. Consequently, the number of training observations for each *Lag* from 1 to 9 is 459, 408, 357, 306, 255, 204, 153, 102, and 51, respectively.

4.3 Popular Time Series Prediction Models

The six machine learning models listed below are widely used in healthcare applications for their ability to capture complex patterns in medical data, including disease progression, patient monitoring, and predictive analytics.

a) Support Vector Machine Regression (SVMReg): SVMReg extends Support Vector Machines for regression by finding a function that minimizes prediction errors within a margin ϵ [2]. The model is defined as:

$$f(x) = \langle w, \phi(x) \rangle + b \tag{1}$$

where w is the weight vector, $\phi(x)$ maps input x into a higher-dimensional space, and b is the bias.

b) Binary Decision Tree (BDTree): BDTree recursively partitions the input space to predict continuous values [24]. It minimizes variance by selecting feature j and threshold t_j that optimize:

$$\min_{j,t_j} \sum_{i \in \text{left}} (y_i - \bar{y}_{\text{left}})^2 + \sum_{i \in \text{right}} (y_i - \bar{y}_{\text{right}})^2 \tag{2}$$

where $\bar{y}_{\text{left}}$ and $\bar{y}_{\text{right}}$ are the mean target values of the child nodes.

c) Least-Squares Boosting (LSBoost): LSBoost enhances regression accuracy by iteratively fitting decision trees to residuals [10]. The final prediction is:

$$\hat{y} = \sum_{m=1}^{M} \beta_m h_m(x) \tag{3}$$

where $h_m(x)$ represents the m-th tree and β_m are weights.

d) Neural Network (NN) NNs model complex patterns using interconnected neurons across layers [12]. A multi-layer NN follows:

$$\hat{y} = \phi_L(\phi_{L-1}(\cdots \phi_1(W_1 x + b_1) \cdots) + b_L) \tag{4}$$

where W_l and b_l are weights and biases, and $\phi_l(\cdot)$ is the activation function.

e) Long Short-Term Memory (LSTM): LSTM addresses long-term dependencies in sequential data using memory cells [18]. It updates states via:

$$\begin{aligned} f_t &= \sigma(W_f x_t + U_f h_{t-1} + b_f) \\ i_t &= \sigma(W_i x_t + U_i h_{t-1} + b_i) \\ \tilde{C}_t &= \tanh(W_c x_t + U_c h_{t-1} + b_c) \\ C_t &= f_t \odot C_{t-1} + i_t \odot \tilde{C}_t \\ o_t &= \sigma(W_o x_t + U_o h_{t-1} + b_o) \\ h_t &= o_t \odot \tanh(C_t) \end{aligned} \tag{5}$$

where σ and tanh are activation functions, $\odot$ denotes element-wise multiplication, and W, U, b are model parameters.

f) Bagging Ensemble Regression (ERMBag): ERMBag is a powerful method for improving the stability and accuracy of regression predictions by combining multiple models [31]. The technique begins by creating multiple bootstrap samples from the original dataset $\{(x_i, y_i)\}_{i=1}^{n}$, where each sample is generated through random sampling with replacement. Each of these bootstrap samples is used to train an individual regression tree model, resulting in a collection of B different models.

The predictions from these individual models are then aggregated to produce a final prediction. The most common method of aggregation is by averaging the predictions:

$$\hat{y} = \frac{1}{B} \sum_{b=1}^{B} f_b(x) \tag{6}$$

where $\hat{y}$ is the final prediction for a given input x, and $f_b(x)$ represents the prediction of the b-th model.

4.4 The ERMBag+ Model

Among the six ML models mentioned above, ERMBag is particularly popular for its robustness and ensemble-based approach, which improves predictive accuracy by reducing variance. In the context of disease prediction in time series, ERMBag offers several advantages. By aggregating multiple regression models, it enhances stability and resilience against overfitting, making it well-suited for handling noisy and imbalanced medical datasets. Additionally, ERMBag can effectively capture temporal dependencies in disease progression, improving early detection and long-term forecasting accuracy. Its adaptability to different feature sets and an ability to integrate various weak learners further enhance its predictive power in healthcare applications.

However, ERMBag still has some disadvantages that need improvement. To address these issues, we propose ERMBag+, as described in Algorithm 1, to enhance the model's robustness and predictive accuracy. This is achieved by refining resampling methods, optimizing prediction aggregation, and implementing adaptive strategies tailored to the dataset's characteristics.

Specifically, given the presence of lagged features in the dataset, it is crucial to adjust the block size (B) in the bootstrap process (`Step 1` of the algorithm) to align with the data's time-based dependencies. For datasets with significant lag, increasing the block size helps preserve temporal patterns, ensuring more accurate predictions. Conversely, for datasets with minimal lag, reducing the block size introduces greater variability, improving the ensemble's ability to generalize across different scenarios.

Instead of relying on a simple bootstrap approach, we recommend adopting the Stratified Block Bootstrap technique (`Step 2` of the algorithm). This method ensures that resampled datasets accurately reflect the distribution of specific subgroups-such as those defined by state or time period—thereby reducing potential biases caused by uneven subgroup representations.

Algorithm 1 Our ERMBag+: Ensemble Bagging-Based Regression using Decision tree for Diabetes Prevalence Forecasting

Require:
Time series dataset $D = \{(X_t, y_t)\}$, where X_t represents the feature vector and y_t denotes the target variable (diabetes prevalence).
Number of base learners: M
Block size for bootstrapping: B
Base model: Decision Trees
Lag parameter: $l \in \{1, 2, \ldots, 9\}$
Performance evaluation metric: Root Mean Square Error (RMSE)

Ensure:
Final ensemble prediction $\hat{y}$ for all 51 states in the year 2021.

Step 1: Data Preprocessing
Transform the dataset into a supervised learning format:

$$X_t = [X_t, X_{t-1}, \ldots, X_{t-l}, y_{t-1}, \ldots, y_{t-l}]$$

Define the total number of predictive features: $d = 89 \times (l + 1) + l$
Normalize all input variables.
Step 2: Stratified Block Bootstrap Sampling
Partition the dataset into non-overlapping time blocks of size B.
Ensure stratification by preserving the distribution across demographic and economic variables.
Generate M bootstrapped datasets $D_1, D_2, \ldots, D_M$.
Step 3: Training an Ensemble of Decision Trees
for $i = 1$ to M **do**
    Train a **Decision Tree** model f_i using bootstrapped dataset D_i.
    Apply an early stopping criterion based on RMSE to mitigate overfitting.
    Store the trained model f_i.
end for
Step 4: Adaptive Model Aggregation via Weighted Averaging
for $i = 1$ to M **do**
    Compute individual model predictions: $\hat{y}_i = f_i(X_{\text{test}})$.
    Compute weight for model f_i: $w_i = \frac{1}{\text{RMSE}(f_i)}$
end for
Normalize model weights: $w_i^* = \frac{w_i}{\sum_{j=1}^{M} w_j}$
Compute final ensemble prediction: $\hat{y} = \sum_{i=1}^{M} w_i^* \hat{y}_i$

To mitigate overfitting, particularly in high-dimensional datasets prone to multicollinearity, we integrate an early stopping criterion for each model in the ensemble (`Step 3` of the algorithm). This mechanism monitors validation error rates and halts training once improvements plateau, thereby preserving the model's ability to generalize to new data.

To further improve the accuracy of combined predictions, we propose a weighted aggregation framework, as outlined in `Step 4` of the algorithm. In this approach, each decision tree's contribution is adjusted based on its performance on validation datasets. Specifically, weights are assigned inversely proportional

to each model's Root Mean Square Error (RMSE), ensuring that models with lower errors have a greater impact on the final prediction.

These enhancements address key challenges like data imbalance, overfitting, and temporal dependencies in time-series and high-dimensional datasets. By integrating adaptive resampling, weighted aggregation, and early stopping, the model improves stability, accuracy, and generalization. Additionally, these improvements enhance the model's ability to generalize across different datasets, reducing the risk of bias and improving robustness in real-world healthcare applications. Ultimately, these refinements contribute to a more reliable and interpretable predictive framework, making it well-suited for complex decision-making tasks in dynamic environments.

5 Experiments and Analysis

This section corresponds to Modules 3 and 4 in Figure 2. For ERMBag+ and all six baseline machine learning models, we present and analyze both their predictive performance and computational efficiency.

5.1 Measures

To assess the performance of prediction models, several widely-used metrics include the Mean Absolute Error, $MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$ [30]; Root Mean Squared Error, RMSE $= \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$ [13]; the Mean Absolute Percentage Error, MAPE $= \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|$ [20]; and the coefficient of determination, $R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$ [7]. where y_i is the actual value, $\hat{y}_i$ is the predicted value, and n is the number of observations. Each of these metrics offers a distinct perspective on the model's prediction accuracy.

The algorithms were executed using MATLAB 2022b[8] on a 64-bit Windows 11 system. All experiments were conducted on a computer equipped with an AMD Ryzen 7 5700U CPU with Radeon Graphics and 24GB of RAM.

5.2 Results and Discussion

Figure 3 presents the average prediction performance of six models—SVMReg, BDTree, LSBoost, ERMBag, NN, and LSTM—evaluated across *Lag* values from 1 to 9, using the metrics MAE, RMSE, R^2, and MAPE. The NN model is configured with two layers, each containing 10 hidden neurons, while the LSTM model employs a dropout rate of 0.005 and a maximum of 500 training epochs. The results highlight each model's optimal performance across different *Lag* values, as summarized in Table 2. Specifically, BDTree and LSBoost achieve their best performance at $Lag = 1$, while ERMBag and our enhanced ERMBag+

[8] https://se.mathworks.com/products/new_products/release2022b.html

perform best at $Lag = 2$. Meanwhile, NN, SVMReg and LSTM reach their optimal performance at Lag values of 4, 6, and 9, respectively.

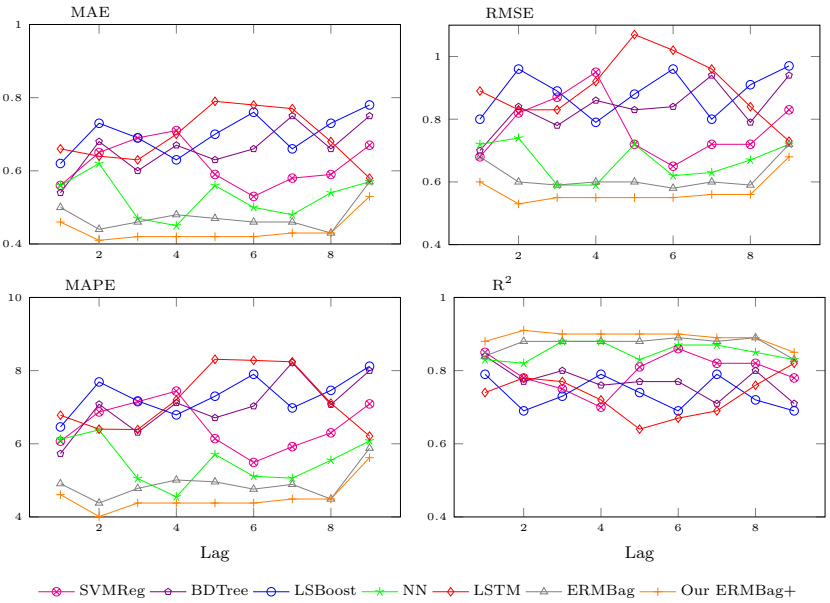

Fig. 3: Average prediction performance of models

The experimental results in Table 2 demonstrate varying performance across models in terms of predictive accuracy and computational efficiency. Our - ERMBag+ achieves the best overall performance at $Lag = 2$, with the lowest error rates (MAE = 0.41, RMSE = 0.53, MAPE = 4.01) and the highest predictive accuracy ($R^2 = 0.91$). However, this requires a substantial computational cost for training.

Among the baseline models, ERMBag ($Lag = 2$) is the second-best performer, achieving MAE = 0.44, RMSE = 0.60, MAPE = 4.38, and $R^2 = 0.88$, while maintaining a more reasonable total execution time of 1.591 seconds. The NN model ($Lag = 4$) also shows strong predictive ability with $R^2 = 0.88$, MAE = 0.45, RMSE = 0.59 and MAPE = 4.55, but its training time (8.182 seconds) is significantly higher than ERMBag.

Other models demonstrate moderate performance with trade-offs between accuracy and efficiency. SVMReg ($Lag = 6$) achieves $R^2 = 0.86$, MAE = 0.53, and RMSE = 0.65, while maintaining a fast total execution time of 0.076 seconds. BDTree and LSBoost (both optimized at $Lag = 1$) perform well for short-term predictions but exhibit relatively higher errors (MAE = 0.54 and 0.62, respectively). LSTM ($Lag = 9$) has an R^2 of 0.82 but requires 11.345 seconds for execution, making it the slowest models.

Table 2: Model Performance Summary

Model	**Lag**	**Prediction Performance**				**Time Performance (second)**		
		MAE	**RMSE**	**MAPE**	**R^2**	Training	Prediction	Total
LSBoost	1	0.62	0.80	6.46	0.79	0.598	0.032	0.63
LSTM	9	0.58	0.73	6.21	0.82	11.309	0.036	11.345
BDTree	1	0.54	0.70	5.73	0.84	0.045	0.011	0.056
SVMReg	6	0.53	0.65	5.49	0.86	0.049	0.027	0.076
NN	4	0.45	0.59	4.55	0.88	8.182	0.013	8.195
ERMBag	2	0.44	0.6	4.38	0.88	1.537	0.054	1.591
ERMBag+	2	0.41	0.53	4.01	0.91	3.235	0.029	3.264

Overall, ERMBag+ achieves the highest predictive performance, surpassing the second-best models (ERMBag and NN) by approximately 3.4% to 13.5% across various evaluation metrics. Additionally, it outperforms the third-best models (SVMReg and BDTree) by 5.8% to 42.9% and exceeds the weakest performing models (LSTM and LSBoost) by 11% to 61.1%.

6 Conclusion and Future Work

We integrated multiple U.S. diabetes-related datasets to construct a comprehensive dataset containing valuable information on diabetes, collected for each state and year from 2011 to 2021. This dataset enables accurate state-level predictions and provides deeper insights into regional trends in diabetes prevalence. Specifically, we proposed EBMBag+, an enhancement of the traditional EBMBag model that incorporates time series techniques for improved diabetes prediction. Our experimental results demonstrated that EBMBag+ achieved the highest predictive performance, outperforming baseline models-LSBoost, LSTM, BDTree, SVMReg, NN, and ERMBag—by 3.4% to 61.1% across the R^2, MAE, RMSE, and MAPE metrics. This approach not only improves prediction accuracy but also offers valuable insights into the evolving dynamics of diabetes at the state level, empowering policymakers to make more informed decisions and implement targeted healthcare strategies.

Future research will see additional features related to diabetes in states including healthcare infrastructure, environmental pollutants, climate and weather, lifestyle factors, and socioeconomic status. In addition, we will apply feature selection techniques to evaluate which features in our dataset have the most significant impact on the performance of our models.

Acknowledgement

This publication has emanated from research conducted with the financial support of Taighde Éireann – Research Ireland under Grant numbers 22/NCF/DR/11244 and 12/RC/2289_P2.

References

1. ADS Authors: Statistics about diabetes. https://diabetes.org/about-diabetes/statistics/about-diabetes (2024), [Accessed 01-October-2024]
2. Awad, M., Khanna, R.: Support Vector Regression, pp. 67–80. Apress, Berkeley, CA (2015)
3. BEA authors: Personal income by state. https://www.bea.gov/data/income-saving/personal-income-by-state (2024), [Accessed 01-October-2024]
4. BLS Authors: Local area unemployment statistics. https://www.bls.gov/lau/tables.htm#mstate (2024), [Accessed 01-October-2024]
5. CDC-Diabetes Authors: National diabetes statistics report. https://www.cdc.gov/diabetes/php/data-research/index.html (2024), [Accessed on 15-May-2024]
6. CDC Population Health Authors: U.S. Chronic Disease Indicators (CDI), 2023 Release. https://data.cdc.gov/Chronic-Disease-Indicators/U-S-Chronic-Disease-Indicators-CDI-2023-Release/g4ie-h725/about_data (2023), [Accessed 01-June-2024]
7. Chicco, D., Warrens, M.J., Jurman, G.: The coefficient of determination r-squared is more informative than smape, mae, mape, mse and rmse in regression analysis evaluation. PeerJ Computer Science **7**(e623) (2021). https://doi.org/10.7717/peerj-cs.623
8. Dharmarathne, G., Jayasinghe, T.N., Bogahawaththa, M., Meddage, D., Rathnayake, U.: A novel machine learning approach for diagnosing diabetes with a self-explainable interface. Healthcare Analytics **5**, 100301 (2024). https://doi.org/https://doi.org/10.1016/j.health.2024.100301
9. El-Sofany, H., El-Seoud, S.A., Karam, O.H., Abd El-Latif, Y.M., Taj-Eddin, I.A.T.F.: A proposed technique using machine learning for the prediction of diabetes disease through a mobile app. International Journal of Intelligent Systems **2024**(1), 6688934 (2024). https://doi.org/https://doi.org/10.1155/2024/6688934
10. Friedman, J.H.: Greedy function approximation: a gradient boosting machine. Annals of Statistics pp. 1189–1232 (2001)
11. Ganie, S.M., Pramanik, P.K.D., Bashir Malik, M., Mallik, S., Qin, H.: An ensemble learning approach for diabetes prediction using boosting techniques. Frontiers in Genetics **14** (2023). https://doi.org/10.3389/fgene.2023.1252159
12. He, L., Madathil, S.C., Servis, G., Khasawneh, M.T.: Neural network-based multitask learning for inpatient flow classification and length of stay prediction. Applied Soft Computing **108**, 107483 (2021)
13. Hodson, T.O.: Root-mean-square error (rmse) or mean absolute error (mae): when to use them or not. Geoscientific Model Development **15**(14), 5481–5487 (2022). https://doi.org/10.5194/gmd-15-5481-2022
14. Iparraguirre-Villanueva, O., Espinola-Linares, K., Flores Castañeda, R.O., Cabanillas-Carbonell, M.: Application of machine learning models for early detection and accurate classification of type 2 diabetes. Diagnostics **13**(14) (2023). https://doi.org/10.3390/diagnostics13142383
15. Khanam, J.J., Foo, S.Y.: A comparison of machine learning algorithms for diabetes prediction. ICT Express **7**(4), 432–439 (2021). https://doi.org/https://doi.org/10.1016/j.icte.2021.02.004
16. Lim, H., Kim, G., Choi, J.H.: Advancing diabetes prediction with a progressive self-transfer learning framework for discrete time series data. Scientific Reports **13**(1), 21044 (2023). https://doi.org/10.1038/s41598-023-48463-0

17. Lugner, M., Rawshani, A., Helleryd, E., Eliasson, B.: Identifying top ten predictors of type 2 diabetes through machine learning analysis of uk biobank data. Scientific Reports **14**(1), 2102 (2024). https://doi.org/10.1038/s41598-024-52023-5
18. Men, L., Ilk, N., Tang, X., Liu, Y.: Multi-disease prediction using lstm recurrent neural networks. Expert Systems with Applications **177**, 114905 (2021)
19. Modak, S.K.S., Jha, V.K.: Diabetes prediction model using machine learning techniques. Multimedia Tools and Applications **83**(13), 38523–38549 (2024). https://doi.org/10.1007/s11042-023-16745-4
20. de Myttenaere, A., Golden, B., Le Grand, B., Rossi, F.: Mean absolute percentage error for regression models. Neurocomputing **192**, 38–48 (jun 2016). https://doi.org/10.1016/j.neucom.2015.12.114
21. National Cancer Institue Authors: Census Tract Population Data Dictionary. https://seer.cancer.gov/censustract-pops/popdictract.html (2024), [Accessed 01-June-2024]
22. Scriney, M., McCarthy, S., McCarren, A., Cappellari, P., , Roantree, M.: Automating data mart construction from semi-structured data sources. The Computer Journal **62**(3), 394–413 (2019). https://doi.org/doi.org/10.1093/comjnl/bxy064
23. Scriney, M., Timilsina, M., Curry, E., Porwol, L., Nie, D., Dahley, D., Fernandez, J.B., D'Aquin, M., Roantree, M.: Engineering data assets for public health applications: A covid-19 case study. In: 2023 IEEE International Conference on Big Data (BigData). pp. 1853–1862 (2023). https://doi.org/10.1109/BigData59044.2023.10386435
24. Silva, M.D.B., de Oliveira, R.D.V.C., da Alves, S.B.D., Melo, E.C.P.: Predicting risk of early discontinuation of exclusive breastfeeding at a brazilian referral hospital for high-risk neonates and infants: A decision-tree analysis. International Breastfeeding Journal **16**(1), 1–13 (2021)
25. Tasin, I., Nabil, T.U., Islam, S., Khan, R.: Diabetes prediction using machine learning and explainable ai techniques. Healthcare Technology Letters **10**(1-2), 1–10 (2023). https://doi.org/https://doi.org/10.1049/htl2.12039
26. U.S. Census: Explore census data 2010 & 2020. https://data.census.gov/table/DECENNIALPL2020.H1?t=Housing:Housing%20Units&d=DEC%20Redistricting%20Data%20(PL%2094-171) (2024), [Accessed on 01-May-2024]
27. U.S. Census Authors: American community survey. https://data.census.gov/advanced (2024), [Accessed 01-October-2024]
28. U.S. Census Authors: State population by characteristics: 2020-2023. https://www.census.gov/data/tables/time-series/demo/popest/2020s-state-detail.html (2024), [Accessed 01-June-2024]
29. WHO Authors: Diabetes overview. https://www.who.int/health-topics/diabetes#tab=tab_1 (2024), [Accessed on Oct 1st, 2024]
30. Willmott, C.J., Matsuura, K.: Advantages of the mean absolute error (mae) over the root mean square error (rmse) in assessing average model performance. Climate Research **30**, 79–82 (2005). https://doi.org/10.3354/cr030079
31. Zhao, C., Peng, R., Wu, D.: Bagging and boosting fine-tuning for ensemble learning. IEEE Transactions on Artificial Intelligence **5**, 1728 (2023)