BMC Medical Research Methodology

RESEARCH

Open Access

Innovative statistical method for longitudinal and hierarchical data modeling: the GMEXGBoost method

Fariba Asadi[1,2], Reza Homayounfar[3,4], Yaser Mehrali[5], Chiara Masci[6] and Farid Zayeri[7*]

Abstract

Introduction and objectives Over recent decades, the exponential growth of data, especially in healthcare, has necessitated advanced analytical methods. Conventional machine learning algorithms often assume independence among data points, limiting their effectiveness with longitudinal and hierarchical data. This study introduces a novel algorithm called GMEXGBoost, a methodological extension of generalized mixed-effects models that leverages the boosting framework of XGBoost for estimating fixed effects while simultaneously accounting for random effects. The innovation lies in GMEXGBoost's ability to explicitly incorporate data correlations while retaining the predictive power of boosted trees.

Methods The GMEXGBoost model was evaluated through extensive simulations and a real-world cohort study, benchmarking against GLMM, GLMMTree, GMERF, and XGBoost. Also, its performance was assessed using predictive mean absolute deviation (PMAD), predictive misclassification rate (PMCR), sensitivity, specificity, accuracy, and AUC. Simulation analyses were conducted using multiple synthetic datasets, each comprising training and testing groups with varying effect structures, including random intercepts and slopes. All computations were performed in RStudio(version 2023.06.0).

Results Our results indicate that while XGBoost achieved the lowest average errors across most scenarios, GMEXGBoost consistently demonstrated superior stability and accuracy when random-effect variance was large or correlations were strong. Also, in real data, GMEXGBoost outperformed other models in terms of the performance metrics.

Conclusion The GMEXGBoost algorithm, by combining the estimates of the GLMM and XGBoost models, leverages the capabilities of both and delivers improved performance in complex problems. Although it is not universally superior, but demonstrates clear advantages in the analysis of hierarchical and longitudinal datasets with strong correlations. These properties make it a valuable tool for decision-making in healthcare and other domains that involve complex, structured data.

Keywords Boosted tree, Generalized linear mixed model, Longitudinal and hierarchical data

*Correspondence:
Farid Zayeri
fzayeri@gmail.com

Full list of author information is available at the end of the article

Introduction

In recent decades, the world has experienced a significant and explosive increase in various types of data. In 2025, the Datasphere is expected to reach 175 ZB, which is more than five times the estimate recorded in 2018 [1]. This phenomenon is known as the "Big Data Era," and it impacts people's daily lives in various ways [2]. A substantial portion of this information explosion has occurred in the fields of medicine and healthcare. The analysis of big data facilitates the examination of large datasets of patients, allowing for the discovery of hidden relationships within the data and the development of predictive models using statistical methods and machine learning (ML) techniques [3–5]. This process leads to the advancement of healthcare services. Parametric statistical methods, such as regression models, are commonly utilized to predict and quantify the relationships between a dependent variable and other associated variables [6]. For analyzing data with non-normal errors and modeling nonlinear relationships, situations that violate the assumptions of most parametric statistical analysis methods [7]—Generalized Linear Models (GLM) [8] and Generalized Additive Models (GAM) [9] have been developed. We have high-dimensional and large datasets, making it challenging to fit parametric models effectively, which may lead to insufficient performance [10, 11].

Consequently, a variety of flexible methods requiring minimal assumptions have recently emerged under the umbrella of "machine learning [3, 10, 12]." Many of these methods have the advantage of being non-parametric, meaning they don't rely on assumptions such as linear relationships or normally distributed residuals. They also permit the inclusion of many predictor variables, which can even exceed the number of observations and enhance predictive accuracy [13]. Statistical approaches begin with an assumed data model, from which parameters are estimated based on the data. In contrast, ML does not start with a predefined model; instead, it employs algorithms to discover the relationships between the response variable and its predictors [11]. ML techniques are suitable for various data types except for longitudinal and hierarchical data. This is because the foundational theory behind most ML algorithms assumes that data points are independent and uniformly distributed [14]. Consequently, when there is a correlation within the data, these methods might underperform. It is evident that classification methods that align with the data structure and effectively manage correlations will enhance prediction accuracy [15].

The current study aims to introduce a new technique based on the decision-tree method called "GMEXGBoost" (Generalized Mixed-Effects eXtreme Gradient Boosting), which draws on insights and techniques of statistical and ML, and is designed to facilitate the analysis of hierarchical and longitudinal data. The main motivation for this work arises from the limitations of existing machine learning algorithms in handling longitudinal and hierarchical data, as well as the limitations of mixed models in managing big data. From a technical standpoint, we aimed to design a model that retains the predictive accuracy and scalability of eXtreme Gradient Boosting (XGBoost) while integrating random-effect structures from Generalized Linear Mixed Model (GLMM), thereby improving performance under correlation. The structure of the paper is as follows: In the remainder of the Introduction, we review and explain the GLMM tree and XGBoost models and their combined model, namely GMEXGBoost, the details framework, including its mathematical formulation and algorithmic implementation. The Structure of the Proposed Method section describes the simulation design and its results. The next section presents the empirical results obtained from the real dataset used to evaluate the method. The discussion section thoroughly examines the findings, highlighting both the strengths and limitations of the approach. In conclusion, the paper suggests directions for future research endeavors.

Contributions of this study

- Integration of GLMM and XGBoost: We propose GMEXGBoost, a hybrid framework that combines generalized linear mixed models with gradient boosting to effectively handle hierarchical and longitudinal data.
- Versatility across response types: GMEXGBoost accommodates various outcomes, including continuous, binary, and count data, within a unified framework.
- Enhanced prediction and scalability: The model maintains high predictive accuracy while remaining scalable to large datasets, outperforming existing hybrid approaches such as Generalized Linear Mixed Model Tree (GLMMTree), Generalized Mixed Effects Random Forest (GMERF), and random-effects XGBoost.

Related works

Recent approaches for biomedical longitudinal data highlight the need for algorithms that can simultaneously capture temporal dependencies and hierarchical clustering. Hu et al. [16] explored the extension of random forest and boosting algorithms for longitudinal data analysis, emphasizing the importance of incorporating random effects into machine learning frameworks. Their study showed that traditional ensemble methods fail to account for intra-subject correlations, which can bias estimates in repeated-measures designs. Their findings

further justify the need for hybrid models that seamlessly integrate mixed-effects estimation and scalable boosting frameworks. Cascarano et al. [17] provided a comprehensive review of machine and deep learning methods for longitudinal biomedical data, emphasizing both the methodological advances and the limitations that remain in integrating ML techniques with mixed-effects modeling. Hu and Szymczak (2023) surveyed extensions of random forest algorithms designed for longitudinal analysis, demonstrating their potential but also noting challenges in interpretability and computational scalability [14]. Recently, Salditt et al. (2023) introduced a gradient tree boosting approach tailored for hierarchical data, demonstrating that boosting frameworks can indeed be adapted to structured datasets. However, their method relies on a gradient boosted tree that is only applicable for continuous response variables [18]. Wang et al. (2024) proposed a novel algorithm, GEEB, which is a gradient boosting framework based on generalized estimating equations (GEE) for correlated or clustered data. Their model effectively handles intra-cluster dependence but remains limited in scalability for very large datasets [19].

Other related works, such as GMERF and GLMMTree, have attempted to combine tree-based methods with mixed-effects modeling [20–22]. While these approaches improve upon classical GLMM in handling non-linearity, they often struggle with efficiency or scalability when applied to large datasets. Compared with these prior efforts, GMEXGBoost offers a unified framework that couples GLMM estimation with XGBoost. This integration allows the model to explicitly account for random effects while preserving the predictive accuracy and scalability of gradient boosting. In this way, GMEXGBoost builds upon existing research while addressing critical gaps related to correlation handling, stability, and computational efficiency. Also, several boosting-based approaches that incorporate random or grouped effects have recently been proposed, such as Gaussian Process Boosting (GPBoost) [23], Model-based Boosting (mboost) [24], and Mixed Effect Gradient Boosting (MEGB) [25]. These methods demonstrate that hybrid frameworks can effectively combine machine learning algorithms with mixed-effects modeling, but most of them are either limited to Gaussian outcomes, computationally demanding, or less straightforward to implement in large-scale clustered datasets. By contrast, GMEXGBoost integrates the efficiency and scalability of XGBoost with the flexibility of generalized mixed-effects models, allowing it to accommodate a variety of responses within a unified framework. Also, Chen et al. (2023) developed a random-effects XGBoost for clustered data, but it only accounts for random intercepts. The proposed GMEXGBoost extends this idea by integrating full GLMM structures to handle complex longitudinal and hierarchical correlations [26].

Structure of the proposed method

Generalized mixed effects XGBoost algorithm

The Generalized Mixed Effects XGBoost algorithm (GMEXGBoost) employs tree-based ensemble methods within the framework of mixed effects models, catering to various classes of response variables in the exponential family. Similar to the GLMM tree, this model mitigates the linear assumptions associated with the fixed effects component of a Generalized Linear Mixed Model (GLMM) by substituting it with an XGBoost algorithm. In the following, we will briefly explain each one. This substitution makes the model more flexible and compatible with different and unknown functional forms. The matrix formula of the GMEXGBoost model is as follows:

$$\begin{array}{ll} g\left(\mu_i\right)=\eta_i=f\left(X_i\right)+Z_i b_i, & \mu_i=\mathrm{E}\left(y_i|b_i\right), \\ b_i \sim \mathrm{N}\left(0, \mathrm{D}\right), & i=1,2,\ldots,p \end{array} \quad (1)$$

Where D is the variance–covariance matrix of the random effects and g(.) is the link function (e.g., logit for binary outcomes, identity for continuous outcomes). In this framework, $f\left(X_i\right)$ represents the prediction function produced by XGBoost, which models the fixed-effects component of the mixed-effects structure. For continuous outcomes, it predicts the response directly. In binary outcomes, $f\left(x_i\right)$ corresponds to the predicted probability of the event via a logistic link. For multiclass categorical outcomes, the vector of class scores is transformed to probabilities using softmax. For count or other non-Gaussian responses, $f\left(x_i\right)$ represents the linear predictor mapped to the mean response through the appropriate link function. This ensures consistency of predictions across different outcome types. The GLMM portion accounts for the random effects, thereby modeling correlations induced by clustering or repeated measurements.

To implement the GMEXGBoost model, it is essential to distinguish between the estimation of fixed effects and random effects. This process involves iteratively refining the model parameters until convergence is achieved. If the random effects are known, the GMEXGBoost model employs an XGBoost to estimate f. However, in numerous practical applications, both the random effects and fixed effects are not known beforehand. It is worth noting that if the random effects were known, the GMEXGBoost model would allow us to fit the XGBoost to estimate f by using $\eta_{ij}-\underline{Z}_{ij}^{T}\underline{b}_i$ as the dependent variable. Conversely, if the population-level function f were known, the random effects could be estimated through a generalized linear mixed-effects model, with the response defined

Table 1 GMEXGBoost algorithm

Initial Input Data Includes:

- Y: Response vector
- Cov: Dataset containing all auxiliary variables
- Group: Vector of grouped variables
- Xnam: Vector of names of variables used as fixed effects
- Znam: Vector of names of variables used as random effects
- Family: Distribution of the response variable
- toll: Threshold value for determining the convergence of estimates
- Itmax: Maximum number of iterations
- b_0: An optional matrix of initial values for each $\underline{b}_i$
- all.b[0] = b

The initial value of b is a zero matrix, where each column is a random coefficient $\underline{b}_i$

1. Fit a GLM model with response *Y* and covariates cov
   - η: The η_{ij} values estimated from the GLM model
   - Set It = 1
2. While convergence has not been achieved, and It < Itmax:
   - Calculate Targ = η − Z* b
   - Fit an XGBoost model with response Targ and the predictor matrix cov
     - *F*(*x*): Values estimated by the XGBoost model
   - Fit a GLMM model in the form $\eta_{ij} - F\left(\underline{x}_{ij}\right) = Z_{ij}^T * \underline{b}_i$
     - all.b[It] = b: Estimated value of *b* from the model
     - M = max(abs(b-all.b [It-1])
     - (i,j) = argmax(abs(b- all.b [It-1])
     - tr = M/all.b [it-1] (i,j)
3. If tr < toll, convergence is correct, then increment it
4. If tr > toll, convergence is incorrect, and exit the loop
5. Repeat steps 3 to 5 until convergence is achieved

Convergence is considered to have occurred when the difference between the estimates of random effects in two consecutive iterations is less than the specified tolerance threshold

$\eta_{ij} - f(\underline{X}_{ij})$. To address this challenge, the GMEXGBoost model incorporates an iterative procedure. This method systematically estimates the random effects in the GLMM while continuously replacing the XGBoost estimate for the fixed effects component until convergence is achieved. In other words, this iterative process continues until the difference between the random effects estimates in consecutive iterations falls below a tolerance threshold set by the researcher.

Convergence is based on the following generalized log-likelihood criterion:

$$\begin{aligned} GLL\left(f, b_i|y\right) = \sum_{i=1}^{n} \{&\widehat{\epsilon}_{i(m)}^T (\widehat{\sigma}_m^2 I_{n_i}) \widehat{\epsilon}_{i(m)} \\ &+ \widehat{b}_{i(m)}^T \widehat{D}_{(m)}^{-1} \widehat{b}_{i(m)} \\ &+ \log \left|\widehat{D}_{(m)}\right| + \log \left|\widehat{\sigma}_m^2 I_{n_i}\right| \end{aligned} \tag{2}$$

This iterative strategy allows the GMEXGBoost model to dynamically estimate random and fixed effects, even when their initial values are unknown. The convergence criterion ensures that the model stabilizes at a point where changes in the random effects estimates are minimal, resulting in robust and dependable outcomes. This flexibility makes the GMEXGBoost model especially useful for longitudinal datasets and hierarchical structures. The pseudo-code of the algorithm's step-by-step structure is detailed in Table 1. Also, (Fig. 1) provides a summary of the GMEXGBoost pipeline, illustrating the entire process from start to finish.

GLMM tree algorithm

A decision tree is one of the famous machine learning algorithms [27]. The focus of the analyst is twofold: to reach a high accuracy in the prediction of a studied phenomenon and to understand the complexity of the underlying data structure [22]. To this purpose, tree-based methods, used for regression and classification, were introduced by Breiman et al. [28]. In recent years, some statistical literature has concentrated on expanding tree-based methods for analyzing nested data, which refers to data with a hierarchical structure, by incorporating them into mixed-effects models [22]. On the other hand, Generalized Linear Mixed Models (GLMMs) are an extension of generalized linear models (GLMs) [8] that allow for the inclusion of both fixed and random effects, making them suitable for analyzing data with hierarchical or hierarchical structures [29]. This flexibility is particularly beneficial in situations where observations are not independent, such as in repeated measures or nested data

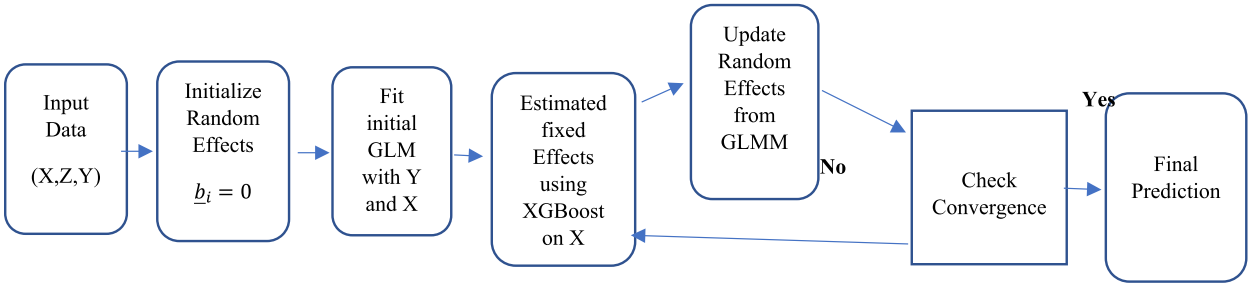

Fig. 1 Schematic diagram of the GMEXGBoost framework, summarizing the start-to-end process of the algorithm

[30]. GLMM trees integrate the flexibility of tree-based algorithms with the hierarchical structure of linear mixed models. This approach facilitates the analysis of complex datasets where observations are nested within groups or collected over time, allowing for the inclusion of both fixed and random effects [13, 21].

XGBoost algorithm

Gradient decision trees are a relatively recent method for forecasting and are a modern topic in data mining that has emerged in the predictive modeling landscape over the past few decades. They find applications across various fields, particularly in medicine and ecology [31]. Boosting is a powerful algorithm in which a collection of weak learners (models with limited predictive power) is combined to create a strong learner (a model with significantly improved performance). The key idea behind boosting is that each tree is grown using information from previously grown trees. By doing so, the overall model can achieve much better performance than any of the individual weak learners on their own [32, 33]. The tree used in this method is based on a CART algorithm, similar to other group methods, which includes a combination of $\widehat{f}^1, \widehat{f}^2, \ldots, \widehat{f}^B$ Trees [34]. XGBoost is a popular gradient-boosting algorithm that utilizes a gradient-based optimization algorithm for training. This algorithm, introduced in 2016 by Chen and Gosterin [35], is known for its excellent performance with large datasets [36]. XGBoost operates over ten times faster than traditional GBM and provides better results than the other popular ML algorithms like artificial neural networks and support vector machines when run on a single machine [37, 38]. The objective function in this algorithm consists of two parts, including 1) a loss function and 2) a regularization parameter, and is defined as follows:

$$Obj\left(\theta\right) = l\left(\theta\right) + \Omega\left(\theta\right) \tag{3}$$

where $\Omega(\theta)$ represents the regularization function, serving as a penalty term to prevent overfitting and manage the complexity of the trees, while $l\left(\theta\right)$ denotes the loss function, which is usually used for regression from the mean squared error (MSE), and for classifying the logarithm as the loss function.

Materials and methods

In this study, we performed a simulation analysis to evaluate the performance of our proposed model, GMEXGBoost (a Generalized Mixed Effects XGBoost), compared to several similar classification techniques. The main goal was to determine GMEXGBoost's ability and flexibility. The simulation experiments allowed us to systematically examine the model's performance under controlled conditions, varying correlation structures and random-effect variances. This provides insights into the robustness and stability of the algorithm.

In addition, we applied GMEXGBoost to a real dataset to demonstrate its practical applicability in real-world scenarios. The real-data analysis enables us to validate the model's predictive performance and comparative advantage over baseline methods.

Simulation framework

To begin, we derive the response variable from a Bernoulli distribution. The setup of our binary data is based on Eq. (4):

$$\begin{array}{l} \eta_{ij} = f\left(X_{ij}\right) + \sum_{q=1}^{Q} b_{iq} z_{ijq} \\ \mu_{ij} = logit^{-1}\left(\eta_{ij}\right) \\ y_{ij} \sim Bernoulli\left(\mu_{ij}\right), \end{array} \tag{4}$$

Where f represents an unspecified functional form of fixed effects, Xij denotes a vector of fixed covariates of dimension P, and $\sum_{q=1}^{Q} b_{iq} z_{ijq}$ accounts for the random effects in the model. We include a moderately high number of covariates, treating f as a linear component estimated by XGBoost. In other words, we design f to include both a linear part and a tree-like part, as well as interactions among covariates. This approach allows us to simulate a complex structure that will effectively test the proposed method's flexibility and adaptability.

Specifically, we set P = 7 and define f as follows:

$$\begin{aligned} f\left(X_1, \ldots, X_7\right) = & \alpha\left(X_1^2 - 3X_2 - X_2 X_3^2\right) \\ & + \beta tree\left(X_4, X_5, X_6\right) \end{aligned} \tag{5}$$

In this formulation, α and β are parameters that adjust the variability of f; tree(×4, ×5, ×6) denotes a function characterized by a tree-structured representation. The final variable, X7, is intentionally insignificant, serving to evaluate if it misleads the algorithm. The seven covariates are generated randomly, adhering to these distributions:

$$\begin{array}{l} X_1, X_2 \sim U\left(-1,1\right); X_3 \sim Weibull\left(3\right); X_4 \sim U\left(-3,3\right); \\ X_5 \sim U\left(-6,6\right); X_6 \sim U\left(-5,5\right); X_7 \sim U\left(-4,4\right) \end{array} \tag{6}$$

For the random effects component, we simulate $N = 10$ groups, each containing $ni = 40$ observations, leading to a total of 400 units. These observations are drawn from a normal distribution in line with GLMM assumptions. We explore two distinct specifications for the random effects:

Random Intercept Only:

$$\sum_{q=1}^{Q} b_{iq} z_{ijq} = b_{i0} \sim N\left(0, \gamma^2\right) \tag{7}$$

Equation (7) indicates a single scalar random effect, where γ influences the variability of this random effect.

Random Intercept and Slope:

$$\sum_{q=1}^{Q} b_{iq} z_{ijq} = b_{i0} + b_{i1} x_{ij1} \sim N\left(0, \gamma^2\right) \tag{8}$$

With x_{ij1} as the first fixed effects covariate, and the random coefficient $\underline{b} \sim N_2(0, \Sigma)$

where $\Sigma \sim diag\left(\gamma^2, \delta^2\right)$.

Notably, b_{i0} and b_{i1} are independent across all groups; δ acts as another variance-regulating parameter. Here, the effect of x is allowed to vary across groups, supporting more tailored modeling. To predict a new observation, we compute $\widehat{\eta}_{ij} = \widehat{f}(X_{ij}) + Z^T{}_{ij}\widehat{b}_i$. where $Z^T{}_{ij}\widehat{b}_i$ captures the contribution of the group-specific random effects, and $\widehat{f}(X_{ij})$ corresponds to the fixed-effects component estimated through XGBoost. The predicted mean response $\widehat{\mu}_{ij}$ is then derived by applying the inverse link function. For new observations not included in the training data, the random effect estimates $(\widehat{b}_i)$ are set to zero, and predictions rely solely on the fixed-effects component,$f(x_i)$. for predicting future observations within clusters already present in the training set, GMEXGBoost can leverage both fixed and random effects, improving prediction accuracy for these clusters.

The parameters introduced govern the variability of the simulated data; we selectively choose their values to ensure that the probability μ_{ij} For each unit, it does not approach extreme values of 0 or 1 (except for some observations). We execute a total of eight different simulation scenarios, adjusting each coefficient's value to create instances of both low and high variance for the respective model components; To evaluate the effectiveness of GMEXGBoost and the ability of the model to predict new individuals, we generated a corresponding test dataset for each training dataset, consisting of 50 observations per group (for a total of 500 observations), and compared its outcomes against those of GLM, GLMM, GLMM tree, XGBoost, and GMERF (Generalized Mixed Effects Random Forest). All models assessed based on their predictive performance using two metrics: predictive mean absolute deviation (PMAD) and predictive misclassification rate (PMCR), which are defined as follows:

$$\mathbf{PMAD} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n_i}\left(\boldsymbol{\mu_{ij}} - \widehat{\boldsymbol{\mu}}_{\boldsymbol{ij}}\right)}{\boldsymbol{N_{test}}} \tag{9}$$

$$\mathbf{PMCR} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n_i}\left(\boldsymbol{y_{ij}} - \widehat{\boldsymbol{y}}_{\boldsymbol{ij}}\right)}{\boldsymbol{N_{test}}} \tag{10}$$

where $N_{test} = 500$ and $n_i = 50$ for each i from 1 to 10; $\widehat{\mu}_{ij}$ represents the predicted probability from the model, y_{ij} is the true value of the response, and $\widehat{y}_{ij}$ signifies the response estimated by the model. When predicting a new observation, we compute $\widehat{\eta}_{ij} = \widehat{f}(X_{ij}) + Z^T{}_{ij}\widehat{b}_i$. where $Z^T{}_{ij}\widehat{b}_i$ captures the contribution of the group-specific random effects, and $\widehat{f}(X_{ij})$ corresponds to the fixed-effects component estimated through XGBoost. The predicted mean response $\widehat{\mu}_{ij}$ is then derived by applying the inverse link function. In this study, the variance of PMAD and PMCR was reported not for inferential purposes, but to evaluate the stability and robustness of predictive performance across repeated simulations. To enhance comparability with the article by Pellagatti et al. [22], all simulation parameters for both fixed and random effects were selected accordingly. Hyperparameter tuning was performed using grid search combined with ten-fold cross-validation, and the same tuning strategy was applied uniformly across all models to ensure a fair comparison. Results were compared across models, with significance set at $p < 0.05$ using RStudio (version 2023.06.0). To train the models, we used a computer equipped with an Intel Core i5-12500H processor, an NVIDIA GeForce GTX 1650 GPU with 4 GB VRAM, and 16 GB of RAM.

The models considered in this study include: GLMER, which estimates generalized linear mixed models and is available in the lme4 package [39]; Random Forest (RF), implemented via the randomForest package [40]; XGBoost using the xgboost package [41]; and the ordinary tree was provided by the rpart package [42], and the GLMM tree was fitted to the data using the glmertree package [43]. Also, for the implementation of GMERF, we used the original code provided by the authors of the method [22]. For GMEXGBoost, the random effects b were initialized as a zero matrix, consistent with related hybrid approaches [21, 22]. We also examined initialization via preliminary GLMM estimates, which produced nearly identical results, confirming that the algorithm is not overly sensitive to starting values. Convergence was monitored through the maximum relative change in random-effects estimates, with tolerance set to toll = 0.02 and maximum iterations to itmax = 30 following Pellagatti et al. [22]. Sensitivity checks using alternative thresholds (e.g., toll = 0.5, itmax = 500) confirmed the robustness of the stopping rule. The iterative procedure was coordinated through custom R functions, and all R scripts used in this study are available at the following GitHub repository: https://github.com/faas34188/GMEXGBoost/tree/main

The real dataset

This study used data from the baseline phase of the Fasa Cohort Study (FACS), which examines health conditions

and risk factors for non-communicable diseases (NCDs) among rural residents in Fasa, Iran. The cohort included 10,146 participants aged 35 to 70 from Sheshdeh and Qarabalag areas, which encompass 29 villages. Demographic, medical, nutrition, and lifestyle data were collected using standard questionnaires. Written consent was obtained from all participants. Further details about the FACS can be found in reference [44]. Data from 9,499 participants were analyzed after excluding those with significant missing data. This study focused on 29 villages as clusters and examined cardiovascular disease (CVD), defined as heart failure (HF) or ischemic heart disease (IHD). Potential predictors included demographic factors, sleep patterns, body mass index (BMI), dietary inflammatory index (DII), biochemical markers, waist-to-height ratio, comorbidities, lifestyle factors, and family history of CVD. The Missing values (< 5%) were imputed using single imputation (mean imputation for numerical features and mode imputation for categorical features), categorical variables were dummy-coded, and the Z-normalization was applied to the numerical variables. For the complex models (XGBoost, RF, GMEXGBoost), hyperparameter tuning was carried out via five-fold cross-validation on the training data, with the optimal parameters selected based on validation AUC. The optimized hyperparameters were set as follows: eta = 0.01, nround = 50, max-depth = 3, and subsample = 0.5. To assess model performance using four metrics: sensitivity, specificity, accuracy, and AUC (Area Under the Curve), we randomly partitioned the dataset, allocating 80% to the training set and 20% to the test set. This division was implemented using a stratified approach, ensuring that the distribution of the outcome variable was consistently represented in the training and testing datasets.

Results

Simulation results

To evaluate the performance of the models, lower values of the metrics—specifically the mean PMAD, mean PMCR, and their variances—indicate a better model. These lower values reflect higher stability and greater accuracy of the models. Overall, based on the results presented in Tables 2 and 3, in terms of PMAD and PMCR across various scenarios, while standard XGBoost frequently achieved the lowest mean error, GMEXGBoost outperformed it in high-correlation scenarios, highlighting its robustness in hierarchical data structures. As illustrated in (Figs. 2 and 3), when the variance of random effects increases, GMEXGBoost consistently

Table 2 Predictive performance of six algorithms in random intercept models across four simulation scenarios with varying variances of fixed and random effects

Model	Variance of fixed effects	Variance of random effects	Algoritm	Mean of PMAD	Variance of PMAD	Mean of PMCR	Variance of PMCR
Random Intercept	Small	Small	GLM	0.1600	0.00019	0.2894	0.00058
Random Intercept	Small	Small	GLMM	0.1622	0.00018	0.2974	0.00045
Random Intercept	Small	Small	GLMMTree	0.1678	0.00048	0.2982	0.00114
Random Intercept	Small	Small	GMERF	0.1625	0.00011	0.2771	0.00033
Random Intercept	Small	Small	XGBoost	0.1532	0.00014	0.2756	0.00028
Random Intercept	Small	Small	GMEXGBoost	0.1614	0.00016	0.2908	0.00067
Random Intercept	Large	Small	GLM	0.1475	0.00015	0.2276	0.00039
Random Intercept	Large	Small	GLMM	0.1452	0.00016	0.2285	0.00033
Random Intercept	Large	Small	GLMMTree	0.1770	0.00030	0.2512	0.00130
Random Intercept	Large	Small	GMERF	0.1669	0.00024	0.2316	0.00054
Random Intercept	Large	Small	XGBoost	0.1425	0.00018	0.2144	0.00060
Random Intercept	Large	Small	GMEXGBoost	0.1551	0.00014	0.2282	0.00054
Random Intercept	Small	Large	GLM	0.1738	0.00053	0.2962	0.00049
Random Intercept	Small	Large	GLMM	0.1720	0.00053	0.2932	0.00042
Random Intercept	Small	Large	GLMMTree	0.1660	0.00113	0.2876	0.00126
Random Intercept	Small	Large	GMERF	0.1775	0.00122	0.2686	0.00236
Random Intercept	Small	Large	XGBoost	0.1672	0.00015	0.2698	0.00058
Random Intercept	Small	Large	GMEXGBoost	0.1656	0.00010	0.2726	0.00031
Random Intercept	Large	Large	GLM	0.1663	0.00082	0.2272	0.00157
Random Intercept	Large	Large	GLMM	0.1585	0.00078	0.2298	0.00158
Random Intercept	Large	Large	GLMMTREE	0.1550	0.00019	0.2186	0.00029
Random Intercept	Large	Large	GMERF	0.1695	0.00043	0.2221	0.00158
Random Intercept	Large	Large	XGBoost	0.1606	0.00040	0.2112	0.00084
Random Intercept	Large	Large	GMEXGBoost	0.1601	0.00011	0.2220	0.00081

Table 3 Predictive performance of six algorithms in random intercept and slope models across four simulation scenarios with varying variances of fixed and random effects

Model	Variance of fixed effects	Variance of random effects	Algoritm	Mean of PMAD	Variance of PMAD	Mean of PMCR	Variance of PMCR
Random Intercept and Slope	Small	Small	GLM	0.1724	0.00024	0.3132	0.00063
Random Intercept and Slope	Small	Small	GLMM	0.1788	0.00022	0.3112	0.00063
Random Intercept and Slope	Small	Small	GLMMTree	0.1845	0.00070	0.3146	0.00126
Random Intercept and Slope	Small	Small	GMERF	0.1793	0.00029	0.3038	0.00097
Random Intercept and Slope	Small	Small	XGBoost	0.1579	0.00008	0.2872	0.00079
Random Intercept and Slope	Small	Small	GMEXGBoost	0.1783	0.00018	0.3092	0.00054
Random Intercept and Slope	Large	Small	GLM	0.1685	0.00028	0.2392	0.00048
Random Intercept and Slope	Large	Small	GLMM	0.1700	0.00032	0.2362	0.00090
Random Intercept and Slope	Large	Small	GLMMTree	0.1775	0.00052	0.2506	0.00178
Random Intercept and Slope	Large	Small	GMERF	0.1733	0.00020	0.2470	0.00054
Random Intercept and Slope	Large	Small	XGBoost	0.1455	0.00015	0.2170	0.00056
Random Intercept and Slope	Large	Small	GMEXGBoost	0.1584	0.00022	0.2264	0.00057
Random Intercept and Slope	Small	Large	GLM	0.1802	0.00074	0.2722	0.00174
Random Intercept and Slope	Small	Large	GLMM	0.1807	0.00088	0.2706	0.00146
Random Intercept and Slope	Small	Large	GLMMTree	0.1816	0.00097	0.2824	0.00150
Random Intercept and Slope	Small	Large	GMERF	0.1805	0.00049	0.2934	0.00147
Random Intercept and Slope	Small	Large	XGBoost	0.1695	0.00083	0.2676	0.00095
Random Intercept and Slope	Small	Large	GMEXGBoost	0.1687	0.00068	0.2836	0.00076
Random Intercept and Slope	Large	Large	GLM	0.1688	0.00069	0.2216	0.00186
Random Intercept and Slope	Large	Large	GLMM	0.1689	0.00155	0.2160	0.00267
Random Intercept and Slope	Large	Large	GLMMTree	0.1710	0.00074	0.2190	0.00206
Random Intercept and Slope	Large	Large	GMERF	0.1769	0.00052	0.2534	0.00063
Random Intercept and Slope	Large	Large	XGBoost	0.1674	0.00030	0.2062	0.00084
Random Intercept and Slope	Large	Large	GMEXGBoost	0.1671	0.00089	0.2094	0.00044

exhibits the lowest variance, which reflects its superior efficiency and performance. All error and variance estimates for GMEXGBoost fall between those of XGBoost and GLMM. Specifically, the error and variance values of GMEXGBoost are slightly higher than those of XGBoost but lower than those of GLMM. Additionally, the results reveal that both GMEXGBoost and GMERF outperform the GLMMTree model, demonstrating improved predictive accuracy and stability.

Real data results

We conducted a comprehensive comparison of the execution time and the predictive performance of Tree, Random Forest (RF), XGBoost, GLMM, GLMMTree, GMERF, and GMEXGBoost models, following a rigorous preprocessing pipeline. A comparison of simple machine learning models in Table 4 revealed that all models performed adequately in predicting cardiovascular disease (CVD). Notably, the XGBoost and Random Forest (RF) models demonstrated superior performance compared to the standard decision tree. The RF model achieved an accuracy of 74.73%, which was one percentage point higher than the 73.63% accuracy of XGBoost, while XGBoost ssforms RF in terms of AUC by approximately 7%. Also, the results indicated that the GMEXGBoost model achieved a sensitivity of 78% and an accuracy of 77%, outperforming the other two models in terms of the performance metrics. With respect to execution time, the analysis showed that XGBoost was approximately twice as fast as RF. Among the hybrid models, GMEXGBoost demonstrated the shortest runtime. In particular, it outperformed the GMERF by 195.57 s and the Mixed Tree model by 387.35 s, highlighting its computational efficiency. while village-level clustering was modeled as a random effect, variables of age, LDL cholesterol, family history of cardiovascular disease in first-degree relatives, physical activity level, and the presence of hypertension were recognized as the most influential fixed effects.

Several boosting-based approaches that incorporate random or grouped effects have recently been proposed, such as GPBoost [23], mboost [24], and MEGB [25]. These methods demonstrate that hybrid frameworks can effectively combine machine learning algorithms with mixed-effects modeling, but most of them are either limited to Gaussian outcomes, computationally demanding, or less straightforward to implement in large-scale clustered datasets. By contrast, GMEXGBoost integrates the efficiency and scalability of XGBoost with the flexibility of generalized mixed-effects models, allowing it

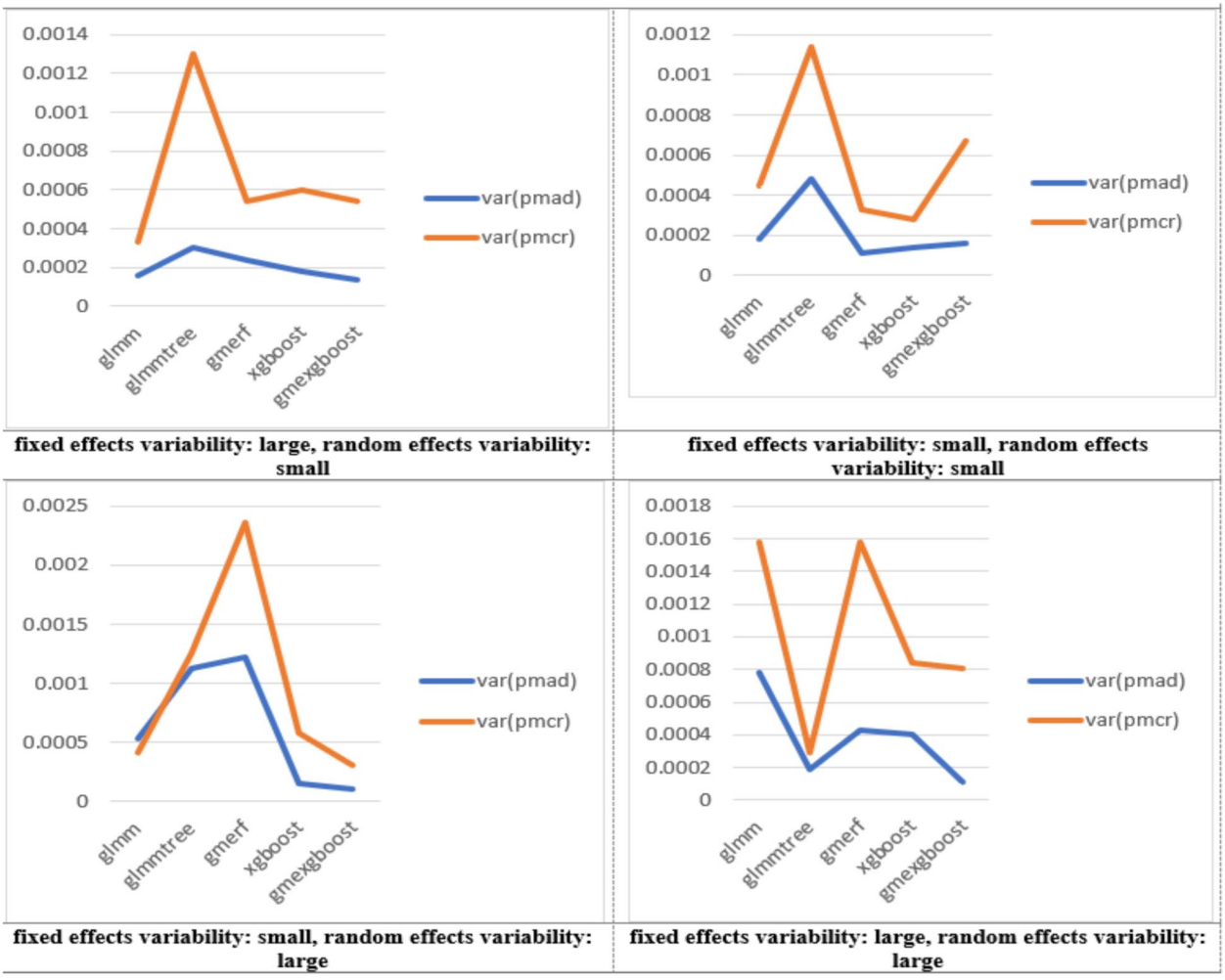

Fig. 2 Predictive variance performance of different algorithms in random intercept model with varying variances

to accommodate a variety of responses within a unified framework.

Discussion

Our current research aimed to introduce a new model for correlated data, named GMEXGBoost, and to compare it with existing methods. The boosted mixed tree model is an advanced machine learning model based on decision trees. Machine learning methods, due to their non-parametric nature, do not rely on assumptions such as linear relationships or normally distributed residuals. Furthermore, these methods allow the inclusion of a wide range of potential predictor variables, even when the number of variables exceeds the number of observations [16, 19].

In this study, the results of the model evaluations in simulation based on two key error metrics, PMAD and PMCR, demonstrate that the two models, XGBoost and GMEXGBoost, performed the best across all conditions, providing the lowest error values for predictions. This highlights the significance of using tree-based boosting algorithms in tackling complex prediction problems and identifying non-linear patterns.

The results demonstrated that the XGBoost model not only achieved the best performance in reducing the mean errors of PMAD and PMCR but also provided the most stable results and minimized the respective variances. These features indicate that XGBoost can accurately and reliably evaluate complex and large-scale data, which is of great importance in practical applications where decision-making based on dependable and low-variance predictions is critical.

Conversely, the GMEXGBoost model ranked second and, compared to XGBoost, exhibited slightly higher fluctuations in the results. While GMEXGBoost does not universally outperform XGBoost in terms of raw predictive error, its strength lies in scenarios where correlation between observations is substantial and cannot be ignored. In such cases, GMEXGBoost provides more reliable and interpretable results, particularly by appropriately modeling both fixed and random effects.

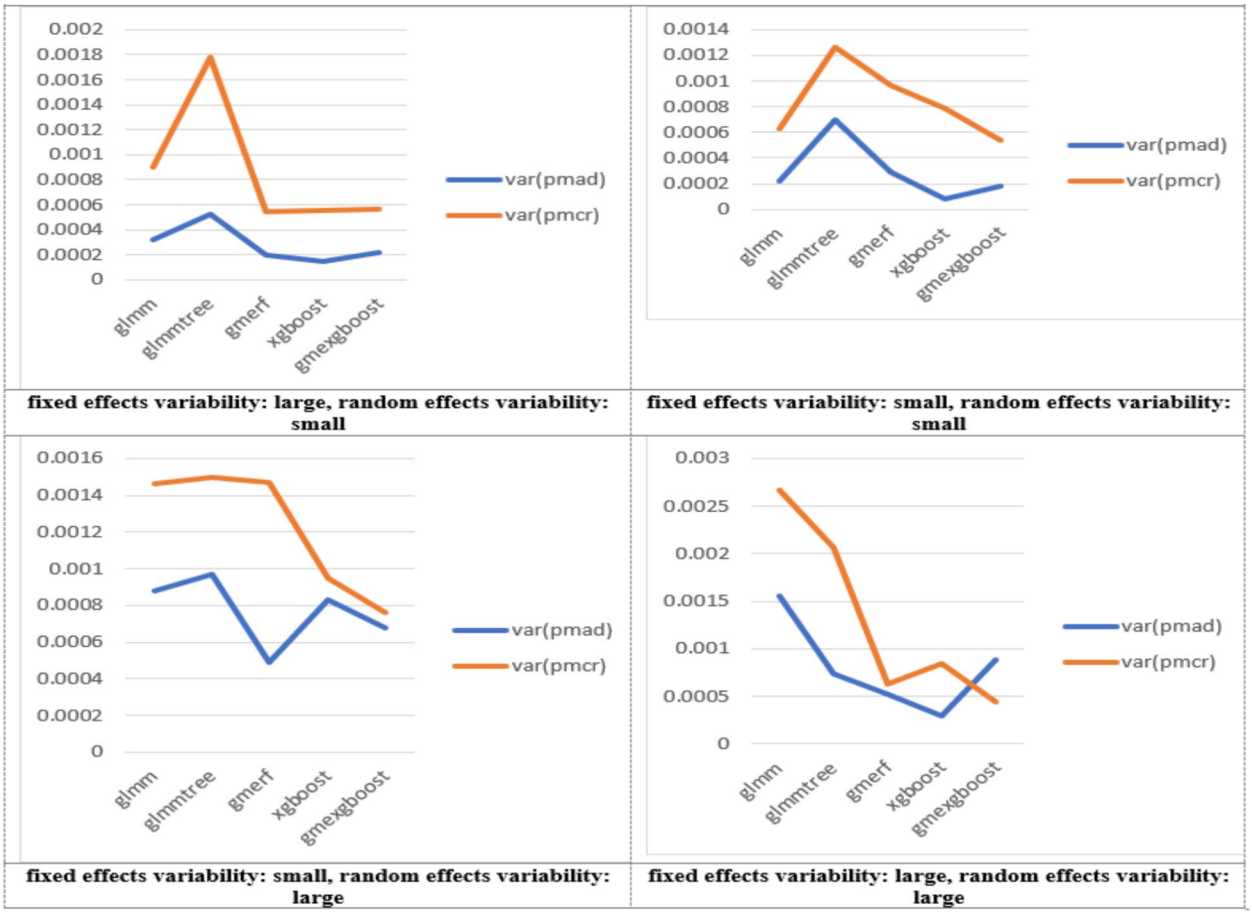

Fig. 3 Predictive variance performance of different algorithms in random intercept and slope model with varying variances

Table 4 Comparison of predictive power indices of machine learning models using fivefold cross-validation

Model	Sensitivity	Specificity	Accuracy	AUC*	Run-time**
Tree	0.83	0.63	0.81 (0.79,0.83)	0.57 (0.54,0.59)	1.27
RF	0.76	0.70	0.76 (0.74,0.77)	0.73 (0.70,0.76)	4.05
XGBoost	0.74	0.72	0.74 (0.71,0.75)	0.80 (0.76,0.84)	1.97
GLMM	0.75	0.72	0.74 (0.72,0.76)	0.79 (0.75,0.82)	68.62
GLMM tree	0.75	0.73	0.74 (0.72,0.76)	0.80 (0.77,0.84)	396.63
GMERF	0.77	0.73	0.76 (0.74,0.78)	0.80 (0.77,0.84)	204.85
GMEXG-Boost	0.78	0.72	0.77 (0.75,0.79)	0.80 (0.77,0.84)	9.28

*Area Under the Curve

**second

In settings with low correlation, XGBoost may appear to perform better due to its direct focus on minimizing prediction error, but this comes at the expense of neglecting within-cluster dependence. Thus, the two models should be viewed as complementary. Additionally, when compared to the GLMM model, GMEXGBoost outperformed it and showed lower variance. This phenomenon could be attributed to the iterative nature of the GMEXGBoost algorithm, which helps stabilize the estimates. On the other hand, the traditional structure and limitations of the GLMM model in capturing complex and nonlinear structures in large datasets might explain this outcome. Although the improvements in predictive performance over GLMM may appear numerically modest, in medical applications even small gains can be clinically meaningful.

In real data, we conducted a comprehensive comparison of the predictive performance metrics of several machine learning models, including Tree, Random Forest (RF), XGBoost, Generalized Linear Mixed Models (GLMM), GLMM Tree, GMERF, and GMEXGBoost. Our analysis revealed that all models performed adequately in predicting cardiovascular disease (CVD), underscoring the utility of these algorithms in medical diagnostics.

Notably, the XGBoost and Random Forest models performed better than standard decision trees. As ensemble methods, XGBoost and RF demonstrated greater

predictive power, paralleling findings from previous studies [45, 46]. Results showed that the Random Forest model attained an accuracy of 74.73%, exceeding the accuracy of XGBoost by one percentage point. This finding is consistent with results reported by Kabiraj et al. (2020) on breast cancer data, who found similar accuracy favoring RF's predictive performance over simpler models [47]. However, in our study, XGBoost outperformed RF in terms of AUC and execution time, indicating that while RF may achieve slightly higher accuracy, XGBoost provides advantages in terms of model efficiency and speed.

Previous research has shown that XGBoost generally outperforms other machine learning algorithms in various healthcare applications. For instance, in work by Zamani et al. (2021), XGBoost exhibited a minimal execution time of 19 s while demonstrating high accuracy and AUC when compared to RF and deep learning models using multi-source remote sensing data [48]. This highlights XGBoost's potential in medical and clinical applications, where both accuracy and computational efficiency are paramount.

In our results, the GMEXGBoost model achieved a sensitivity of 78% and an accuracy of 77%, outperforming the other models assessed. This high sensitivity is particularly significant in clinical settings, where accurately identifying patients at risk of CVD can directly impact treatment decisions and outcomes. These results align with those of Dehghan et al. (2020), who noted that XGBoost achieved superior performance in terms of preventing future CVD events, thereby emphasizing its effectiveness in identifying critical cases [49].

The analysis of the execution time of each model revealed that the GMEXGBoost model demonstrated the shortest execution duration, completing its tasks significantly faster than the other models. These results are consistent with the findings of Chen and Guestrin (2016), who noted that XGBoost is designed for speed and performance in large datasets, making it particularly suitable for time-sensitive medical applications [35].

The primary focus of the present study was on the methodological contribution, namely the development and evaluation of the GMEXGBoost model, and demonstrating its superior predictive performance and computational efficiency compared to existing approaches. While mixed-effects models inherently allow for the interpretation of fixed and random effects, a comprehensive investigation of the underlying risk factors and their clinical implications was beyond the scope of this work. We deliberately restricted our analysis to performance evaluation to ensure a rigorous and fair methodological comparison. In line with Pellagatti et al. [22], who concentrated on extending GMET into GMERF rather than benchmarking all possible alternatives, our aim was not to exhaustively evaluate other hybrid strategies, but to improve upon GMERF by integrating XGBoost into the mixed-effects framework. A detailed exploration of the influential fixed and random effects identified by the proposed model will be addressed in a separate study, where the focus will shift from methodological development to epidemiological and clinical. However, in our previous work [50] on the Fasa Cohort Study, we have already provided a detailed interpretation of the fixed and random effects identified by some hybrid models. Several boosting-based approaches that incorporate random or grouped effects have recently been proposed, such as GPBoost [23], mboost [24], and MEGB [25]. These methods demonstrate that hybrid frameworks can effectively combine machine learning algorithms with mixed-effects modeling, but most of them are either limited to Gaussian outcomes, computationally demanding, or less straightforward to implement in large-scale clustered datasets. By contrast, GMEXGBoost integrates the efficiency and scalability of XGBoost with the flexibility of generalized mixed-effects models, allowing it to accommodate a variety of responses within a unified framework.

Conclusion

The GMEXGBoost algorithm, by combining the estimates of the GLMM and XGBoost models, leverages the capabilities of both and delivers improved performance in complex problems. Results showed that GMEXGBoost achieves higher sensitivity and stability than GLMM-based methods, and it performs better than XGBoost in highly correlated settings. Compared with related approaches such as GMERF and GLMMTree, GMEXGBoost provides a more scalable and computationally efficient solution. The combination of high accuracy and rapid computation makes the GMEXGBoost model a compelling choice for applications requiring reliable predictions and quick processing capabilities. These findings suggest that GMEXGBoost is particularly valuable for applications like personalized medicine, where data are clustered and reliable predictions are essential. This study has several strengths, including methodological innovation, rigorous simulations across diverse scenarios, and application to a large biomedical dataset. Nevertheless, the proposed GMEXGBoost model has certain limitations. First, it requires careful parameter tuning, which may pose challenges for researchers without a strong statistical background, potentially hindering its immediate adoption in fields such as medical research, where ease of use is critical. Second, we did not investigate simulation scenarios with a large number of irrelevant (junk) variables, which would be valuable to further assess robustness under high-dimensional noise. A more systematic evaluation of this aspect will be pursued in future work.

Third, in the real-data analysis, performance was evaluated using random individual-level splits, which may allow partial cluster overlap. More rigorous cluster-aware validation strategies will be considered in future work to address this issue. Fourth, we did not include calibration metrics or cluster-level summaries of predictive performance, as our primary aim was a methodological comparison. In this context, improvements in execution time and predictive metrics relative to existing models were the main focus, while calibration and clinical trade-offs can be addressed in future work. Future research can extend GMEXGBoost to multiclass and survival outcomes and evaluate its robustness in high-dimensional data. We also plan to expand its implementation as a user-friendly statistical software package to facilitate adoption in healthcare and social sciences. In addition, we plan to implement similar open-source models, such as GPBoost, and directly compare their performance with GMEXGBoost to further benchmark its effectiveness.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12874-025-02751-7.

Supplementary Material 1.

Acknowledgements
The authors express their gratitude towards all individuals who patiently contributed to this study.

Authors' contributions
F.A.: Supervision, methodology, investigation, conceptualization, validation, formal analysis, writing—original draft preparation. F.Z.: Supervision, conceptualization, validation, reviewing, and editing. Y.M. and CH.M.: Methodology, conceptualization. R.H.: Data curation, reviewing, and editing.

Funding
The authors declare that no funding was received for this research.

Data availability
The datasets generated and/or analysed during the current study are not publicly available due to privacy and security considerations, but are available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate
This study was conducted as part of a PhD dissertation in Biostatistics approved by the Ethics Committee of Shahid Beheshti University of Medical Sciences (approval number: IR.SBMU.RETECH.REC.1402.137). The study was conducted in accordance with the Declaration of Helsinki. The data were derived from a previously approved human cohort study, and ethical approval and informed consent were obtained in the original study.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Author details
[1]Ferdows Faculty of Medical Sciences, Birjand University of Medical Sciences, Birjand, Iran
[2]Department of Biostatistics, School of Allied Medical Sciences, Shahid Beheshti University of Medical Sciences, Tehran, Iran
[3]Noncommunicable Diseases Research Center, Fasa University of Medical Sciences, Fasa, Iran
[4]Food Technology Research Institute, Faculty of Nutrition Sciences and Food Technology, Shahid Beheshti University of Medical Sciences, Tehran, Iran
[5]Statistical Center of Iran, Tehran, Iran
[6]Department of Economics, Management and Quantitative Methods, University of Milan, Milan, Italy
[7]Proteomics Research Center and Department of Biostatistics, School of Allied Medical Sciences, Shahid Beheshti University of Medical Sciences, Qods Square, Darband Street, Tehran, Iran

Received: 14 July 2025 / Accepted: 17 December 2025
Published online: 08 January 2026

References

1. Reinsel D, Gantz J, Rydning J. The digitization of the world from edge to core, vol. 16. Framingham: International Data Corporation; 2018. p. 1–28.
2. BashariRad B, Akbarzadeh N, Ataei P, Khakbiz Y. Security and privacy challenges in big data era. Int J Control Theory Appl. 2016;9(43):437–48.
3. Gupta MK, Chandra P. A comprehensive survey of data mining. Int J Inf Technol. 2020;12(4):1243–57.
4. Razzak MI, Imran M, Xu G. Big data analytics for preventive medicine. Neural Comput Appl. 2020;32(9):4417–51. https://doi.org/10.1007/s00521-019-04095-y.
5. Xiaoling S, Yiwan Y. Knowledge discovery: methods from data mining and machine learning. Soc Sci Res. 2023;110(102817):1–16. https://doi.org/10.1016/j.ssresearch.2022.102817.
6. Young DS. Handbook of Regression Methods. 1st ed. Chapman and Hall/CRC; 2017.
7. Lee DK. Data transformation: a focus on the interpretation. Korean J Anesthesiol. 2020;73(6):503–8. https://doi.org/10.4097/kja.20137.
8. McCullagh P, Nelder J. Generalized linear models. 2nd ed. Boca Raton: Chapman and Hall/CRC; 2019.
9. Wood S. Generalized additive models: an introduction with R. 2nd ed. Boca Raton, FL: CRC Press; 2017.
10. Hastie T, Tibshirani R, Friedman JH. The elements of statistical learning: data mining, inference, and prediction. Statistics SSi, editor. New York: Springer; 2009.
11. Elith J, Leathwick JR, Hastie T. A working guide to boosted regression trees. J Anim Ecol. 2008;77(4):802–13.
12. Jovel J, Greiner R. An introduction to machine learning approaches for biomedical research. Front Med. 2021;8:771607. https://doi.org/10.3389/fmed.2021.771607.
13. Fokkema M, Edbrooke-Childs J, Wolpert M. Generalized linear mixed-model (GLMM) trees: a flexible decision-tree method for multilevel and longitudinal data. Psychother Res. 2021;31(3):329–41. https://doi.org/10.1080/10503307.2020.1785037. https://www.tandfonline.com/doi/full/10.1080/10503307.2020.1785037.
14. Jianchang H, Silke S. A review on longitudinal data analysis with random forest. Briefings in Bioinformatics. 2023;24(2). https://doi.org/10.1093/bib/bbad00210.
15. Mangino AA, Finch WH. Prediction with mixed effects models: a Monte Carlo simulation study. Educ Psychol Meas. 2021;81(6):1118–42.
16. Hu S, Wang Y-G, Drovandi C, Cao T. Predictions of machine learning with mixed-effects in analyzing longitudinal data under model misspecification. Stat Methods Appl. 2023;32(2):681–711. https://doi.org/10.1007/s10260-022-00658-x.
17. Cascarano A, Mur-Petit J, Hernandez-Gonzalez J, Camacho M, de Toro Eadie N, Gkontra P, et al. Machine and deep learning for longitudinal biomedical data: a review of methods and applications. Artif Intell Rev. 2023;56(Suppl 2):1711–71.
18. Salditt M, Humberg S, Nestler S. Gradient tree boosting for hierarchical data. Multivar Behav Res. 2023;58(5):911–37.
19. Wang Y-W, Yang H-C, Chen Y-H, Guo C-Y. Generalized estimating equations boosting (GEEB) machine for correlated data. J Big Data. 2024;11(1):20. https://doi.org/10.1186/s40537-023-00875-5.

20. Cascarano A, Mur-Petit J, Hernández-González J, Camacho M, de Toro EN, Gkontra P, et al. Machine and deep learning for longitudinal biomedical data: a review of methods and applications. Artif Intell Rev. 2023;56(2):1711–71. https://doi.org/10.1007/s10462-023-10561-w10.1007/s10462-023-10561-w.
21. Hajjem A, Larocque D, Bellavance F. Generalized mixed effects regression trees. Stat Probability Lett. 2017;126:114–8. https://doi.org/10.1016/j.spl.2017.02.033. https://www.sciencedirect.com/science/article/abs/pii/S0167715217300895?via%3Dihub.
22. Pellagatti M, Masci C, Ieva F, Paganoni AM. Generalized mixed-effects random forest: a flexible approach to predict university student dropout. Stat Anal Data Min ASA Data Sci J. 2021;14(3):241–57. https://doi.org/10.1002/sam.11505.
23. Sigrist F, Gyger T, Kuendig P. gpboost: combining tree-boosting with Gaussian process and mixed effects models. CRAN: Contributed Packages. 2021
24. Hofner B, Mayr A, Robinzonov N, Schmid M. Model-based boosting in R: a hands-on tutorial using the R package mboost. Comput Stat. 2014;29(1):3–35.
25. Olaniran OR, Olaniran SF, Allohibi J, Alharbi AA, Alharbi NM. Mixed effect gradient boosting for high-dimensional longitudinal data. Sci Rep. 2025;15(1):30927.
26. Dong T, Oronti IB, Sinha S, Freitas A, Zhai B, Chan J, et al. Enhancing cardiovascular risk prediction: development of an advanced xgboost model with hospital-level random effects. Bioengineering. 2024;11(10):1039. https://www.mdpi.com/2306-5354/11/10/1039.
27. Song YY, Lu Y. Decision tree methods: applications for classification and prediction. Shanghai Arch Psychiatry. 2015;27(2):130–5. https://doi.org/10.11919/j.issn.1002-0829.215044.
28. Breiman L, Friedman J, Stone CJ, Olshen RA. Classification and regression trees. Cambridge, MA: CRC Press; 1984.
29. Fallahzadeh H, Asadi F. Generalized linear mixed models: Introduction, Estimation method and Application in medical studies. Paramed Sci Military Health. 2019;14(1):33–8. http://jps.ajaums.ac.ir/article-1-176-fa.html.
30. Bolker BM, Brooks ME, Clark CJ, Geange SW, Poulsen JR, Stevens MHH, et al. Generalized linear mixed models: a practical guide for ecology and evolution. Trends Ecol Evol. 2009;24(3):127–35. https://doi.org/10.1016/j.tree.2008.10.008.
31. Carty DM. An analysis of boosted regression trees to predict the strength properties of wood composites. Master's Thesis, The University of Tennessee; 2011. https://trace.tennessee.edu/utk_gradthes/954.
32. Zhao C, Wu D, Huang J, Yuan Y, Zhang H-T, Peng R, et al. BoostTree and BoostForest for ensemble learning. IEEE Trans Pattern Anal Mach Intell. 2022;45(7):8110–26.
33. Bahad P, Saxena P. Study of AdaBoost and Gradient Boosting Algorithms for Predictive Analytics. International Conference on Intelligent Computing and Smart Communication 2019. Algorithms for Intelligent Systems. Springer: Singapore; 2020:235-44. https://doi.org/10.1007/978-981-15-0633-8_22.
34. Jiang J, Cui B, Zhang C. Distributed machine learning and gradient optimization. Singapore: Springer; 2022.
35. Chen T, Guestrin C. Xgboost: a scalable tree boosting system. Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. New York; 2016. p. 785–94.
36. Wang C, Deng C, Wang S. Imbalance-XGBoost: leveraging weighted and focal losses for binary label-imbalanced classification with XGBoost. Pattern Recogn Lett. 2020;136:190–7.
37. Dhaliwal SS, Nahid A-A, Abbas R. Effective intrusion detection system using XGBoost. Information. 2018;9(7):149. https://doi.org/10.3390/info9070149.
38. Ghafarian F, Wieland R, Lüttschwager D, Nendel C. Application of extreme gradient boosting and Shapley Additive explanations to predict temperature regimes inside forests from standard open-field meteorological data. Environ Model Softw. 2022;156:105466.
39. Bates D, Mächler M, Bolker B, Walker S. Fitting linear mixed-effects models using lme4. J Stat Softw. 2015;67(1):1–48. https://doi.org/10.18637/jss.v067.i01.
40. Liaw A, Wiener M. Classification and regression by randomForest. R News. 2002;2(3):18–22.
41. Chen T, He T, Benesty M, Khotilovich V, Tang Y, Cho H, et al. Xgboost: extreme gradient boosting. R package version 04–2. 2015;1(4):1–4
42. Therneau T, Atkinson B, Ripley B, Ripley MB. Package 'rpart'. 2015;2:5–32. Available online: cranmaicacuk/web/packages/rpart/rpart.pdf. Accessed on 20 Apr 2016.
43. Fokkema M, Zeileis A, Fokkema MM. Package 'glmertree'. 2019. Retrived from: https://cran.r-project.org/web/packages/glmertree/glmertree.pdf.
44. Farjam M, Bahrami H, Bahramali E, Jamshidi J, Askari A, Zakeri H, et al. A cohort study protocol to analyze the predisposing factors to common chronic non-communicable diseases in rural areas: Fasa Cohort Study. BMC Public Health. 2016;16(1):1–8. https://doi.org/10.1186/s12889-016-3760-z. https://bmcpublichealth.biomedcentral.com/articles/10.1186/s12889-016-3760-z.
45. Dong X, Zhiwen Y, Wenming C, Yifan S, Qianli M. A survey on ensemble learning. Front Comp Sci. 2020;14:241–58.
46. Mehra A, Tripathy P, Faridi A, Chinmay A. Ensemble learning approach to improve existing models. Int J Innov Sci Res Technol. 2019;4(12):25–9.
47. Kabiraj S, Raihan M, Alvi N, Afrin M, Akter L, Sohagi SA, et al. Breast cancer risk prediction using XGBoost and random forest algorithm. 11th international conference on computing, communication and networking technologies (ICCCNT). IEEE; Kharagpur, India, 2020, pp. 1–4. https://doi.org/10.1109/ICCCNT49239.2020.9225451.
48. Zamani Joharestani M, Cao C, Ni X, Bashir B, Talebiesfandarani S. PM2. 5 prediction based on random forest, XGBoost, and deep learning using multisource remote sensing data. Atmosphere. 2019;10(7):373.
49. Nghiem N, Wilson N, Krebs J, Tran T. Predicting the risk of diabetes complications using machine learning and social administrative data in a country with ethnic inequities in health: Aotearoa New Zealand. BMC Med Inform Decis Mak. 2024;24(1):274. https://doi.org/10.1186/s12911-024-02678-x.
50. Asadi F, Homayounfar R, Mehrali Y, Masci C, Talebi S, Zayeri F. Detection of cardiovascular disease cases using advanced tree-based machine learning algorithms. Sci Rep. 2024;14(1):22230. https://doi.org/10.1038/s41598-024-72819-9.

Publisher's Note