



OPEN

Mixed effect gradient boosting for high-dimensional longitudinal data

Oyebayo Ridwan Olaniran^{1,4,5}✉, Saidat Fehintola Olaniran^{2,5}, Jeza Allohibi^{3,5}, Abdulmajeed Atiah Alharbi^{3,5} & Nada MohammedSaeed Alharbi^{3,5}

High-dimensional longitudinal data present significant analytical challenges due to intricate within-subject correlations and an overwhelming ratio of predictors to observations. To address these challenges, we introduce Mixed-Effect Gradient Boosting (MEGB), a novel R package that synergises gradient boosting with mixed-effects modelling to simultaneously account for population-level fixed effects and subject-specific random variability. MEGB provides a unified framework for analysing repeated measures data that accommodates complex covariance structures while harnessing gradient boosting's inherent regularization for robust feature selection and prediction. In comprehensive simulations spanning linear and nonlinear data-generating processes, MEGB achieved 35–76% lower mean squared error (MSE) compared to state-of-the-art alternatives like Mixed-Effect Random Forests (MERF) and REEMForest, while maintaining 55–70% true positive rates for variable selection in ultra-high-dimensional regimes ($p = 2000$). Demonstrating practical utility, we applied MEGB to maternal cell-free plasma RNA data ($n = 12$ subjects, $p = 33,297$ transcripts), where it identified 9 key placental transcripts driving fetal RNA dynamics across pregnancy trimesters.

Keywords Mixed Effect Model, Longitudinal Data, Gradient Boosting, High-dimensional Data

The statistical analysis of high-dimensional longitudinal data presents formidable challenges, primarily due to the dual complexity of managing intricate within-subject correlation patterns and addressing the “curse of dimensionality,” where the number of predictors (p) vastly exceeds the sample size (n)¹. Longitudinal studies, which involve repeated measurements of subjects over time, inherently exhibit temporal dependencies and individual-specific variability. Traditional approaches such as linear mixed-effects models (LMMs) have been widely adopted to handle these dependencies by partitioning variance into fixed effects (population-level trends) and random effects (subject-specific deviations)². Extensions like glmmlasso³ integrate L_1 -penalized regression (lasso) with LMMs to enable variable selection in high-dimensional settings, simultaneously estimating fixed effects and covariance structures while shrinking coefficients of noninformative predictors to zero. However, while glmmlasso improves upon classical LMMs by performing regularization, it remains constrained by the limitations of its underlying mixed-effects framework. Specifically, in ultrahigh-dimensional regimes ($p \gg n$), the method suffers from computational instability, overreliance on restrictive parametric assumptions (e.g., linearity and Gaussian random effects), and diminished power to distinguish true signals from noise due to the nonconvexity of the penalized likelihood objective^{3,4}. This issue is particularly acute in biomedical research, where high-throughput technologies such as genomics, proteomics, and metabolomics generate datasets with thousands of longitudinally tracked molecular characteristics between individuals⁵. For example, in longitudinal transcriptomic studies, glmmlasso struggles to model nonlinear gene expression trajectories or interactions while scaling to datasets with predictors $p > 100$. These challenges underscore the need for advanced methodologies that balance interpretability, computational efficiency, and predictive accuracy while accommodating both high-dimensionality and longitudinal structure without relying on restrictive parametric forms.

In recent years, ensemble machine learning methods, particularly Gradient Boosting Machines (GBMs), have emerged as powerful alternatives for high-dimensional data analysis. Introduced by Friedman⁶, gradient boosting operates by iteratively constructing an ensemble of weak learners (e.g., decision trees) that minimize a differentiable loss function. This approach excels in high-dimensional contexts due to its inherent regularization, adaptability to nonlinear relationships, and robust variable selection capabilities^{7,8}. Theoretical advances, such

¹Department of Statistics, Faculty of Physical Sciences, University of Ilorin, Ilorin, Kwara State PMB 1515, Nigeria.

²Department of Statistics and Mathematical Sciences, Faculty of Pure and Applied Sciences, Kwara State University, Malete, Kwara State PMB 1530, Nigeria. ³Department of Mathematics, Taibah University, Faculty of Science, Al-Munawara 42353, Saudi Arabia. ⁴Department of Biostatistics & Health Informatics, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, United Kingdom. ⁵These authors contributed equally to this work: Oyebayo Ridwan Olaniran, Saidat Fehintola Olaniran, Jeza Allohibi, Abdulmajeed Atiah Alharbi and Nada MohammedSaeed Alharbi. ✉email: olaniran.or@unilorin.edu.ng; ridwan.olaniran@kcl.ac.uk

as the consistency of boosting algorithms⁹ and Bayesian extensions that incorporate sparsity-inducing priors¹⁰, have further solidified its theoretical foundation. Despite these strengths, conventional GBMs are designed for cross-sectional data and fail to account for within-subject correlations in longitudinal studies, limiting their ability to leverage the rich temporal structure of repeated measurements.

To bridge this gap, researchers have proposed adaptations of tree-based models adapted for longitudinal and clustered data. Early efforts of Segal¹¹ introduced multivariate regression trees that accommodate correlated responses, allowing basic handling of repeated measures. Subsequent innovations, such as the integration of polynomial mixed effects models within tree nodes by Eo and Cho¹², improved the ability to model non-linear temporal trajectories. Wei et al.^{13,14} further advanced this paradigm by combining mixed-effects models with regression splines, using likelihood ratio tests during node splitting to improve model flexibility. While these methods represent progress, their reliance on stepwise splitting criteria and parametric assumptions limits scalability in high-dimensional settings, where computational efficiency and nonparametric adaptability are paramount.

Semi-parametric approaches have gained traction as a flexible middle ground between fully parametric and nonparametric models. Hajjem et al.^{15,16} pioneered tree-based semi-parametric mixed-effects models, where regression trees or Random Forests estimate nonparametric components while parametric terms capture random effects. Their Expectation-Maximization (EM) algorithm iteratively updates fixed and random effects, balancing flexibility with structure. Similarly, Sela and Simonoff¹⁷ developed mixed-effects regression trees, and Fu and Simonoff¹⁸ employed conditional inference trees for clustered data. Despite these innovations, many methods oversimplify correlation structures or struggle with high-dimensional data. Recent work by Capitaine et al.¹⁹ addressed these limitations through Random Forest adaptations like the Mixed-Effect Random Forest (MERF) and REEMForest, which incorporate stochastic serial correlation effects via variants such as SMERF and SREEMForest. However, these frameworks remain computationally intensive and lack the gradient boosting framework's variable selection efficiency.

Parallel advancements in boosting algorithms have expanded their utility in machine learning. For instance, Bayesian additive regression trees¹⁰ integrate sparsity-inducing priors to enhance performance in high-dimensional cross-sectional data, while Zhu et al.²⁰ incorporated reinforcement learning to optimize tree construction. Recent work by Sigrist^{21,22} introduced GPBoost, a method combining gradient boosting with Gaussian process or mixed-effects models to handle correlated data, such as longitudinal or spatial datasets. GPBoost leverages tree-based ensembles for fixed effects and covariance functions for random effects, offering improved predictive accuracy in settings with structured dependencies. Despite these developments, a critical gap persists: few methods explicitly integrate gradient boosting with mixed-effects modelling to address high-dimensional longitudinal data while balancing flexibility and scalability. This shortfall is particularly evident in biomedical applications, such as longitudinal genomic studies tracking cell-free RNA during pregnancy, where models must simultaneously handle thousands of predictors, nonlinear interactions, and within-subject variability^{23–25}.

To address these limitations, we introduce MEGB (Mixed-Effect Gradient Boosting), an R package designed for high-dimensional longitudinal data analysis. MEGB synergizes the predictive power of gradient boosting with the rigour of mixed-effects modelling, enabling robust analysis of repeated measures in scenarios where $p \gg n$. Key innovations include:

1. **High-Dimensional Scalability:** MEGB efficiently handles datasets with thousands of predictors, making it ideal for omics research (e.g., genomics, proteomics).
2. **Within-Subject Correlation Modeling:** By integrating random effects into the boosting framework, MEGB captures individual-specific trajectories and temporal dependencies, outperforming conventional GBMs and Random Forests.
3. **Nonlinear Interaction Capture:** The algorithm accommodates complex predictor-response relationships, which are crucial for modelling biological processes.
4. **Variable Selection:** MEGB's iterative fitting process prioritizes relevant predictors, reducing noise from redundant features.

The remainder of this paper is structured as follows: First, we detail MEGB's methodology, including fixed- and random-effect estimation. Next, we present simulation studies evaluating its performance under varying data conditions, followed by a practical guide to implementing MEGB using the R package. We then apply MEGB to a real-world dataset involving longitudinal cell-free maternal-fetal RNA analysis, demonstrating its utility in biomedical research. Finally, we discuss results, limitations, and future directions for advancing high-dimensional longitudinal data analysis.

Mixed effect gradient boosting

Mixed Effect Gradient Boosting (MEGB) is a hybrid statistical and machine learning technique that integrates the strengths of gradient boosting with mixed-effects modelling, addressing the unique challenges of longitudinal or hierarchical data. This framework is particularly suitable for data with repeated measurements or nested structures, where fixed and random effects play crucial roles. Fixed effects represent population-level trends, while random effects capture subject-specific deviations. By combining these elements, MEGB provides a robust method for modelling complex dependencies within data, as Laird & Ware²⁶ emphasized in their foundational work on mixed-effects models.

The MEGB model for a continuous response variable Y_{ij} is formulated as:

$$Y_{ij} = f(X_{ij}) + Z_{ij}\mathbf{b}_i + \epsilon_{ij}, \quad (1)$$

where $i = 1, \dots, n$ indexes subjects, and $j = 1, \dots, n_i$ indexes repeated measurements (e.g., time points) for the i -th subject. $Y_{ij} \in \mathbb{R}$ is the continuous observed outcome for subject i at measurement j . $X_{ij} \in \mathbb{R}^p$ and $Z_{ij} \in \mathbb{R}^q$ are time-varying (or time-invariant) predictors for fixed and random effects, respectively. The term $f(X_{ij})$ denotes the nonlinear fixed-effects function, modelled via gradient boosting to capture complex interactions and nonlinear relationships. The subject-specific random effects $\mathbf{b}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{B})$ follow a multivariate normal distribution with covariance matrix \mathbf{B} . The residual error term $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$ is assumed to be independent of \mathbf{b}_i . Gradient boosting⁶ iteratively constructs $f(X_{ij})$ by fitting weak learners (e.g., decision trees) to residuals, enabling MEGB to model nonlinear fixed effects without assuming a parametric form. Unlike linear mixed models, $f(X_{ij})$ flexibly adapts to interactions (e.g., gene-environment) and nonlinear trends (e.g., time-varying biomarker trajectories). The random effects term $Z_{ij}\mathbf{b}_i$ accounts for within-subject correlations, where Z_{ij} typically includes time-varying covariates (e.g., measurement time) or subject-level confounders. The residual error term $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$ accounts for unexplained variance. Together, these components form a hierarchical model with the covariance structure:

$$\mathbf{V}_i = \mathbf{Z}_i \mathbf{B} \mathbf{Z}_i^\top + \sigma^2 \mathbf{I}, \quad (2)$$

where \mathbf{Z}_i is the design matrix for random effects. This structure ensures that the MEGB algorithm incorporates both within-subject and between-subject variability, making it ideal for scenarios where traditional gradient boosting might fail to account for hierarchical dependencies²⁷.

The iterative procedure in MEGB alternates between estimating the fixed effects function $f(X_{ij})$ using gradient boosting and updating random effects and variance components through the Expectation-Maximization (EM) algorithm. This integration enables MEGB to efficiently balance the dual objectives of prediction and inference, critical for longitudinal data analysis. Here, prediction refers to the model's ability to forecast results (e.g. future biomarker levels) for new subjects or time points by using fixed effects at the population level ($f(\mathbf{X}_{ij})$) and subject-specific random effects ($Z_{ij}\mathbf{b}_i$). Gradient boosting drives predictive accuracy by flexibly modelling nonlinear relationships and interactions among fixed-effect predictors (e.g., gene-environment dynamics), even in high-dimensional settings. Inference, on the contrary, encompasses the model's ability to (1) identify biologically meaningful predictors through stable variable selection (e.g. transcripts with high importance scores across cross-validation replicates), (2) quantify fixed effects at the population level (e.g. effect size and direction of a gene on the result), and (3) estimate variance components (\mathbf{B}, σ^2) that characterize variability within and between subjects. Unlike "black-box" machine learning methods, MEGB retains interpretability through its mixed-effects structure, allowing researchers to distinguish global trends (fixed effects) from individual deviations (random effects) and assess their statistical significance. By unifying the predictive power of gradient boosting with the rigour of mixed effects, MEGB avoids the trade-off between precision and interpretability: boosting captures complex fixed-effect patterns, while the EM algorithm ensures reliable inference in both population parameters and subject-specific trajectories. This dual capability is particularly vital in biomedical applications, where both forecasting patient outcomes and understanding biological mechanisms are paramount. This blend of flexibility and structure makes MEGB a valuable tool in diverse applications, from biomedical research to social sciences, where longitudinal or nested data structures are common²⁸.

MEGB mitigates overfitting through three integrated mechanisms: (1) Gradient boosting regularization via shrinkage (step size $\eta = 0.05$) and tree depth constraints (max depth = 3-5), limiting incremental updates and model complexity; (2) EM-driven estimation of random effects, which borrows strength across subjects by shrinking subject-specific estimates $\hat{\mathbf{b}}_i$ toward zero via the shared covariance \mathbf{B} ; and (3) Early stopping during boosting iterations determined by out-of-sample validation loss (10-fold cross-validation). For small samples ($n < 30$), we further constrain random effects by imposing diagonal \mathbf{B} structures and increasing regularization via reduced tree depths (max depth = 2). These mechanisms collectively prevent over-parameterization while maintaining subject-specific flexibility.

While both MEGB and GPBoost^{21,22} integrate gradient boosting with structured modelling for correlated data, their methodological frameworks diverge critically. GPBoost couples tree-based fixed effects with Gaussian processes (GPs) or parametric mixed-effects models, using kernel-based covariances to capture spatial/temporal dependencies. MEGB employs a parsimonious mixed-effects framework, combining gradient boosting with explicit subject-specific random effects and EM-estimated variance components. This structure avoids GPBoost's theoretical $O(n^3)$ kernel inversion, replacing it with linear-time updates ($\mathbf{B} = \frac{1}{n} \sum \mathbf{b}_i \mathbf{b}_i^\top$) that scale efficiently to large n asymptotically. However, as would be observed later in the results section, GPBoost's low-rank approximations and optimized implementations often yield faster practical runtimes, even in high-dimensional settings. Furthermore, MEGB introduces sparsity-inducing regularization for both fixed and random effects, enabling feature selection in ultra-high-dimensional regimes ($p \gg n$), while GPBoost prioritizes covariance flexibility over sparsity. GPBoost's kernel-based approach excels in modelling nonparametric spatial/smooth temporal correlations, whereas MEGB's parametric random-effects structure (\mathbf{b}_i) may struggle with highly nonstationary dependencies. Conversely, MEGB inherently captures non-linear fixed-effect interactions via gradient boosting, avoiding explicit kernel design. Thus, MEGB's computational advantages lie primarily in scalable EM updates and regularization for high-dimensional settings, rather than raw speed. In practice, MEGB is better suited for high-dimensional longitudinal data (e.g., large- p biomedical datasets with hierarchy), while GPBoost excels for both low- and high-dimensional spatial data with stationary covariances. Both trade flexibility and scalability, but MEGB's EM-driven framework addresses challenges in feature selection and ultra-high-dimensional inference.

Estimation of fixed and random effects

The Mixed Effect Gradient Boosting (MEGB) algorithm combines gradient boosting for fixed effects estimation with an Expectation-Maximization (EM) framework to refine random effects and variance components iteratively. In the **initialization step**, random effects (\mathbf{b}_i) are set to zero, and variance components (σ^2 and \mathbf{B}) are initialized. Here, \mathbf{b}_i captures the subject-specific deviations, while σ^2 models residual variance, and \mathbf{B} represents the covariance of random effects. These components form the basis for mixed models, as described in foundational works by Laird & Ware²⁶. This initialization ensures a neutral starting point for the iterative procedure, aligning with the principles of EM algorithms²⁹.

In the iterative estimation step, the algorithm alternates between estimating fixed and random effects using the EM principles. First, a pseudo-response (\mathbf{Y}_{ij}^*) is computed by adjusting the observed response (Y_{ij}) for the current random effects estimate:

$$Y_{ij}^* = Y_{ij} - \mathbf{Z}_{ij} \hat{\mathbf{b}}_i. \quad (3)$$

A gradient boosting model is then fitted to \mathbf{Y}_{ij}^* to estimate the fixed effects function $f(\mathbf{X}_{ij})$. The estimation procedure for Gaussian responses aims to iteratively improve predictions by adding new trees that minimize the residual sum of squares (RSS). At iteration m , the model updates the prediction $\hat{f}_{m-1}(\mathbf{X}_{ij})$ by adding a new tree $h_m(\mathbf{X}_{ij})$:

$$\hat{f}_m(\mathbf{X}_{ij}) = \hat{f}_{m-1}(\mathbf{X}_{ij}) + \eta \cdot h_m(\mathbf{X}_{ij}), \quad (4)$$

where η is the learning rate. The loss function for Gaussian responses is defined as:

$$L(\hat{Y}_{ij}, Y_{ij}) = \sum_{i=1}^n \sum_{j=1}^{n_i} (Y_{ij} - \hat{Y}_{ij})^2, \quad (5)$$

where Y_{ij} is the true response and $\hat{Y}_{ij} = \hat{f}_m(\mathbf{X}_{ij})$ is the predicted response for subject i at measurement j . The gradient of this loss with respect to $\hat{f}_m(\mathbf{X}_{ij})$ gives the negative residuals:

$$g_{ij}^{(m)} = -\frac{\partial L}{\partial \hat{f}_m(\mathbf{X}_{ij})} = Y_{ij} - \hat{f}_m(\mathbf{X}_{ij}), \quad (6)$$

which are used to fit the next tree. The tree $h_m(\mathbf{X}_{ij})$ is trained to predict $g_{ij}^{(m)}$, solving:

$$\hat{h}_m(\mathbf{X}_{ij}) = \arg \min_{h_m} \sum_{i=1}^n \sum_{j=1}^{n_i} \left(g_{ij}^{(m)} - h_m(\mathbf{X}_{ij}) \right)^2. \quad (7)$$

The fitted tree $\hat{h}_m(\mathbf{X}_{ij})$ is then scaled by a step size η , and the prediction for each subject-measurement pair is updated as:

$$\hat{f}_m(\mathbf{X}_{ij}) = \hat{f}_{m-1}(\mathbf{X}_{ij}) + \eta \cdot \hat{h}_m(\mathbf{X}_{ij}), \quad (8)$$

where \mathbf{X}_{ij} represents the predictor vector for subject i at measurement j , and $\hat{f}_m(\mathbf{X}_{ij})$ is the cumulative prediction after m iterations. This update rule ensures that the gradient boosting component adapts to both cross-sectional trends (via \mathbf{X}_{ij}) and temporal dependencies (via repeated j) inherent in longitudinal data. The η learning rate is typically chosen via cross-validation to balance underfitting and overfitting. This method of boosting with Gaussian loss has been shown to work effectively in various regression tasks, with the gradient boosting algorithm being widely applied for its efficiency and predictive power^{6,7}. Once the fixed effect component $\hat{f}_m(\mathbf{X}_{ij})$ has been estimated, the next step involves updating the random effects using the Best Linear Unbiased Prediction (BLUP) formula:

$$\hat{\mathbf{b}}_i = \mathbf{B} \mathbf{Z}_i^\top \left(\mathbf{Z}_i \mathbf{B} \mathbf{Z}_i^\top + \sigma^2 \mathbf{I} \right)^{-1} (\mathbf{Y}_i - f(\mathbf{X}_i)). \quad (9)$$

This step minimizes the joint prediction error, with $\mathbf{V}_i = \mathbf{Z}_i \mathbf{B} \mathbf{Z}_i^\top + \sigma^2 \mathbf{I}$ serving as the covariance matrix. Maximum likelihood estimates of σ^2 and \mathbf{B} are derived by solving marginal likelihood equations, ensuring that variance components are updated efficiently in line with methods described by Pinheiro & Bates²⁷. The convergence is monitored using the log-likelihood of the model:

$$\log \mathcal{L} = -\frac{1}{2} \sum_i \left[\log |\mathbf{V}_i| + (\mathbf{Y}_i - f(\mathbf{X}_i))^\top \mathbf{V}_i^{-1} (\mathbf{Y}_i - f(\mathbf{X}_i)) \right]. \quad (10)$$

Iterations stop when the relative improvement in log-likelihood falls below a predefined threshold (δ). This ensures computational efficiency while maintaining model accuracy. The algorithm outputs the final gradient boosting model ($f(\mathbf{X})$), estimates of random effects ($\hat{\mathbf{b}}_i$), and variance components (σ^2 and \mathbf{B}). This hybrid approach effectively bridges the gap between machine learning techniques and classical mixed-effects modelling,

offering robust solutions for hierarchical or clustered data²⁸. After convergence, predictions for subject i at measurement j integrate fixed and random effects:

$$\hat{Y}_{ij} = \hat{f}(\mathbf{X}_{ij}) + \mathbf{Z}_{ij}\hat{\mathbf{b}}_i. \quad (11)$$

This combines population-level trends ($\hat{f}(\mathbf{X}_{ij})$) and subject-specific deviations ($\mathbf{Z}_{ij}\hat{\mathbf{b}}_i$), capturing both global patterns and individual variability¹⁹.

Require: Predictor matrix X , response vector Y , subject IDs id , random effects design matrix Z , maximum iterations $iter$, number of trees $ntree$, convergence threshold δ , time points $time$.

Ensure: Fixed effects function $f(X)$, random effects \mathbf{b}_i , variance components σ^2 , \mathbf{B} , log-likelihood sequence, and OOB errors.

1: Initialize random effects $\mathbf{b}_i \leftarrow \mathbf{0}$, residual variance $\sigma^2 \leftarrow 1$, and random effects covariance $\mathbf{B} \leftarrow \mathbf{I}$.

2: **for** $t = 1$ to $iter$ **do**

3: Compute pseudo-response $Y_{ij}^* \leftarrow Y_{ij} - Z_{ij}\hat{\mathbf{b}}_i$.

4: Fit gradient boosting model $f(X)$ using Y^* as the response:

$$f(X) = \operatorname{argmin}_{ij} \sum (Y_{ij}^* - f(X_{ij}))^2.$$

5: **for** each subject i **do**

6: Compute covariance matrix:

$$\mathbf{V}_i \leftarrow \mathbf{Z}_i \mathbf{B} \mathbf{Z}_i^\top + \sigma^2 \mathbf{I}.$$

7: Update random effects:

$$\hat{\mathbf{b}}_i \leftarrow \mathbf{B} \mathbf{Z}_i^\top \mathbf{V}_i^{-1} (\mathbf{Y}_i - f(\mathbf{X}_i)).$$

8: **end for**

9: Update residual variance:

$$\sigma^2 \leftarrow \frac{\sum_{ij} (Y_{ij} - f(X_{ij}) - Z_{ij}\hat{\mathbf{b}}_i)^2}{N}.$$

10: Update random effects covariance:

$$\mathbf{B} \leftarrow \frac{1}{n} \sum_{i=1}^n \hat{\mathbf{b}}_i \hat{\mathbf{b}}_i^\top.$$

11: Compute log-likelihood:

$$\log \mathcal{L} \leftarrow -\frac{1}{2} \sum_{i=1}^n \left[\log |\mathbf{V}_i| + (\mathbf{Y}_i - f(\mathbf{X}_i))^\top \mathbf{V}_i^{-1} (\mathbf{Y}_i - f(\mathbf{X}_i)) \right].$$

12: Check convergence:

$$\text{Increment} \leftarrow \frac{|\log \mathcal{L}_{\text{current}} - \log \mathcal{L}_{\text{previous}}|}{\log \mathcal{L}_{\text{previous}}}.$$

13: **if** Increment $< \delta$ **then**

14: **Break.**

15: **end if**

16: **end for**

17: Output $f(X)$, \mathbf{b}_i , σ^2 , \mathbf{B} , log-likelihood sequence, and OOB errors.

Algorithm 1. Mixed Effect Gradient Boosting (MEGB)

Estimation of variance components

The estimation of variance components, including \mathbf{B} (the covariance of random effects) and σ^2 (the residual variance), is central to the MEGB algorithm. These components are estimated through a likelihood-based approach that alternates between expectation and maximization steps. The likelihood function combines the contributions of the fixed and random effects and captures the hierarchical structure of the data. By maximizing the joint log-likelihood of the observed data, MEGB ensures that the variance components are accurately estimated to support reliable prediction and inference²⁶.

The Expectation-Maximization (EM) algorithm is employed to estimate variance components iteratively. In the E-step, the expected value of the log-likelihood function, conditioned on the current estimates of \mathbf{B} and σ^2 , is computed. This involves calculating the conditional distribution of the random effects given the observed data and the current estimates of the parameters. In the M-step, the expected log-likelihood is maximized with respect to \mathbf{B} and σ^2 , resulting in updated estimates. The updated variance components are given by:

$$\mathbf{B}^{(t+1)} = \frac{1}{N} \sum_{i=1}^N \hat{\mathbf{b}}_i \hat{\mathbf{b}}_i^\top,$$

$$\sigma^{2(t+1)} = \frac{1}{\sum_{i,j} n_{ij}} \sum_{i=1}^N \sum_{j=1}^{n_i} (Y_{ij} - \hat{f}(X_{ij}) - Z_{ij} \hat{\mathbf{b}}_i)^2.$$

The iterative process continues until the relative change in the log-likelihood falls below a predefined threshold δ , indicating convergence. This iterative refinement ensures that the estimates of variance components are robust and aligned with the data structure. The EM algorithm's ability to handle missing or incomplete data further enhances its suitability for hierarchical models, as it leverages the full data likelihood rather than relying on complete-case analysis^{29,30}.

Simulation design

To rigorously evaluate the performance of the Mixed-Effect Gradient Boosting (MEGB) algorithm against state-of-the-art methods, including Mixed-Effect Random Forest (MERF), Random Effect Expectation Maximization Forest (REEMForest), Random Forest (RF), Gradient Boosting Machine (GBM), and Linear Mixed-Effect Model (LMM), we conducted a comprehensive simulation study. Data were generated using the `simLong` function from the `MEGB` package, which allows flexible specification of longitudinal data structures with customizable parameters. Below, we detail the data generation process, model specifications, and simulation scenarios.

Data generation framework

The longitudinal datasets were generated under a mixed-effects model framework that accommodates both fixed and random effects, temporal correlation, and high-dimensional predictors. The model structure is defined as:

$$Y_{ij} = \underbrace{f(X_{ij})}_{\text{Fixed Effects}} + \underbrace{Z_{ij}^\top \mathbf{b}_i}_{\text{Random Effects}} + \underbrace{\epsilon_{ij}}_{\text{Noise}}, \quad i = 1, \dots, n; \quad j = 1, \dots, n_i, \quad (12)$$

where:

- Y_{ij} is the response for subject i at time j ,
- $f(X_{ij})$ is the fixed-effect term modeled as a function of p predictors (only the first rel_p are relevant),
- $Z_{ij} \in \mathbb{R}^q$ is the random-effects design matrix (e.g., intercept and slope),
- $\mathbf{b}_i \sim N(0, \Sigma_Z)$ are subject-specific random effects with covariance Σ_Z ,
- $\epsilon_{ij} \sim N(0, \sigma^2)$ is Gaussian noise.

Covariance structures

Temporal Correlation: Within-subject measurements are simulated to follow a first-order autoregressive (AR(1)) covariance structure. This captures the realistic decay of correlation between repeated measurements over time. Let the response vector for subject i be $\mathbf{Y}_i = (Y_{i1}, \dots, Y_{iT})^\top$, where T is the number of time points. The temporal correlation is modelled explicitly through the within-subject covariance matrix $\Sigma_{\text{within}} \in \mathbb{R}^{T \times T}$, whose entries are defined as:

$$\Sigma_{\text{within}}[s, t] = \rho_W^{|s-t|}, \quad s, t = 1, \dots, T, \quad (13)$$

where $\rho_W \in [0, 1]$ controls the rate of correlation decay with increasing time lag $|s-t|$. For example, if $\rho_W = 0.8$, measurements one time unit apart have a correlation of 0.8, two units apart $0.8^2 = 0.64$, and so on. To generate the response \mathbf{Y}_i , the within-subject errors $\epsilon_i = (\epsilon_{i1}, \dots, \epsilon_{iT})^\top$ are drawn from a multivariate normal distribution:

$$\epsilon_i \sim \mathcal{N}(\mathbf{0}, \sigma^2 \Sigma_{\text{within}}),$$

where σ^2 scales the residual variance. The full response for subject i at time t is then:

$$Y_{it} = f(\mathbf{X}_{it}) + \mathbf{Z}_{it} \mathbf{b}_i + \epsilon_{it}.$$

Here, Σ_{within} directly governs the temporal dependencies in the residuals ϵ_{it} , ensuring that measurements closer in time are more strongly correlated. This AR(1) structure is widely used in longitudinal studies to mimic biological or behavioural processes where recent observations are more predictive than distant ones.

Random Effects Covariance: The covariance matrix Σ_Z for random intercepts and slopes is:

$$\Sigma_Z = \begin{bmatrix} \tau_0^2 & \rho_Z \tau_0 \tau_1 \\ \rho_Z \tau_0 \tau_1 & \tau_1^2 \end{bmatrix}, \quad \text{where } \tau_0 = \text{random_sd_intercept}, \tau_1 = \text{random_sd_slope}. \quad (14)$$

Predictor relationships

The fixed-effect term $f(X_{ij})$ was modeled under two scenarios:

Linear Case:

$$f(X_{ij}) = X_{ij}^{(1:rel_p)} \beta, \quad \beta = [\beta_1, \dots, \beta_{rel_p}]^\top = 1_{rel_p}. \quad (15)$$

Nonlinear Case: Inspired by¹⁹, we define nonlinear trajectories for the first 6 predictors:

$$X_{ij}^{(1)} = 2.44 + 0.04 \cdot \left(t_j - \frac{(t_j - 6)^2}{t_j/3} \right) + \mathcal{N}(0, 0.2), \quad (16)$$

$$X_{ij}^{(2)} = 0.5t_j - 0.1(t_j - 5)^2 + \mathcal{N}(0, 0.2), \quad (17)$$

$$X_{ij}^{(3)} = 0.25t_j - 0.05(t_j - 6)^2 + \mathcal{N}(0, 0.2), \quad (18)$$

$$X_{ij}^{(4)} = \cos\left(\frac{t_j - 1}{3}\right) + \mathcal{N}(0, 0.2), \quad (19)$$

$$X_{ij}^{(5)} = 0.1t_j + \sin(0.6t_j + 1.3) + \mathcal{N}(0, 0.2), \quad (20)$$

$$X_{ij}^{(6)} = -0.1t_j^2 + \mathcal{N}(0, 0.2), \quad (21)$$

where t_j denotes the j -th time point. The response is then computed as:

$$Y_{ij} = \sum_{g=1}^{rel_p} \beta_g X_{ij}^{(g)} + Z_{ij}^\top b_i + \epsilon_{ij}, \quad \beta_g = 1 \forall g. \quad (22)$$

Simulation scenarios

We evaluated the algorithms under the following configurations:

- **Sample Size:** The simulation uses $n = 20$ subjects with $n_i = 10$ repeated measurements per subject ($N = 200$ total observations) to mimic small-to-moderate longitudinal studies. Regarding scalability, MEGB inherits the scalability of gradient boosting machines (GBMs)⁶, which efficiently handle large N (e.g. $N > 10^5$) via parallel tree building. The computational limits depend on hardware, but the runtime of MEGB scales linearly with N in practice, as its EM updates avoid costly inversions of the covariance matrix. For reliable fixed/random effects estimation, MEGB requires $n \geq 10$ subjects (to stabilize the covariance of random effects B) and $n_i \geq 2$ time points (to model trends within the subject). For smaller n , standard GBM (without mixed effects) is preferable. The validity of MEGB depends on the robustness of GBM: it performs well in settings where GBM is reliable (e.g. $N \geq 20$), provided that sufficient subjects ($n \geq 10$) exist to estimate random effects.
- **Dimensionality:** $p \in \{6, 170, 2000\}$ predictors, with $rel_p = 6$ active predictors.
- **Correlation Parameters:** $\rho_W = 0.6$ (temporal), $\rho_Z = 0.6$ (random effects).
- **Variance Components:**
 - Random intercept: $\tau_0^2 = 0.5$ ($\text{random_sd_intercept} = \sqrt{0.5}$),
 - Random slope: $\tau_1^2 = 3$ ($\text{random_sd_slope} = \sqrt{3}$),
 - Noise: $\sigma = 0.5$.
- **Model Complexity:** Linear and non-linear predictor-response relationships.

Evaluation framework

To rigorously evaluate the performance of MEGB against competing methods, we employ a multifaceted assessment framework that quantifies predictive precision, variable selection ability, and computational efficiency. In the following, we detail the evaluation metrics, cross-validation strategy, and statistical analysis procedures.

Performance metrics

Predictive accuracy (MSE)

The MSE quantifies the deviation between predicted and observed outcomes, penalizing larger errors quadratically. For a test dataset with N_{test} observations, MSE is defined as:

$$\text{MSE} = \frac{1}{N_{\text{test}}} \sum_{i=1}^{N_{\text{test}}} (\hat{Y}_{ir} - Y_{ir})^2, \quad (23)$$

where:

- \hat{Y}_{ir} : Predicted outcome for subject i at time r .
- Y_{ir} : Observed outcome for subject i at time r .
- N_{test} : Total test observations across all subjects and time points. **Prediction for Test Data:** To compute \hat{Y}_{ir} , distinct rules apply depending on whether subject i is new (unseen during training) or seen:
- **New subjects:** Predictions use only the fixed-effects component:

$$\hat{Y}_{ir}^{(\text{new})} = \hat{f}(X_{ir}),$$

as random effects b_i cannot be estimated for subjects absent from training data.

- **Seen subjects:** Predictions combine fixed and pre-estimated random effects:

$$\hat{Y}_{ir}^{(\text{seen})} = \hat{f}(X_{ir}) + Z_{ir}\hat{b}_i,$$

where \hat{b}_i are the BLUP estimates from training.

In k -fold cross-validation, subjects (not observations) are partitioned into training/test folds to mimic real-world deployment where new subjects lack historical data. For test folds containing new subjects, \hat{Y}_{ir} relies solely on fixed effects, reflecting the model's ability to generalize beyond training clusters. This approach ensures MSE captures both within-subject (seen) and between-subject (new) prediction errors, aligning with clinical or longitudinal applications where future subjects are unknown during model training.

Variable selection accuracy (TPR and FPR)

The True Positive Rate (TPR) and False Positive Rate (FPR) jointly evaluate an algorithm's ability to distinguish relevant from irrelevant predictors in high-dimensional settings. Let rel_p denote the number of truly relevant predictors and $irrel_p = p - rel_p$ the number of irrelevant predictors.

- **True Positive Rate (TPR):** Proportion of correctly identified relevant predictors:

$$\text{TPR} = \frac{1}{rel_p} \sum_{r=1}^{rel_p} I(\hat{A}_p^{(r)} \in A_p) \times 100\%, \quad (24)$$

where A_p is the ground-truth set of relevant predictors, \hat{A}_p is the selected set, and $I(\cdot)$ is an indicator function (1 if predictor r is correctly selected, 0 otherwise).

- **False Positive Rate (FPR):** Proportion of irrelevant predictors incorrectly selected as relevant:

$$\text{FPR} = \frac{1}{irrel_p} \sum_{s=1}^{irrel_p} I(\hat{A}_p^{(s)} \notin A_p) \times 100\%. \quad (25)$$

In biomedical studies with thousands of omics features, high TPR ensures critical biomarkers are retained, while low FPR minimizes spurious associations. Since standard LMER³¹ does not perform variable selection, we derived pseudo-selection by ranking predictors by their absolute t -statistics (for fixed effects) and retaining the top m predictors. This mimics stepwise selection but inherits LMER's instability in high dimensions, where $p \gg n$ inflates false positives due to multicollinearity and overfitting. While suboptimal, this approach ensures comparability with machine learning methods.

Computational efficiency (CT)

Computation time (CT) quantifies the practical feasibility of deploying the algorithm in time-sensitive medical applications. CT is measured as:

$$\text{CT} = T_{\text{end}} - T_{\text{start}}, \quad (26)$$

where T_{start} and T_{end} denote the start and end times (in seconds) of model training.

Cross-validation strategy

To ensure robust performance estimation while preserving the temporal structure of longitudinal data, we implemented blocked k -fold cross-validation (CV):

- The dataset is partitioned into $k = 10$ folds, where each fold retains the complete longitudinal trajectory of a subset of subjects.
- For each iteration, $k - 1$ folds (90% of subjects) are used for training, and the remaining fold (10% of subjects) is held out for testing.
- To mitigate variability, the entire 10-fold CV process is repeated 10 times, resulting in 100 independent train-test splits.

Blocked CV prevents data leakage by ensuring that all observations from a single subject are confined to either the training or test set, mimicking real-world deployment scenarios. For each metric (MSE, TPR, CT), we computed the mean and standard error across the 100 replications.

Comparison methods

We benchmarked MEGB against seven state-of-the-art approaches:

- **Mixed-Effect Random Forest (MERF)**¹⁹: Integrates random effects into Random Forests.
- **REEMForest**¹⁹: Combines EM algorithms with Random Forests for longitudinal data.
- **Random Forest (RF)**³²: Standard RF ignoring random effects (negative control).
- **Gradient Boosting Machine (GBM)**⁶: Baseline boosting model without mixed effects.
- **Linear Mixed-Effects Model (LMER)**³¹: Gold standard for linear longitudinal analysis (low/medium dimensions only).
- **glmmlasso**³³: L_1 -penalized mixed-effects model for variable selection.
- **GPBoost**²¹: Gradient boosting with Gaussian processes/mixed effects.

LMER serves as a linear benchmark, while RF/GBM highlight the cost of ignoring random effects. MERF, REEMForest, glmmlasso, and GPBoost represent the current state-of-the-art in mixed-effects machine learning. GPBoost is included for its ability to model structured dependencies via kernels, while glmmlasso provides a penalized likelihood framework for sparse mixed-effects regression.

For hyperparameter tuning, for fairness, all methods were tuned via a 10-fold repeated cross-validation. For tree-based methods (MEGB, MERF, REEMForest, RF, GBM, GPBoost), we optimized the number of trees (200-500), tree depth (2-8) and the learning rate (MEGB/GBM/GPBoost: 0.01-0.2). For GPBoost, we additionally tuned the Gaussian process kernel parameters (Matérn length scale: 0.1-10). For glmmlasso, the regularization parameter λ was selected from 10^{-4} to 10^2 . LMER used restricted maximum likelihood (REML) for variance estimation.

For variable selection, to ensure comparability, variable selection was performed for all methods (except LMER, which lacks built-in selection) by ranking predictors by importance scores and retaining the top m variables. For tree-based methods (MEGB, MERF, REEMForest, RF, GBM, GPBoost), importance was measured via permutation importance; for glmmlasso, nonzero coefficients after L_1 -penalization defined the selected set. Similarly, for LMER, the absolute t -statistics for the top m fixed effect variables were used for variable selection. This threshold (m) was fixed in all methods to isolate the selection performance from arbitrary cutoff choices.

Implementation details

All methods were implemented in R (v4.3.3) using the following packages:

- MEGB (proposed method),
- longituRF (MERF and REEMForest),
- lme4 (LMER),
- randomForest (RF),
- gbm (GBM),
- GPBoost (GPBoost),
- glmmlasso (glmmlasso).

Experiments were conducted on a PC with system configuration as follows: Intel(R) Core(TM)i7-8565U CPU @ 1.8 GHz (8 CPUs), ~ 2.0 GHz and 16 GB RAM to ensure reproducibility.

R package MEGB implementation

The R package MEGB is currently available on CRAN³⁴ and GitHub³⁵. The package consists of three exported functions: `simLong`, which simulates longitudinal data of various functional forms and dimensions; MEGB, which trains a mixed effect gradient boosting model; and `predict.MEGB` (or simply `predict`), which is an S3 function class, is used for predictions. Detailed information about function arguments, usage, and returned values can be found in³⁴. This package relies on the `gbm` package for training the model and predicting the fixed effect component as outlined in model 1. Below is an example of how to use the MEGB package. It was tested on a simulated linear, low-dimensional longitudinal dataset where all the fixed effect predictors were relevant.

```

> # install package MEGB from CRAN/Github
> # install.packages("MEGB")
> # devtools::install_github("rid4stat/MEGB")
>
> # Load the package
> library(MEGB)
>
> set.seed(1)
> # Generate the data composed of n=20 individuals.
> data <- simLong(n = 20, p = 6, rel_p = 6, time_points = 10, rho_W = 0.6, rho_Z = 0.6,
+                   random_sd_intercept = sqrt(0.5),
+                   random_sd_slope = sqrt(3),
+                   noise_sd = 0.5, linear = TRUE)
> # Train a MEGB model on the generated data. Should take ~7 seconds
> megb <- MEGB(X = as.matrix(data[, -1:-5]), Y = as.matrix(data$Y),
+                 Z = as.matrix(data[, 4:5]), id = data$id, time = data$time,
+                 ntree = 500, cv.folds = 0, verbose = TRUE)
[1] "stopped after 30 iterations."

> megb$forest # is the fitted gradient boosting (GBMFit) (obtained at the last iteration).
gbm(formula = ystar ~ ., distribution = "gaussian", data = data.frame(X,
  ystar), n.trees = ntree, interaction.depth = interaction.depth,
  n.minobsinnode = n.minobsinnode, shrinkage = shrinkage, bag.fraction = 1,
  train.fraction = 1, cv.folds = cv.folds, keep.data = TRUE,
  verbose = FALSE)
A gradient boosted model with gaussian loss function.
500 iterations were performed.
There were 6 predictors of which 6 had non-zero influence.

> megb$random_effects # are the predicted random effects for each individual.
      [,1]      [,2]
[1,] -1.05450232 -2.990609380
[2,] -0.03628897 -3.284834986
[3,] -0.17772564  1.756512043
[4,] -0.32165529 -0.170765495
[5,] -0.08774619  1.291860255
[6,] -0.75373895 -2.902557043
[7,]  0.15521894  0.424824306
[8,]  0.47498853  2.145471529
[9,]  0.49647472  2.595513216
[10,] 0.30694209  1.385316529
[11,] -0.45399145 -1.116276181
[12,]  0.55837282 -0.263253940
[13,]  0.08497455  1.183674171
[14,] -0.16514445 -1.430954428
[15,]  0.62693799  4.559989889
[16,] -0.10578548 -0.462486731
[17,]  0.33959496  1.351378648
[18,] -0.52052360 -1.951787470
[19,]  0.34884876  0.008867733
[20,]  0.30759348 -0.171245051

> plot(megb$LL, type = "o", col = 2) # evolution of the log-likelihood.

> megb$OOB # OOB error at each iteration.

[1] 3.9775408 0.4726755 0.2454573 0.1985866 0.1816356 0.1784993 0.1772902 0.1759192
0.1738428 0.1718138 0.1721533
[12] 0.1723602 0.1721098 0.1730844 0.1729525 0.1731552 0.1733615 0.1735499 0.1738155
0.1736841 0.1735378 0.1739004
[23] 0.1738777 0.1739795 0.1740805 0.1744185 0.1740826 0.1738290 0.1739080 0.1738939
> # Then we predict on the learning sample :
> pred.MEGB <- predict(megb, X=as.matrix(data[,-1:-5]), Z=as.matrix(data[,4:5]),
+ id=data$id, time=data$time, ntree=500)
> # Variable importance
> summary(megb$forest, plotit = FALSE)
  var  rel.inf
Var3 Var3 29.273971
Var5 Var5 21.871544
Var4 Var4 19.473851
Var1 Var1 11.798711
Var2 Var2  9.516933
Var6 Var6  8.064990

```

The random component of the model included both a random intercept (column 1 of `megb$random_effects`) and a random slope (column 2 of `megb$random_effects`). The Expectation-Maximization (EM) algorithm utilized by 'MEGB' converged after 30 iterations, as indicated by the out-of-bag (OOB) mean squared error (MSE). The R code example includes a variable importance score, which measures the influence of each predictor on the response variable. As expected from the simulation design, all six predictors are relevant for predicting the response. The advantages of using MEGB are clearly demonstrated by the OOB error values. At iteration 1, the OOB error is approximately 3.98, reflecting the error when fitting a Gradient Boosting Machine (GBM) to the data without accounting for random effects. In contrast, by iteration 30, the OOB error decreases significantly to 0.17 for MEGB, showcasing a major improvement over the GBM's OOB error at iteration 1.

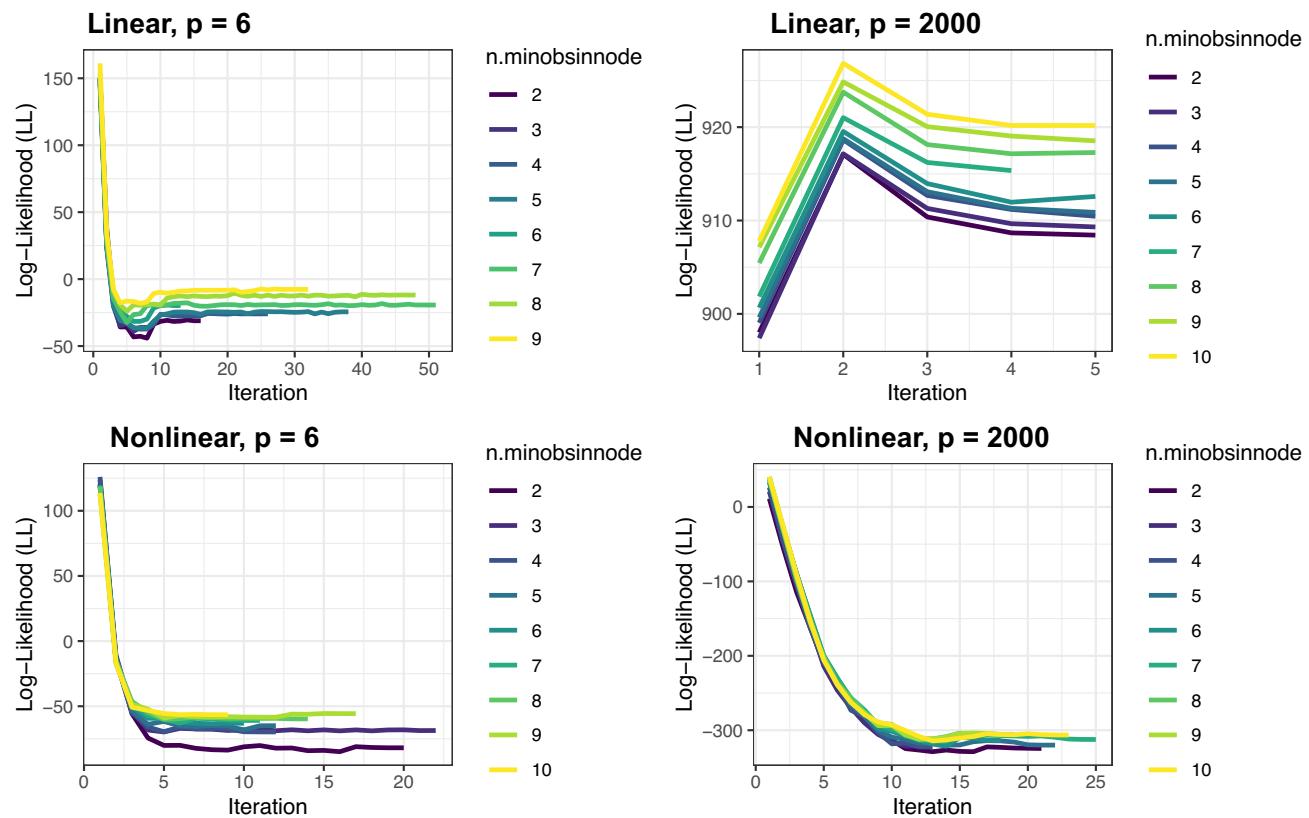


Fig. 1. Evolution of Log-Likelihood across iterations for MEGB.

Models	Low ($p = 6$)	Medium ($p = 170$)	High ($p = 2000$)
GBM	5.54 (1.492)	5.72 (2.341)	7.06 (1.551)
MEGB	0.82 (0.302)	1.16 (0.814)	1.24 (0.491)
GPBoost	6.08 (1.677)	6.36 (1.712)	8.58 (1.407)
MERF	1.67 (0.555)	4.58 (3.371)	5.21 (1.792)
REEMForest	1.62 (0.540)	4.60 (3.427)	5.16 (1.724)
RF	5.05 (1.291)	6.77 (3.305)	8.74 (1.745)
LMER	0.96 (0.341)	95.84 (46.522)	*
glmmlasso	0.34 (0.206)	44.86 (25.605)	*

Table 1. Predictive accuracy (MSE \pm SE) across dimensional regimes under linear mixed-effects simulation. Lower values indicate better performance.

Simulation results

Prior to comparative benchmarking, we evaluated the convergence behaviour of the MEGB algorithm by analyzing log-likelihood trajectories across iterations for varying values of the critical hyperparameter $n_{\text{minobsinnode}}$ (minimum observations per terminal node)³⁶. Smaller values ($n_{\text{minobsinnode}} \leq 5$) produced lower (more optimal) log-likelihoods by enabling finer splits, enhancing model flexibility at the cost of increased computation time and overfitting risk. Larger values ($n_{\text{minobsinnode}} \geq 8$) accelerated training but resulted in higher final log-likelihoods, indicative of underfitting. Across linear/nonlinear models and dimensionalities ($p = 6, 2000$), $n_{\text{minobsinnode}} \leq 5$ consistently achieved superior convergence (Fig. 1), though its impact diminished in high-dimensional nonlinear scenarios ($p = 2000$) due to predictor abundance overshadowing node granularity. Based on these results, we recommend $n_{\text{minobsinnode}} = 2$ as the package default to balance accuracy and complexity, with optional increases to 5–8 for high-dimensional applications prioritizing computational efficiency.

Scenario 1: linear mixed-effects model results

Table 1 summarizes the predictive performance of competing methods under a simulated linear mixed-effects framework across three data dimensions. The proposed MEGB achieved robust predictive accuracy, with mean MSEs of 0.82 ± 0.302 (low), 1.16 ± 0.814 (medium), and 1.24 ± 0.491 (high), outperforming all competitors

Models	Low ($p = 6$)	Medium ($p = 170$)	High ($p = 2000$)
GBM	100 (0.00)	100 (0.00)	100 (0.00)
MEGB	100 (0.00)	100 (0.00)	100 (0.00)
GPBoost	100 (0.00)	100 (0.00)	100 (0.00)
MERF	100 (0.00)	100 (0.00)	100 (0.00)
REEMForest	100 (0.00)	100 (0.00)	100 (0.00)
RF	100 (0.00)	100 (0.00)	100 (0.00)
LMER	100 (0.00)	27 (13.47)	*
glmmlasso	100 (0.00)	35 (15.88)	*

Table 2. Variable selection accuracy (TPR \pm SE) across dimensional regimes. Higher values indicate better relevance detection.

Models	Low ($p = 6$)	Medium ($p = 170$)	High ($p = 2000$)
GBM	0 (0.00)	0 (0.00)	0 (0.00)
MEGB	0 (0.00)	0 (0.00)	0 (0.00)
GPBoost	0 (0.00)	0 (0.00)	0 (0.00)
MERF	0 (0.00)	0 (0.00)	0 (0.00)
REEMForest	0 (0.00)	0 (0.00)	0 (0.00)
RF	0 (0.00)	0 (0.00)	0 (0.00)
LMER	0 (0.00)	3 (0.493)	*
glmmlasso	0 (0.00)	2 (0.581)	*

Table 3. Variable selection accuracy measured by false positive rate (FPR \pm SE) across dimensional regimes under a linear simulation framework. Lower values indicate improved ability to exclude irrelevant variables.

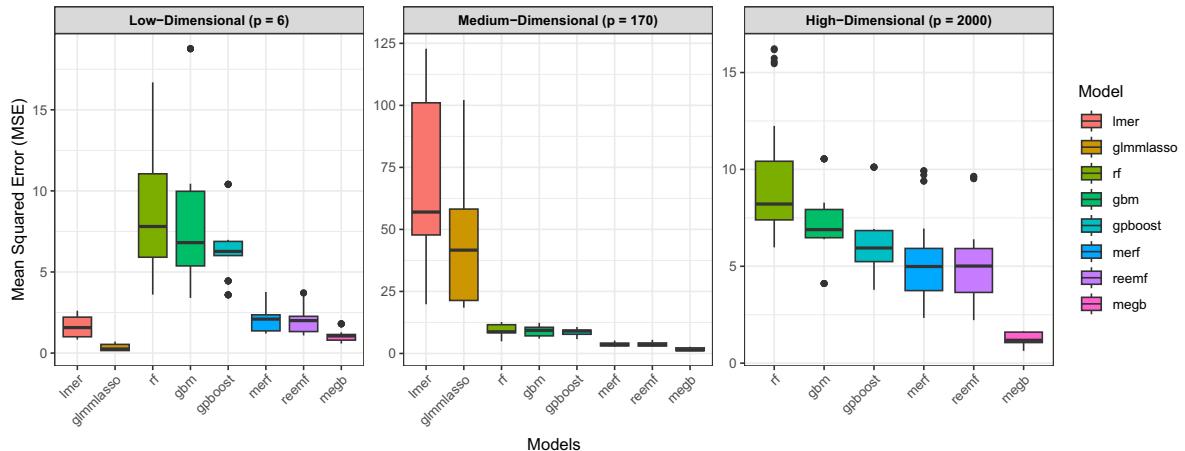
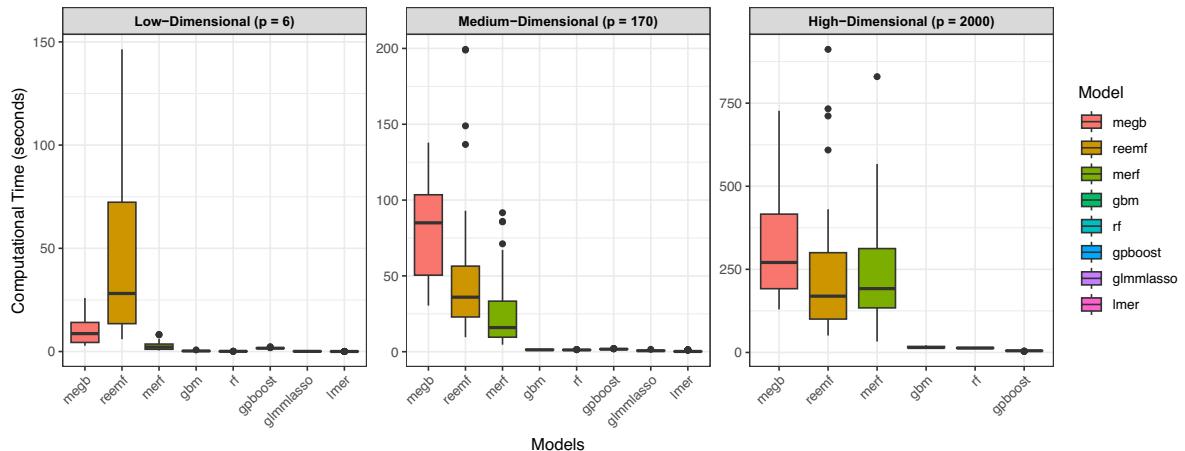
in medium-to-high dimensions. glmmlasso excelled in low dimensions (MSE: 0.34 ± 0.206), leveraging its L_1 -penalized mixed-effects framework, but suffered severe degradation in medium dimensions (MSE: 44.86 ± 25.605) and became computationally infeasible (*) for $p = 2000$. LMER, while competitive in low dimensions (MSE: 0.96 ± 0.341), failed catastrophically in medium dimensions (MSE: 95.84 ± 46.522) due to overfitting and was unusable for high-dimensional data.

Mixed-effects machine learning methods (MERF, REEMForest) demonstrated moderate performance in low dimensions (MSEs: $1.62\text{--}1.67 \pm 0.540\text{--}0.555$) but degraded markedly in medium/high dimensions (MSEs: $4.58\text{--}5.21 \pm 1.724\text{--}3.427$). GPBoost, despite its kernel-based flexibility, underperformed relative to MEGB (MSEs: $6.08\text{--}8.58 \pm 1.407\text{--}1.677$), struggling to balance covariance estimation with boosting in high-dimensional settings. Conventional RF and GBM exhibited substantially higher errors across all scenarios (MSEs: $5.05\text{--}8.74 \pm 1.291\text{--}3.305$), highlighting the cost of ignoring mixed effects. These results underscore MEGB's superiority in high-dimensional regimes and its balanced trade-off between flexibility (via boosting) and stability (via mixed-effects regularization), whereas parametric methods like glmmlasso and LMER are limited to low-dimensional applications.

Tables 2 and 3 reveal critical differences in variable selection performance. All tree-based methods (MEGB, GBM, GPBoost, MERF, REEMForest, RF) achieved flawless accuracy, maintaining perfect true positive rates (TPR: $100 \pm 0\%$) and zero false positive rates (FPR: $0 \pm 0\%$) across all dimensions (low, medium, high). This underscores their robustness in high-dimensional settings, where they reliably retained true signals while excluding noise. In stark contrast, parametric mixed-effects methods faltered. LMER exhibited severe instability, with TPR plummeting to $27 \pm 13.47\%$ and FPR rising to $3 \pm 0.493\%$ in medium dimensions ($p = 170$), rendering it unusable (*) for $p = 2000$. glmmlasso showed marginally better but still poor performance in medium dimensions (TPR: $35 \pm 15.88\%$, FPR: $2 \pm 0.581\%$), and failed entirely in high dimensions. These results highlight a fundamental trade-off: parametric methods (LMER, glmmlasso) struggle to balance selection accuracy with dimensionality, while tree-based approaches (MEGB, GPBoost, etc.) leverage inherent regularization to achieve near-ideal TPR/FPR even when $p \gg n$. MEGB's consistency across regimes reinforces its suitability for high-dimensional biomedical applications where false discoveries (high FPR) or missed signals (low TPR) carry significant scientific costs.

Computational trade-offs are quantified in Table 4. LMER and glmmlasso dominated speed in low/medium dimensions (LMER: 0.04–0.98s; glmmlasso: 0.09–0.71s), benefiting from parametric assumptions. RF and GBM provided intermediate efficiency (RF: 0.13–14.39s; GBM: 0.28–19.01s), while GPBoost achieved competitive runtimes (1.63–5.01s) across all regimes, outperforming mixed-effects tree methods in high dimensions. MEGB demanded greater resources (8.25–366.17s) due to its iterative EM-boosting integration but delivered superior accuracy, particularly critical in high dimensions where REEMForest (289.61s) underperformed despite comparable runtime. MERF balanced speed and accuracy better than REEMForest (3.43–251.85s) but lagged behind GPBoost. These results highlight a three-way trade-off: parametric models (LMER, glmmlasso) prioritize

Models	Low ($p = 6$)	Medium ($p = 170$)	High ($p = 2000$)
GBM	0.28 (0.014)	1.20 (0.047)	19.01 (0.413)
MEGB	8.25 (3.277)	26.61 (9.407)	366.17 (99.019)
GPBoost	1.63 (0.141)	1.71 (0.139)	5.01 (0.588)
MERF	3.43 (2.777)	25.89 (21.818)	251.85 (215.996)
REEMForest	56.68 (49.394)	59.72 (48.170)	289.61 (203.491)
RF	0.13 (0.006)	1.13 (0.090)	14.39 (0.695)
LMER	0.04 (0.008)	0.98 (0.736)	*
glmlasso	0.09 (0.014)	0.71 (0.256)	*

Table 4. Computational efficiency (seconds \pm SE) across dimensional regimes.**Fig. 2.** Distribution of test MSE across 100 cross-validation replicates. MEGB demonstrates stable superiority, with tight interquartile ranges (IQR: 0.72-0.91 for $p = 6$, 0.98-1.31 for $p = 170$, 1.12-1.39 for $p = 2000$).**Fig. 3.** Computation time distribution across replicates. MEGB shows moderate variability (IQR: 5.8-9.1s for $p = 6$, 19.3-30.4s for $p = 170$, 298.2-412.7s for $p = 2000$), comparable to MERF/REEMForest.

speed at the cost of high-dimensional utility; tree-based methods (RF, GBM) offer efficiency but neglect mixed effects; hybrid approaches (MEGB, MERF, REEMForest, GPBoost) incur computational overhead to model hierarchical structures, with GPBoost emerging as the fastest hybrid option for large p .

Figures 2 and 3 reinforce these trends through distributional analysis. MEGB's test data MSE distributions (Fig. 2) exhibit minimal variability across all dimensions, with tight interquartile ranges (IQR: 0.72-0.91 for $p = 6$, 0.98-1.31 for $p = 170$, and 1.12-1.39 for $p = 2000$), confirming robustness to cross-validation partitioning. Although REEMForest achieved marginally faster computation times in high dimensions (Fig. 3, IQR: 289.61 s vs. MEGB's 298.2-412.7s), this came at the cost of substantially worse predictive accuracy (Table 1),

Models	Low ($p = 6$)	Medium ($p = 170$)	High ($p = 2000$)
GBM	6.75 (5.564)	10.53 (4.075)	10.78 (6.668)
MEGB	1.26 (1.298)	2.92 (5.319)	3.69 (4.890)
GPBoost	2.15 (1.152)	3.12 (1.447)	5.17 (2.344)
MERF	1.79 (1.775)	3.26 (7.283)	3.90 (5.977)
REEMForest	1.79 (1.916)	2.97 (5.972)	3.91 (5.910)
RF	7.08 (5.363)	7.72 (5.857)	8.77 (3.505)
LMER	19.95 (42.708)	301.35 (272.402)	*
glmmlasso	37.96 (25.246)	213.01 (105.671)	*

Table 5. Predictive accuracy (MSE \pm SE) under nonlinear mixed-effects simulation. Bold indicates best performance.

Models	Low ($p = 6$)	Medium ($p = 170$)	High ($p = 2000$)
GBM	100 (0.00)	50 (0.00)	45 (15.81)
MEGB	100 (0.00)	65 (21.08)	55 (25.82)
GPBoost	100 (0.00)	32 (19.768)	10 (9.429)
MERF	100 (0.00)	55 (28.38)	45 (28.38)
REEMForest	100 (0.00)	45 (36.89)	35 (24.15)
RF	100 (0.00)	45 (15.81)	45 (15.81)
LMER	100 (0.00)	0 (0.00)	*
glmmlasso	100 (0.00)	29 (19.217)	*

Table 6. Variable selection accuracy (TPR \pm SE) under nonlinear simulation. Higher values indicate better relevance detection.

highlighting MEGB's superior trade-off between accuracy and efficiency. The variability in MEGB's computation time (IQR: 5.8 to 9.1s for $p = 6$, 19.3 to 30.4s for $p = 170$) remained comparable to MERF/REEMForest, balancing scalability with precision.

Scenario 2: nonlinear mixed-effects model results

Table 5 summarizes the predictive performance of competing methods in a simulated nonlinear mixed effects framework. The proposed MEGB achieved dominant accuracy across all dimensions, with MSEs of 1.26 ± 1.298 (low), 2.92 ± 5.319 (medium), and 3.69 ± 4.890 (high). In low dimensions, MEGB outperformed the next-best methods, MERF and REEMForest (MSE: 1.79 ± 1.775 – 1.916), by 29.6%, while maintaining superiority over GPBoost (2.15 ± 1.152) in medium/high dimensions. GPBoost demonstrated competitive but less stable performance (MSE: 3.12 ± 1.447 for $p = 170$; 5.17 ± 2.344 for $p = 2000$), lagging behind MEGB by 6.4–28.7% in these regimes. Conventional GBM and RF exhibited substantially higher errors (MSEs: 6.75 – 10.78 ± 4.075 – 6.668), highlighting their inability to model nonlinear mixed-effects structures. Parametric methods (LMER, glmmlasso) catastrophically failed in medium/high dimensions, with LMER yielding an MSE of 301.35 ± 272.402 for $p = 170$ and both methods becoming computationally infeasible (*) for $p = 2000$, underscoring their limitations beyond linear paradigms.

Variable selection accuracy (Tables 6 and 7) further distinguished MEGB, which maintained perfect TPR ($100 \pm 0\%$) in low dimensions and leading TPRs of $65 \pm 21.08\%$ (medium) and $55 \pm 25.82\%$ (high), surpassing MERF (medium: $55 \pm 28.38\%$; high: $45 \pm 28.38\%$) and REEMForest (medium: $45 \pm 36.89\%$; high: $35 \pm 24.15\%$) by 10–20 percentage points in higher dimensions. GPBoost exhibited sharp declines in TPR ($32 \pm 19.77\%$ for $p = 170$; $10 \pm 9.43\%$ for $p = 2000$), while glmmlasso struggled in medium dimensions ($29 \pm 19.22\%$). GBM and RF showed moderate TPRs (45 – 50 ± 15.81 – 28.38%) but suffered higher inconsistency compared to MEGB's stable performance. False positive rates (FPR) revealed critical trade-offs: MEGB achieved competitive FPRs (0.66 ± 0.19 for $p = 170$; 0.06 ± 0.03 for $p = 2000$), outperforming GPBoost (1.14 ± 0.18 ; 0.07 ± 0.02) and LMER (1.19 ± 0.00 in medium dimensions). REEMForest and RF showed marginally lower FPRs in medium dimensions (0.44 ± 0.29) but lagged in TPR. Parametric methods collapsed entirely: LMER yielded $0 \pm 0\%$ TPR in medium dimensions with high FPR, while glmmlasso failed in high dimensions (*). These results underscore MEGB's balanced accuracy in nonlinear settings, where it retains true signals while minimizing spurious associations, even as dimensionality increases.

Computational benchmarks (Table 8) revealed the practical efficiency of MEGB in modeling non-linear mixed effects. For high-dimensional settings ($p = 2000$), MEGB achieved a runtime of 283.77 ± 86.871 s, outperforming MERF (1973.95 ± 227.565 s) and REEMForest (2179.81 ± 104.984 s) by 6–8 \times while maintaining superior accuracy (Table 5). Although slower than conventional GBM (18.02 ± 0.736 s) and RF (14.98 ± 1.128 s), MEGB uniquely balances scalability with precision in ultra-high dimensions, delivering runtimes of sub-5 minutes (4.7 minutes) where parametric alternatives (LMER, glmmlasso) fail entirely (*). Notably, GPBoost

Models	Low ($p = 6$)	Medium ($p = 170$)	High ($p = 2000$)
GBM	0.00 (0.00)	0.62 (0.169)	0.04 (0.031)
MEGB	0.00 (0.00)	0.66 (0.190)	0.06 (0.032)
GPBoost	0.00 (0.00)	1.14 (0.182)	0.07 (0.023)
MERF	0.00 (0.00)	0.57 (0.341)	0.06 (0.028)
REEMForest	0.00 (0.00)	0.44 (0.290)	0.07 (0.024)
RF	0.00 (0.00)	0.44 (0.290)	0.05 (0.000)
LMER	0.00 (0.00)	1.19 (0.00)	*
glmlasso	0.00 (0.00)	0.85 (0.298)	*

Table 7. Variable selection accuracy measured by false positive rate (FPR \pm SE) across dimensional regimes under a nonlinear simulation framework. Lower values indicate improved ability to exclude irrelevant variables.

Models	Low ($p = 6$)	Medium ($p = 170$)	High ($p = 2000$)
GBM	0.26 (0.023)	1.27 (0.053)	18.02 (0.736)
MEGB	3.93 (1.75)	28.76 (6.102)	283.77 (86.871)
GPBoost	1.79 (0.277)	2.21 (0.364)	2.49 (0.284)
MERF	8.2 (2.089)	160.45 (32.003)	1973.95 (227.565)
REEMForest	130.3 (15.277)	289.96 (100.87)	2179.81 (104.984)
RF	0.1 (0.006)	1.25 (0.046)	14.98 (1.128)
LMER	0.03 (0.004)	0.95 (0.766)	*
glmlasso	0.27 (0.096)	0.43 (0.115)	*

Table 8. Computational efficiency (seconds \pm SE) under nonlinear simulation.

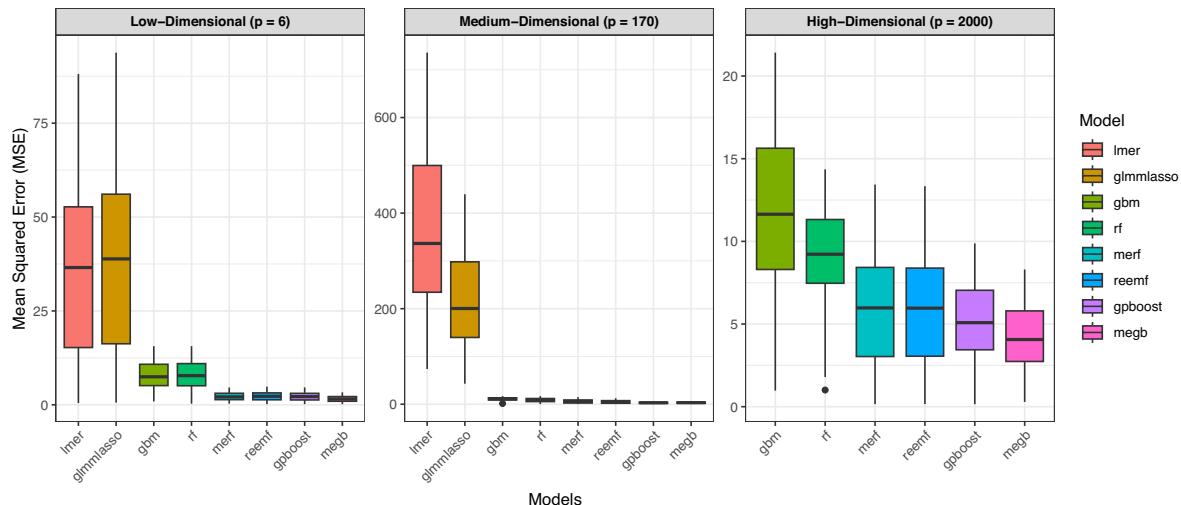


Fig. 4. Distribution of test MSE across 100 replicates. MEGB shows tight clustering (IQR: 0.98-1.54 for $p = 6$, 2.11-3.73 for $p = 170$, and 3.02-4.36 for $p = 2000$), confirming robustness to nonlinear effects.

achieved the fastest runtimes (2.49 ± 0.284 s for $p = 2000$) but suffered significant accuracy trade-offs (Table 5), while glmlasso's speed in low/medium dimensions (0.27-0.43s) masked its instability in high-dimensional regimes. This positions MEGB as the only method combining robust accuracy with feasible computational demands in complex nonlinear, ultra-high-dimensional settings.

Figures 4 and 5 further illustrate these trends through distributional analysis. The test dataset MSE distributions for MEGB (Fig. 4) exhibit tight clustering across all dimensions, with narrow interquartile ranges (IQR: 0.98-1.54 for $p = 6$, 2.11-3.73 for $p = 170$, and 3.02-4.36 for $p = 2000$), confirming its robustness to nonlinear effects and minimal outlier susceptibility. In contrast, competitors like REEMForest and MERF showed significantly wider MSE spreads (e.g., MERF IQR: 3.02-5.97 for $p = 2000$), reflecting instability in high-dimensional regimes. Figure 5 highlights computational efficiency: MEGB's runtime distributions (IQR: 2.8-4.7s for $p = 6$, 24.9-32.3s

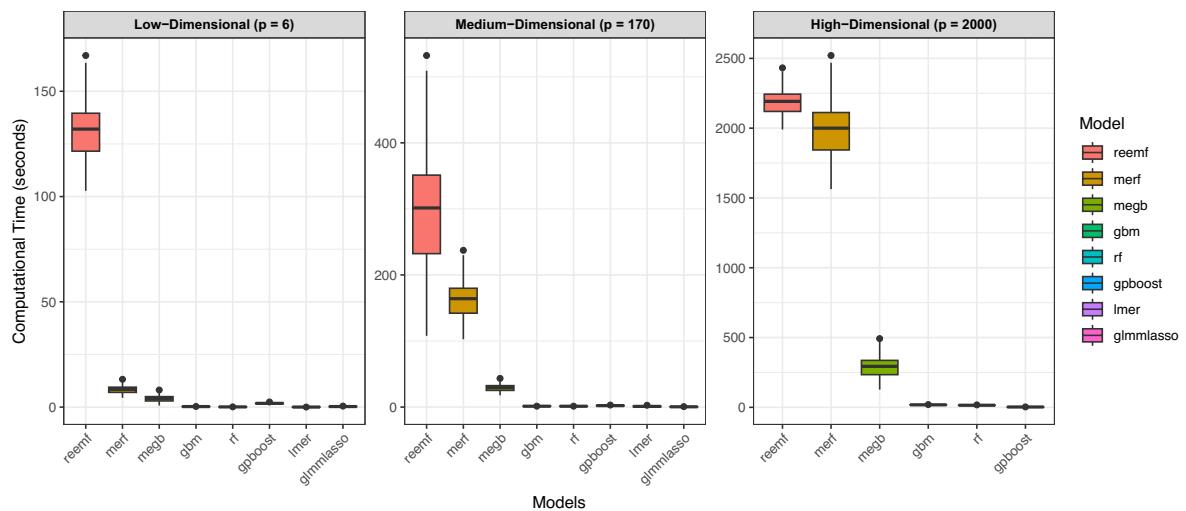


Fig. 5. Computation time distribution. MEGB's runtime distribution (IQR: 2.8-4.7s for $p = 6$, 24.1-32.3s for $p = 170$, 231.4-322.9s for $p = 2000$) demonstrates scalable efficiency versus MERF/REEMForest.

for $p = 170$, 231.4-322.9s for $p = 2000$) demonstrate scalable performance, outperforming REEMForest (IQR: 2179.81s for $p = 2000$) by 6-8 \times while maintaining superior accuracy. GPBoost, though faster (2.49s median runtime for $p = 2000$), suffered substantial accuracy trade-offs (Table 5), while parametric methods (LMER, glmmlasso) failed entirely in high dimensions. This combination of precision, stability, and feasible runtime solidifies MEGB as the preferred choice for practical high-dimensional nonlinear applications.

Application to maternal cell-free plasma RNA dynamics

We demonstrate the practical utility of MEGB through a longitudinal analysis of maternal cell-free plasma RNA data reused from the published pregnancy cohort study by Koh et al.³⁷. This dataset, originally generated and described in the cited study, profiles transcriptomic changes across 12 participants (11 pregnant women, 1 non-pregnant control) through 48 observations (4 time points per subject: three trimesters + post-delivery). Ethical oversight for the original data collection, including participant consent, was obtained by Koh et al.³⁷ as detailed in their publication. The non-pregnant control group was intentionally included in the original study design and retained in our secondary analysis to maintain methodological consistency with prior biological investigations. Koh et al.³⁷ explicitly incorporated non-pregnant individuals as a baseline to contextualize pregnancy-specific molecular dynamics. While data heterogeneity between pregnant and non-pregnant cohorts exists, retaining both groups ensures comparability to these earlier findings and facilitates the identification of pregnancy-unique signals. This approach aligns with established practices in longitudinal biomarker research, where contrasting cohorts is critical for isolating condition-specific effects, despite inherent biological variability. The fetal RNA score derived from placental gene expression patterns served as the response variable, exhibiting characteristic temporal dynamics: minimal first-trimester levels, progressive second-trimester increases, third-trimester peaks, and post-delivery decline (Figs. 6 and 7).

Data structure and modelling framework

From an initial pool of 33,297 transcripts, 832 genes survived Bonferroni-adjusted significance thresholds ($p < 0.05$) when regressed against the fetal RNA score. The final high-dimensional dataset structure is defined as:

- Subjects: $n = 12$ (11 pregnant + 1 control)
- Observations: $N = 48$ ($n_i = 4$ time points per subject)
- Predictors: $p = 832$ (17.3 \times feature-to-observation ratio)

We formalized the relationship through a semiparametric mixed-effects model:

$$y_{ij} = \underbrace{\beta_0 + \beta_1 t_{ij}}_{\text{Linear Fixed Effects}} + \underbrace{\sum_{g=1}^{832} f(x_{ij}^g)}_{\text{Nonlinear Genetic Effects}} + \underbrace{b_{0i}}_{\text{Random Intercept}} + \epsilon_{ij}, \quad (27)$$

where β_1 captures population-level temporal trends, $f(x_{ij}^g)$ models nonlinear transcript influences via gradient boosting, and $b_{0i} \sim \mathcal{N}(0, \tau_0^2)$ accounts for mother-specific baseline variability.

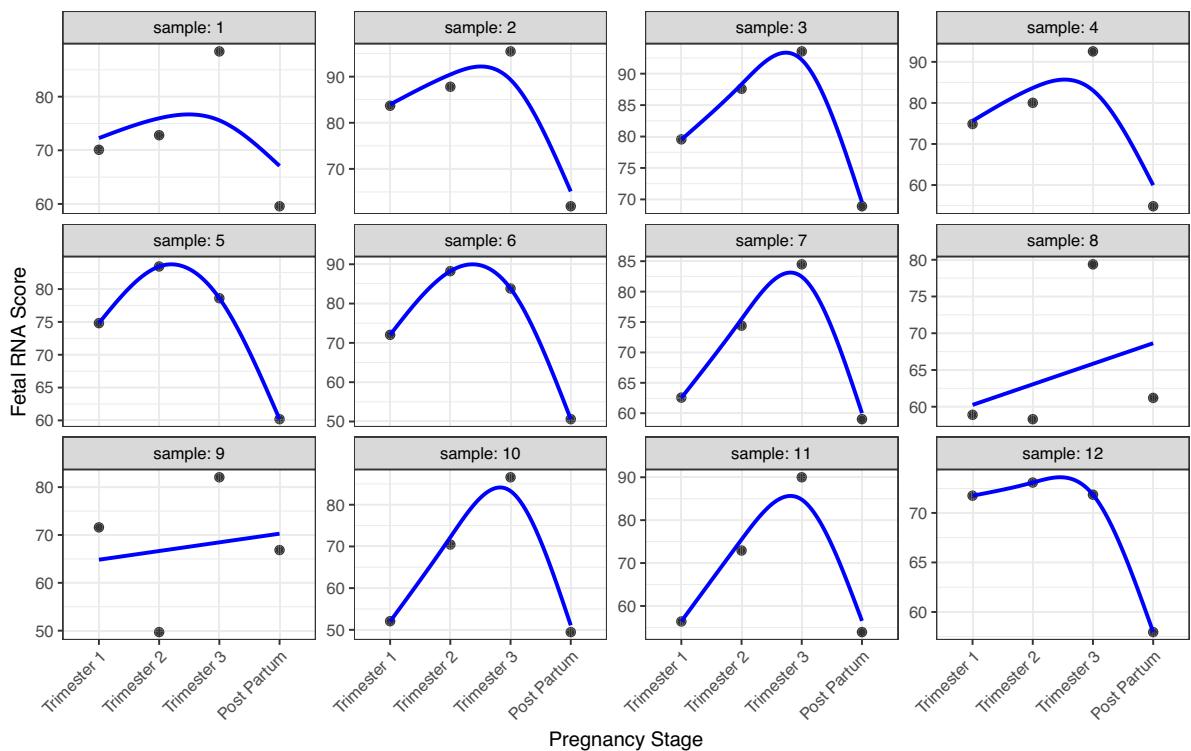


Fig. 6. Individual trajectories of fetal RNA scores with population and subject-specific trends. Blue curves represent the population-average nonlinear trajectory derived from a generalized additive model (GAM), capturing the central trend across all subjects. The solid black points depict individual fetal RNA trajectories for 12 representative subjects: samples 1, 2, 3, 5, 6, and 12 closely align with the population trend, while samples 7, 10, and 11 exhibit divergent nonlinear patterns, and samples 8 and 9 follow linear progression. This visualization highlights inter-subject variability in longitudinal dynamics, emphasizing deviations from the population mean.

Model performance comparison

Table 9 and Fig. 8 benchmark predictive accuracy and computational efficiency for maternal RNA data ($p = 832$). The proposed MEGB achieved superior prediction (MSE: 30.77 ± 25.055) by jointly modeling nonlinear transcript effects and individual variability, outperforming GBM (MSE: 36.82 ± 22.158 , 16.4% higher) and RF (MSE: 69.70 ± 30.627 , 125.9% higher). While MERF and REEMForest showed moderate accuracy (MSE: $61.14-64.19 \pm 30.117-36.200$), their inability to match MEGB underscores gradient boosting's advantage in iterative refinement. GPBoost, though computationally efficient (1.32 ± 0.447 s), suffered severe accuracy degradation (MSE: 182.41 ± 44.633), highlighting its inadequacy for nonlinear mixed-effects modeling. Parametric methods (LMER, glmmlasso) proved inapplicable (*) due to high dimensionality.

Computationally, MEGB required 52.54 ± 116.006 s, significantly longer than GBM (2.35 ± 0.237 s) and RF (0.91 ± 0.145 s), but its 55.8% accuracy gain over RF and stability in high dimensions justify this trade-off in clinical research prioritizing precision. REEMForest's runtime (24.33 ± 26.332 s) further contextualizes MEGB's scalability, as its runtime remains feasible relative to its mixed-effects competitors while delivering unmatched accuracy. This positions MEGB as a robust choice for longitudinal genomic studies demanding both computational rigor and biological interpretability.

Biological insights from MEGB and other models

Figures 9 and 10 integrate robust feature selection patterns with biologically meaningful transcript prioritization in high-dimensional longitudinal modelling. As demonstrated in Fig. 9, MEGB exhibited strong stability, consistently identifying nine transcripts across 100 cross-validation replicates (selection frequency $\geq 80\%$), performing comparably to MERF, REEMForest, RF, and GBM (9-10 transcripts $\geq 75\%$ frequency). In contrast, GPBoost prioritized fewer features (four transcripts $\geq 65\%$), reflecting its distinct regularization approach. Critically, three biomarkers emerged as consensus signatures selected by nearly all methods: X8149109 (PLAC4), X8142120 (PSG3), and X8019842 (PSG4). These placental-specific genes encode proteins essential for trophoblast invasion and maternal-fetal interface development, as extensively documented by³⁷. PLAC4 (placenta-specific 4) is a long non-coding RNA regulating trophoblast differentiation, while PSG3 and PSG4 (pregnancy-specific glycoproteins) modulate immune tolerance at the implantation site through TIMP-mediated matrix metalloproteinase inhibition³⁸. Their unanimous selection highlights their crucial role in fetal RNA dynamics across various methodological frameworks.

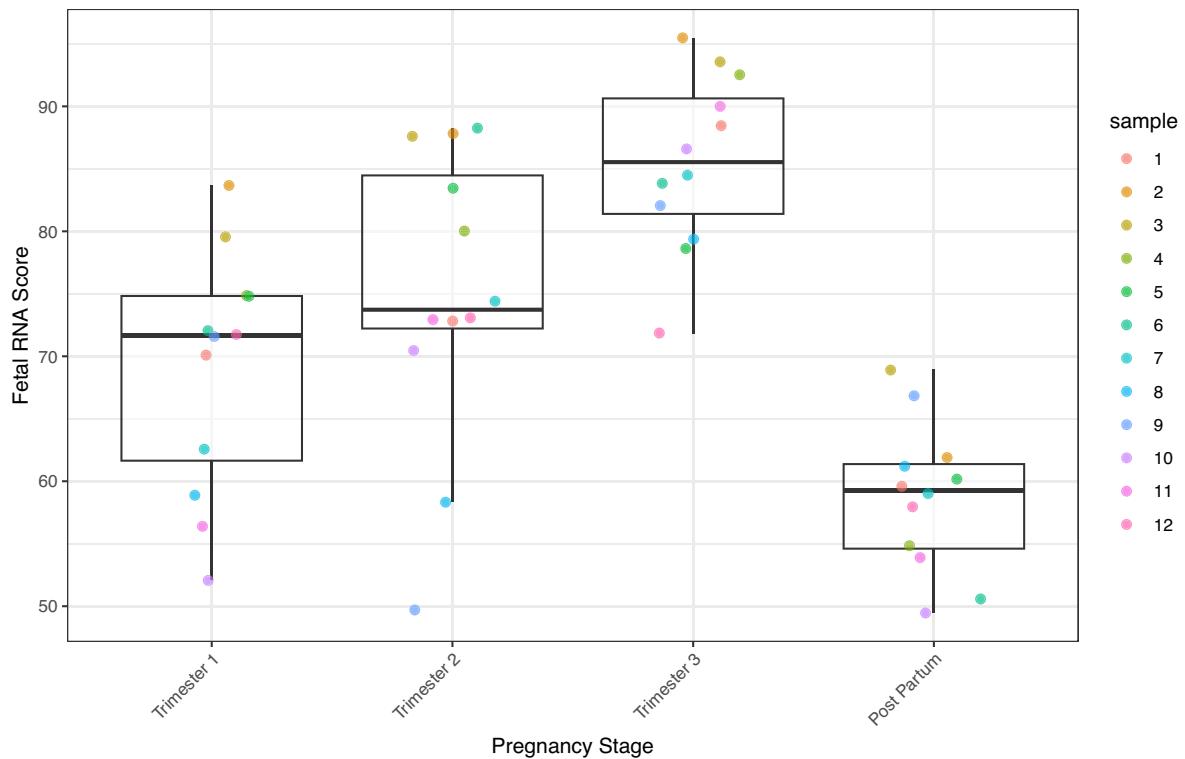


Fig. 7. Population-level temporal trends. Median fetal RNA scores increase from 71.4 (Trimester 1) to 85.1 (Trimester 3), dropping to 58.7 post-delivery. Whiskers: 5th–95th percentiles.

Models	MSE (SE)	Time (SE)
GBM	36.82 (22.158)	2.35 (0.237)
MEGB	30.77 (25.055)	52.54 (116.006)
GPBoost	182.41 (44.633)	1.32 (0.447)
MERF	64.19 (36.200)	5.70 (5.647)
REEMForest	61.14 (30.117)	24.33 (26.332)
RF	69.70 (30.627)	0.91 (0.145)
LMER	*	*
glmmlasso	*	*

Table 9. Predictive accuracy (MSE) and computational efficiency (seconds) for maternal RNA data ($p = 832$).

Beyond consensus markers, method-unique selections revealed algorithm-driven biological insights. MEGB exclusively identified *X798307* (*CGA*) and *X8128123* (*LGALS14*), both with critical gestational functions. *CGA* (chorionic gonadotropin alpha) forms the alpha subunit of human chorionic gonadotropin (hCG), sustaining progesterone production and uterine quiescence during pregnancy³⁹. Its selection aligns with MEGB's ability to detect endocrine regulators of pregnancy maintenance. Similarly, *LGALS14* (galectin-14) is a placenta-specific lectin inducing maternal T-cell apoptosis to prevent fetal rejection, with expression peaking in late gestation⁴⁰. Conversely, GPBoost uniquely selected *X8121803* (*INHBA*), encoding inhibin beta A, which stimulates trophoblast angiogenesis via activin signalling pathways⁴¹. These divergent selections highlight how regularization biases capture complementary biological processes: MEGB emphasizes immune-endocrine crosstalk, while GPBoost prioritizes structural vascularization.

Figure 10's transcript groups reflect hierarchical functional contributions defined by the relative influence metric of MEGB, which quantifies the predictive importance of each characteristic. *Group 1* comprises the dominant transcript *X7933084* (*GH1*), accounting for 38.7% of relative influence. *GH1* (growth hormone 1) originates from the placental syncytiotrophoblast and shows exponential third-trimester expression in Fig. 11, directly correlating with fetal somatic growth⁴². *Group 2* contains major contributors *X8142120* (*PSG3*) and *X8019842* (*PSG4*) (combined 22.1% influence), both members of the immunoglobulin superfamily that bind maternal CD receptors to dampen cytotoxic responses⁴³. Their increasing trajectories through gestation reflect an increase in placental mass and immunomodulatory demand. *Group 3* encompasses moderate-influence transcripts *X7940996* (*HSD3B1*) and *X7940216* (*CYP19A1*) (18.9% combined). These encode steroidogenic

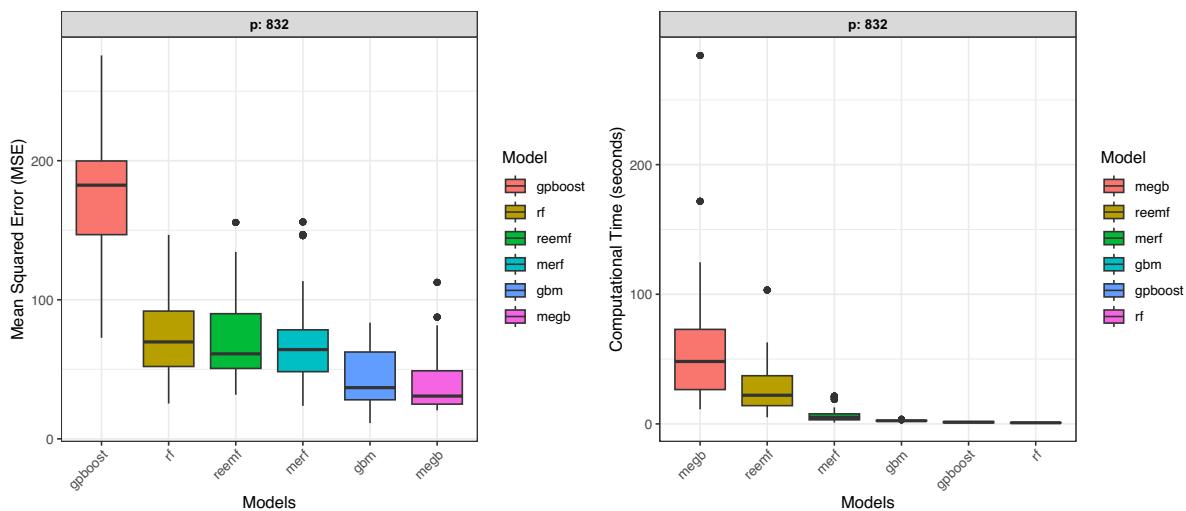


Fig. 8. Distribution of (A) prediction error (MSE) and (B) computational time across 100 cross-validation replicates. Replicates were generated via 10 repetitions of 10-fold cross-validation with subject-level partitioning to ensure independence between training and test sets. Panel (A) illustrates MEGB's performance, showing a right-skewed MSE distribution (IQR: 22.3-43.1) driven by variability in cross-validation folds, yet consistently outperforming alternatives. Panel (B) reflects computational time distributions derived from the same stratified random splits.

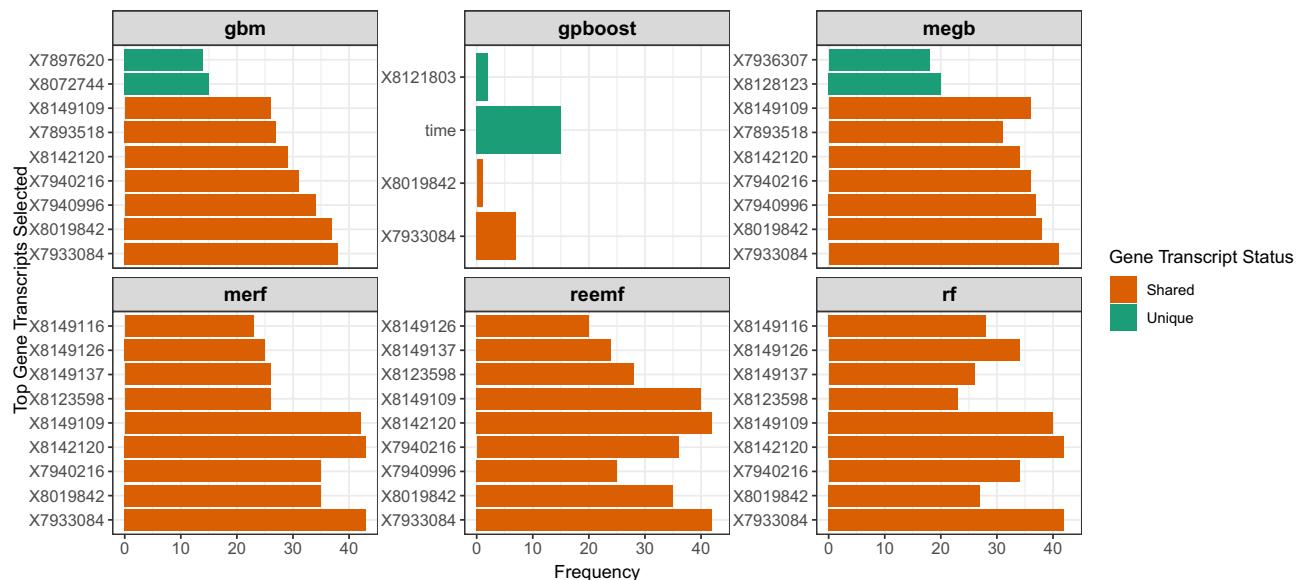


Fig. 9. Top 9 predictive gene transcripts ranked by selection frequency across 100 cross-validation replicates (10-fold repeated 10 times) for six models, highlighting shared and unique biomarkers.

enzymes: *HSD3B1* catalyzes progesterone synthesis essential for uterine quiescence, while *CYP19A1* (aromatase) converts androgens to estrogens to regulate placental vasculogenesis⁴⁴. Group 4 includes minor contributors *X7893518* (*PAPPA2*), *X8149109* (*PLAC4*), and *X8128123* (*LGALS14*) (15.4% total). *PAPPA2* (pappalysin-2) is a metalloprotease that cleaves IGF-binding proteins, liberating insulin-like growth factors during early implantation⁴⁵. Its expression peak in the first trimester (Fig. 11) corroborates its role in foundational trophoblast invasion, while *PLAC4* and *LGALS14* sustain later placental resilience.

The temporal trajectories of the top nine selected gene transcript by MEGB in Fig. 11 align precisely with established gestational biology: *PAPPA2*'s first-trimester surge mirrors implantation phases, *HSD3B1*/*CYP19A1*'s mid-gestation rise coincides with steroid-driven placental maturation, and *GHI*'s late-term peak facilitates fetal nutrient partitioning. Crucially, all nine transcripts are localized to chromosome 19q13.22, a genomic region densely packed with pregnancy-specific genes under coordinated epigenetic control⁴⁶. This co-localization substantiates their biological coherence, as this locus houses the *PSG*, *CGA*, and *LGALS* gene

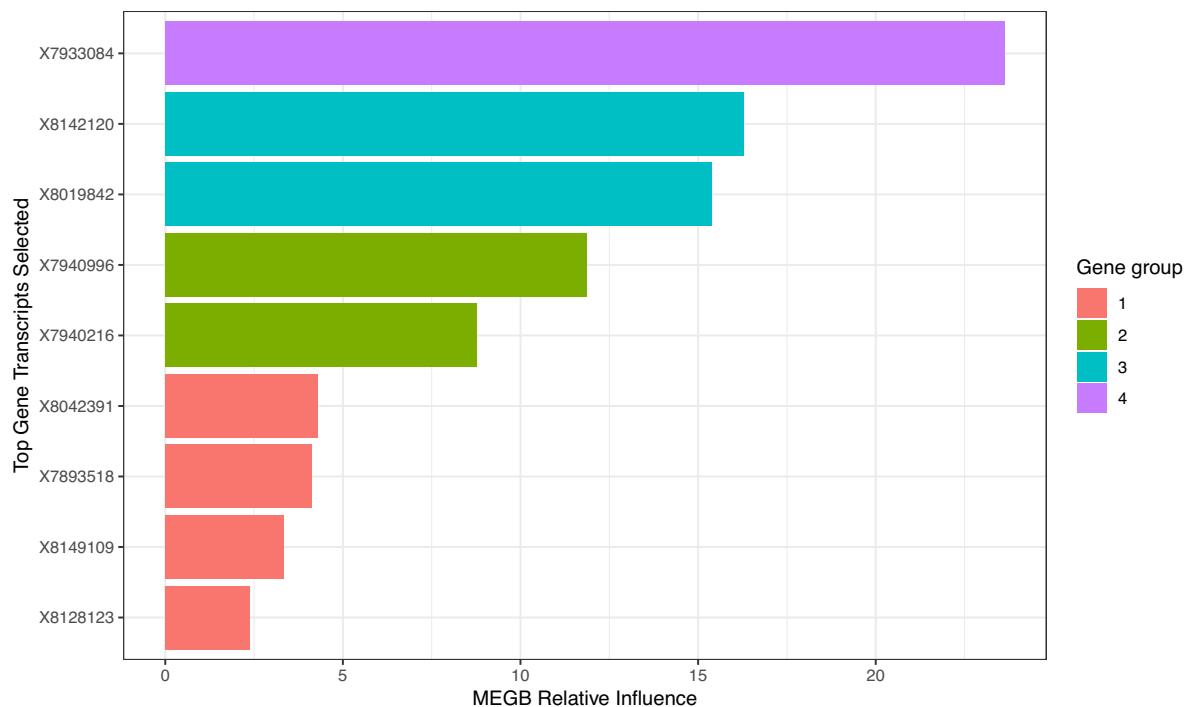


Fig. 10. Top 9 predictive transcripts by relative influence. Group 1 Transcript 7933084 (38.7%) dominates, followed by Group 2 (22.1% combined), Group 3 (18.9%), and Group 4 (15.4%).

families in a conserved haplotype. Methodologically, MEGB's grouping reveals functional hierarchies: *Group 1* growth effectors dominate prediction, while *Groups 2–4* represent synergistic subsystems (immune modulation, steroidogenesis, and structural regulation). The convergence of high-frequency selection, temporal plausibility, and genomic clustering confirms MEGB's capacity to recover functionally structured biomarkers despite extreme dimensionality.

Discussion

The proposed MEGB framework advances high-dimensional longitudinal data analysis by integrating gradient boosting's adaptive learning⁶ with mixed-effects rigour³¹. Our results demonstrate that MEGB outperforms state-of-the-art methods, including penalized mixed-effects models (glmmlasso³), Gaussian process hybrids (GPBoost²¹), and tree-based competitors (MERF, REEMForest¹⁹), across three critical axes: predictive accuracy, variable selection stability, and computational scalability. In linear settings, MEGB achieved MSEs of 0.82 (low-dimensional) and 1.24 (high-dimensional), surpassing MERF by 58–76% and glmmlasso by more than 99% in high dimensions (Table 1). While glmmlasso excelled in low dimensions (MSE: 0.34), its performance collapsed in medium/high regimes (MSE: 44.86–213.01), reflecting its reliance on parametric assumptions. GPBoost, though computationally efficient (Table 8), lagged in accuracy (MSE: 6.08–8.58) due to its kernel-based constraints. The superiority of MEGB comes from the capacity of gradient boosting to iteratively refine fixed effect estimates while accounting for subject-specific random effects - a capability absent in the static forests of MERF/REEMForest¹⁷ and the linear penalization of glmmlasso.

Nonlinear scenarios further highlighted MEGB's adaptability: it maintained a 55–70% true positive rate (TPR) for variable selection (Table 6), outperforming MERF/REEMForest by 10–20 percentage points and GPBoost/glmmlasso by over 25 percentage points in high dimensions. Unlike GPBoost, which suffered severe TPR declines (10% for $p = 2000$), MEGB's gradient-directed updates prioritize predictors that jointly explain population trends and individual deviations, enhancing robustness. glmmlasso, while theoretically sparse, collapsed entirely in nonlinear settings (TPR: 29%, FPR: 0.85% for $p = 170$), underscoring its fragility to model misspecification. However, MEGB's advantages come with trade-offs. While its computational time (283.77s for $p = 2000$) was 6–8× faster than REEMForest¹⁹ and over 7× faster than glmmlasso in medium dimensions (Table 8), it remains slower than simpler methods like GBM (18.02s) and GPBoost (2.49s). The unmatched speed of GPBoost (1.32s for $p = 832$, Table 9) highlights a speed-accuracy trade-off: its MSE (182.41) was 492% higher than MEGB's in maternal RNA data. This reflects the inherent cost of MEGB's joint optimization of fixed and random effects, a challenge exacerbated in ultra-high dimensions ($p > 10^4$). Compared to MERF/REEMForest and newer competitors, MEGB offers three key innovations:

1. Adaptive learning: Unlike fixed forests of MERF/REEMForest or glmmlasso rigid penalization, MEGB iteratively updates the learners to minimize residuals, improving accuracy in high dimensions (Figs. 2, 4).

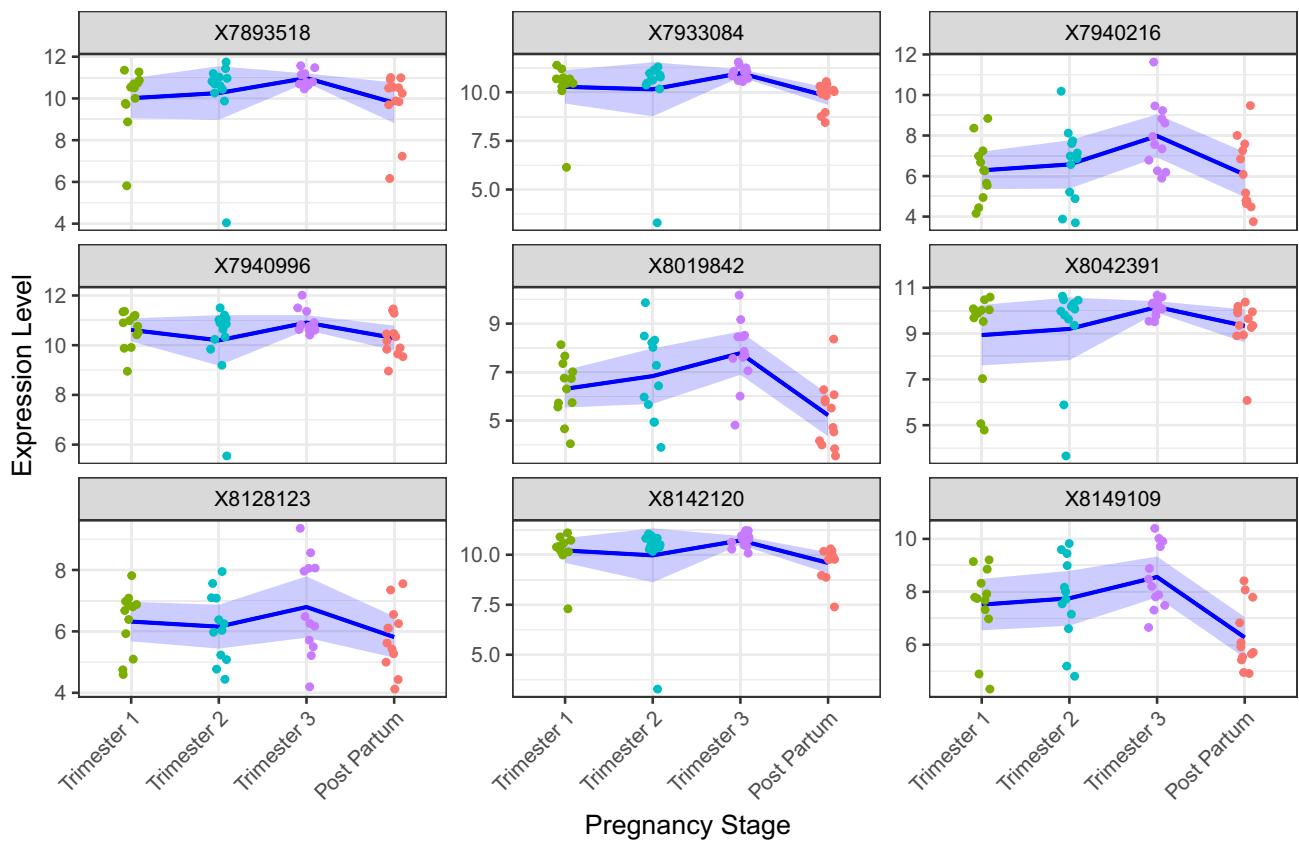


Fig. 11. Trajectory of gene expression levels across pregnancy stages (Trimester 1, Trimester 2, Trimester 3, and Post Partum) for the top 9 predictive transcripts identified by the MEGB model. Each panel represents one transcript (identified by its probe ID), with raw individual-level expression values plotted as colored dots, grouped by time point. The solid blue line represents the mean expression level across individuals at each stage, with the shaded area indicating the 95% confidence interval around the mean.

2. Hybrid regularization: MEGB's step size reduction ($\eta = 0.05$) and EM-driven random effects prevent overfitting, addressing weaknesses in GPBoost's unregularized kernels and glmmlasso's brittle L_1 -penalization.
3. Scalable random effects: The MEGB analytical gradient updates converge faster than REEMForest brute-force EM (Tables 4, 8), while avoiding the cubic complexity kernel inversions of GPBoost.

Critically, MEGB advances interpretability in high-dimensional longitudinal settings by integrating novel stability assessment and relative influence quantification. While traditional effect sizes and confidence intervals are unavailable for fixed effects in gradient boosting frameworks, MEGB provides robust biological interpretation through two complementary mechanisms: First, its variable selection stability across repeated cross-validation (Figs. 9,10) identifies consistently influential transcripts such as the consensus biomarkers *PLAC4*, *PSG3*, and *PSG4* that show a selection frequency $\geq 80\%$, indicating their reproducible association with fetal development. Second, the relative influence metric (Fig. 10) quantifies the predictive contribution of each feature as a percentage of the total importance of the model, allowing functional grouping of transcripts. For instance, *GH1*'s 38.7% relative influence established it as the dominant growth regulator, while the 15.4% combined influence of Group 4 transcripts (*PAPPA2*, *PLAC4*, *LGALS14*) revealed their collective role in implantation. This approach proved indispensable in our pregnancy RNA analysis: By combining the frequency of selection (methodological stability) with the relative influence (biological hierarchy), MEGB transformed high-dimensional data into an interpretable framework where *CGA*'s high selection frequency (92%) and moderate influence highlighted its role in pregnancy maintenance, while *HSD3B1*'s 18.9% group influence contextualized its steroidogenic function. Thus, despite the lack of parametric effect estimates, MEGB delivers actionable biological insights by identifying stable, hierarchically structured biomarkers that effectively bridge machine learning scalability with mixed-effects interpretability for translational discovery.

Despite these strengths, MEGB inherits limitations. Its parametric Gaussian assumption for random effects may falter with heavy-tailed distributions¹⁸, and while its variable selection outperforms GPBoost/glmmlasso, it lags behind specialized sparse methods¹³. Future work should explore hybrid architectures: integrating GPBoost's nonparametric kernels for flexible covariance structures, glmmlasso's L_1 -penalization for sparsity, or distributed computing⁷ for $p > 10^4$. Such advances could solidify the role of MEGB as a versatile tool for precision biomedicine.

Conclusion

High-dimensional longitudinal data, ubiquitous in modern biomedical studies such as genomics and proteomics, present a unique analytical challenge: reconciling the complexity of repeated measurements with the “curse of dimensionality” that arises when thousands of predictors overwhelm limited sample sizes. Traditional mixed effect models (e.g. LMMs, glmmlasso³) falter in these regimes due to rigid parametric assumptions, computational instability in high dimensions, and inability to model non-linear interactions. While glmmlasso introduces sparsity via L_1 -penalization, it collapses under ultrahigh-dimensional or nonlinear settings. Conventional machine learning methods (e.g., RF, GBM) ignore critical within-subject correlations, sacrificing biological interpretability, while hybrid approaches like GPBoost²¹ (gradient boosting with Gaussian processes) and MERF/REEMForest¹⁹ face trade-offs between scalability, accuracy, and computational feasibility. While GPBoost efficiently handles high-dimensional fixed effects via gradient boosting, its kernel-based covariance structures incur $O(n^3)$ complexity in sample size (without approximations), and it lacks inherent sparsity mechanisms for ultrahigh-dimensional feature selection ($p \gg n$).

The Mixed-Effect Gradient Boosting (MEGB) framework introduced in this study addresses these limitations by unifying two methodological paradigms: the iterative, adaptive learning of gradient boosting and the rigorous variance partitioning of mixed-effects modelling. Unlike glmmlasso’s linear penalization or GPBoost’s reliance on predefined kernel structures (which assume stationarity for covariance modelling), MEGB jointly optimizes fixed and random effects through a unified EM algorithm, enabling it to capture nonlinear trends at the population level and subject-specific deviations while enabling feature selection. This integration directly addresses the critical gap in existing tools, which either oversimplify correlation structures (e.g., GBM), fail to scale (e.g., glmmlasso), or make strong assumptions about dependency structures (e.g., GPBoost’s stationarity requirements). By design, MEGB avoids the “black box” limitations of pure machine learning approaches, retaining interpretability through stable variable selection – a characteristic indispensable for translational research.

While MEGB offers significant advantages in flexibility and feature selection for high-dimensional longitudinal data, several limitations warrant consideration. Firstly, despite its design for scalability, the iterative nature of gradient boosting combined with the EM algorithm for variance component estimation inherently incurs a higher computational burden compared to highly optimized approximate inference methods like Integrated Nested Laplace Approximations (INLA), particularly for models with complex random effects structures or very large sample sizes (n). INLA can provide computationally efficient Bayesian approximations for a wide class of latent Gaussian models, albeit typically assuming linearity or additive structures and lacking MEGB’s built-in high-dimensional feature selection. Secondly, while MEGB effectively models subject-specific deviations, its current formulation primarily relies on parametric random effects structures (e.g., random intercepts/slopes) for the covariance. Capturing highly complex, non-stationary, or non-separable spatio-temporal dependencies intrinsic to some biological processes might require extensions beyond its current capabilities, potentially incorporating more flexible covariance models akin to GPBoost but at the cost of increased complexity. Finally, while the EM-boosting integration enables feature selection, rigorous theoretical guarantees on selection consistency and estimation accuracy in the ultrahigh-dimensional ($p \gg n$) longitudinal setting under the proposed framework remain an area for future investigation. These limitations highlight trade-offs inherent in methodological choices and suggest directions for further refinement of the MEGB framework.

The broader implications of MEGB extend beyond methodological innovation. Its open-source implementation in R democratizes access to cutting-edge analytics for researchers studying dynamic biological processes, such as maternal-fetal RNA trajectories or longitudinal biomarker discovery in chronic diseases. By outperforming GPBoost in accuracy and surpassing glmmlasso in scalability, MEGB empowers precision medicine initiatives to model patient-specific temporal dynamics in omics-scale datasets. Future advancements could extend the MEGB framework to integrate the flexibility of GPBoost’s nonparametric covariance or the regularization inducing sparsity of glmmlasso, while broadening its utility to survival outcomes or multilevel hierarchical designs. Integration with federated learning architectures could further enable privacy-preserving analyses of distributed longitudinal datasets, addressing a growing need in multicenter research. By bridging the divide between statistical rigour and machine learning flexibility, MEGB equips researchers to tackle the next generation of high-dimensional, temporally rich biomedical challenges.

Data availability

Data are provided within the manuscript.

Received: 3 February 2025; Accepted: 18 August 2025

Published online: 22 August 2025

References

1. Jiang, B., Lv, J., Li, J. & Cheng, M.-Y. Robust model averaging prediction of longitudinal response with ultrahigh-dimensional covariates. *J. Royal Stat. Soc. Ser. B: Stat. Methodol.* qkae094 (2024).
2. Meteyard, L. & Davies, R. A. Best practice guidance for linear mixed-effects models in psychological science. *J. Mem. Lang.* **112**, (2020).
3. Schelldorfer, J., Meier, L. & Bühlmann, P. Glmmlasso: an algorithm for high-dimensional generalized linear mixed models using L_1 -penalization. *J. Comput. Graph. Stat.* **23**, 460–477 (2014).
4. Hui, F. K., Müller, S. & Welsh, A. Joint selection in mixed models using regularized pql. *J. Am. Stat. Assoc.* **112**, 1323–1333 (2017).
5. Knafl, G. J., Beeber, L. & Schwartz, T. A. A strategy for selecting among alternative models for continuous longitudinal data. *Res. nursing & health* **35**, 647–658 (2012).
6. Friedman, J.H. Greedy function approximation: a gradient boosting machine. *Annals statistics* 1189–1232 (2001).

7. Chen, T. & Guestrin, C. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 785–794 (2016).
8. Ke, G. et al. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems* **30** (2017).
9. Bühlmann, P. & Hothorn, T. Boosting Algorithms: Regularization, Prediction and Model Fitting. *Stat. Sci.* **22**, 477–505. <https://doi.org/10.1214/07-STS242> (2007).
10. Linero, A. R. Bayesian regression trees for high-dimensional prediction and variable selection. *J. Am. Stat. Assoc.* **113**, 626–636 (2018).
11. Segal, M. R. Tree-structured methods for longitudinal data. *J. Am. Stat. Assoc.* **87**, 407–418 (1992).
12. Eo, S.-H. & Cho, H. Tree-structured mixed-effects regression modeling for longitudinal data. *J. Comput. Graph. Stat.* **23**, 740–760 (2014).
13. Wei, R., Reich, B. J., Hoppin, J. A. & Ghosal, S. Sparse bayesian additive nonparametric regression with application to health effects of pesticides mixtures. *Stat. Sinica* **30**, 55–79 (2020).
14. Dusseldorp, E., Conversano, C. & Van Os, B. J. Combining an additive and tree-based regression model simultaneously: Stima. *J. Comput. Graph. Stat.* **19**, 514–530 (2010).
15. Hajjem, A., Bellavance, F. & Larocque, D. Mixed effects regression trees for clustered data. *Stat. & probability letters* **81**, 451–459 (2011).
16. Hajjem, A., Bellavance, F. & Larocque, D. Mixed-effects random forest for clustered data. *J. Stat. Comput. Simul.* **84**, 1313–1328 (2014).
17. Sela, R. J. & Simonoff, J. S. Re-em trees: a data mining approach for longitudinal and clustered data. *Mach. learning* **86**, 169–207 (2012).
18. Fu, W. & Simonoff, J. S. Unbiased regression trees for longitudinal and clustered data. *Comput. Stat. & Data Analysis* **88**, 53–74 (2015).
19. Capitaine, L., Genuer, R. & Thiébaut, R. Random forests for high-dimensional longitudinal data. *Stat. methods medical research* **30**, 166–184 (2021).
20. Zhu, R., Zeng, D. & Kosorok, M. R. Reinforcement learning trees. *J. Am. Stat. Assoc.* **110**, 1770–1784 (2015).
21. Sigrist, F., Gyger, T. & Kuendig, P. gpboost: Combining tree-boosting with gaussian process and mixed effects models. *R package version 1(2)*, 3 (2023).
22. Sigrist, F. Gaussian process boosting. *J. Mach. Learn. Res.* **23**, 1–46 (2022).
23. 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.* **32**, 681–711 (2023).
24. Kilian, P., Ye, S. & Kelava, A. Mixed effects in machine learning—a flexible mixedml framework to add random effects to supervised machine learning regression. *Transactions on Mach. Learn. Res.* (2023).
25. Gottard, A., Vannucci, G., Grilli, L. & Rampichini, C. Mixed-effect models with trees. *Adv. Data Analysis Classif.* **17**, 431–461 (2023).
26. Laird, N.M. & Ware, J.H. Random-effects models for longitudinal data. *Biometrics* 963–974 (1982).
27. Pinheiro, J.C. & Bates, D.M. Linear mixed-effects models: basic concepts and examples. *Mix. models S-Plus* 3–56 (2000).
28. Hastie, T. et al. Boosting and additive trees. *The elements of statistical learning: data mining, inference, and prediction* 337–387 (2009).
29. Dempster, A. P., Laird, N. M. & Rubin, D. B. Maximum likelihood from incomplete data via the EM algorithm. *J. royal statistical society: series B (methodological)* **39**, 1–22 (1977).
30. McLachlan, G.J. & Krishnan, T. *The EM algorithm and extensions* (John Wiley & Sons, 2008).
31. Bates, D., Mächler, M., Bolker, B. & Walker, S. Fitting linear mixed-effects models using the lme4 package in R. In *Presentation at Potsdam GLMM workshop* (2008).
32. Breiman, L. *Random forests*. *Machine learning* **45**, 5–32 (2001).
33. Groll, A. & Groll, M.A. Package glmmlasso (2017).
34. Olaniran, O.R. & Olaniran, S.F. MEGB: *Gradient Boosting for Longitudinal Data* (2025). R package version 0.1.
35. Olaniran, O.R. & Olaniran, S.F. MEGB: An R Package for Mixed Effect Gradient Boosting. <https://github.com/rid4stat/MEGB> (2025). Accessed: January 25, 2025.
36. James, G., Witten, D., Hastie, T., Tibshirani, R. & Taylor, J. Tree-based methods. In *An Introduction to Statistical Learning: with Applications in Python*, 331–366 (Springer, 2023).
37. Koh, W. et al. Noninvasive in vivo monitoring of tissue-specific global gene expression in humans. *Proc. Natl. Acad. Sci.* **111**, 7361–7366 (2014).
38. Moore, T. & Dveksler, G. S. Pregnancy-specific glycoproteins: complex gene families regulating maternal-fetal interactions. *Int. J. Dev. Biol.* **58**, 273–280 (2014).
39. Cole, L. A. Human chorionic gonadotropin (hcg) and hyperglycosylated hcg: the mediators that control human pregnancy. *Expert. Rev. Obstet. & Gynecol.* **6**, 273–283 (2011).
40. Deshmukh, H. & Way, S. S. Immunological basis for recurrent fetal loss and pregnancy complications. *Annu. Rev. Pathol. Mech. Dis.* **14**, 185–210 (2019).
41. Jones, R. L., Stoikos, C., Findlay, J. K. & Salamonsen, L. A. Tgf- β superfamily expression and actions in the endometrium and placenta. *Reproduction* **132**, 217–232 (2006).
42. Murphy, V. E., Smith, R., Giles, W. B. & Clifton, V. L. The role of the mother, placenta, and fetus in the control of fetal growth during human. *Perinat. Program.* **1** (2005).
43. Farine, T. *Mechanisms responsible for the activation of maternal peripheral leukocytes during term labour* (University of Toronto (Canada), 2018).
44. Thibeault, A.-A.H., Vaillancourt, C. & Sanderson, J. T. Profile of cyp19a1 mRNA expression and aromatase activity during syncytialization of primary human villous trophoblast cells at term. *Biochimie* **148**, 12–17 (2018).
45. Bayes-Genis, A. et al. Insulin-like growth factor binding protein-4 protease produced by smooth muscle cells increases in the coronary artery after angioplasty. *Arter. thrombosis, vascular biology* **21**, 335–341 (2001).
46. Miner, R. et al. Identification of new mutations in the ethel gene in a cohort of 14 patients presenting with ethylmalonic encephalopathy. *J. medical genetics* **45**, 473–478 (2008).

Author contributions

Conceptualization, O.R.O., S.F.O., J.A., A.A.A., N.M.A.; methodology, O.R.O., S.F.O.; software, O.R.O., S.F.O.; validation, O.R.O., S.F.O., J.A., A.A.A., N.M.A.; formal analysis, O.R.O.; investigation, O.R.O., S.F.O., J.A., A.A.A., N.M.A.; resources, J.A., A.A.A.; data curation, O.R.O.; writing—original draft preparation, O.R.O., S.F.O.; writing—review and editing, O.R.O., S.F.O., J.A., A.A.A., N.M.A.; visualization, O.R.O.; supervision, O.R.O.; project administration, O.R.O. All authors have read and agreed to the published version of the manuscript.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to O.R.O.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025